Modelling, Simulation and Identification

edited by
Azah Mohamed

SCIYO

Modelling, Simulation and Identification
Edited by Azah Mohamed

Published by Sciyo
Janeza Trdine 9, 51000 Rijeka, Croatia

Publishing Process Manager Iva Lipovic
Technical Editor Sonja Mujacic
Cover Designer Martina Sirotic
Image Copyright ultimathule, 2010. Used under license from Shutterstock.com

First published September 2010
Printed in India

A free online edition of this book is available at **www.sciyo.com**
Additional hard copies can be obtained from **publication@sciyo.com**

Modelling, Simulation and Identification, Edited by Azah Mohamed
p. cm.
ISBN 978-953-307-136-7

Contents

Preface

Modelling, simulation and identification is a topic that has been most actively researched and has yielded practical engineering applications. In modelling, mathematical models are usually derived from prior knowledge concerning the physics, describing a system which may be linear, nonlinear, continuous and discrete. Simulation is then considered as a numerical tool for calculating time responses of almost any mathematical model. Studies in the area of modelling, simulation and identification have provided a lot of useful methods and knowledge related to dynamic modelling, real-time computer-assisted simulation, on-line and off-line identification of engineering systems.

This book aims to bring together selected recent advances, applications and new ideas in the areas of modelling, simulation and identification. This book covers various methods such as signal processing, adaptive control, non-linear system identification, multi-agent simulation, eigenvalue analysis, risk assessment, modelling of dynamic systems, finite difference time domain modelling and visual feedback techniques. The scientific topics in the book play an increasingly dominant part in many areas such as electrical engineering, mechanical engineering, civil engineering, computer science and information technology.

Chapter 1 reviews several advanced signal processing techniques that are crucial for identifying power quality disturbance and predicting the source of voltage sags and incipient faults. Chapter 2 presents a comprehensive system study on the effect of voltage sag on sensitive loads such as personal computers, fluorescent lightings and ac contactors. Chapter 3 describes modelling and simulation using the finite-difference time-domain method for electromagnetic transient analysis of lightning discharge. Chapter 4 presents simulation of a power system stabilizer for a simple power system to investigate its performance in damping low frequency local oscillation using eigenvalue analysis. Chapter 5 provides an overview and comparison of the widely used wireless network simulation tools and makes recommendations on the protocols, models and simulators. Chapter 6 introduces wavelet atomic functions for reconstructing super-resolution images and video sequences. Chapter 7 describes processing of multichannel images and video sequences using order statistics and fuzzy approach. Chapter 8 introduces a new implicit multivariable pole-placement PID self-tuning controller for controlling a process under set point changes. Chapter 9 addresses a nonlinear system identification using generalized total least squares methodologies in local model networks. Chapter 10 introduces real-time visual feedback techniques using Microsoft Visual C++ and OpenGL as a graphic library for manipulation tasks. Chapter 11 presents multi-agent geo-simulation of a crowd and control forces in urban environments in order to assess different intervention strategies using non lethal weapons. Chapter 12 also presents a multi-agent geo-simulation by developing an agent-oriented language PLAMAGS. Chapter 13 introduces analytical solutions of the elastic thermal displacements and stresses of the homogeneous cross-anisotropic stratum subjected to a deep buried point heat source.

Chapter 14 presents modelling of transient ground surface displacements due to a point heat source. Chapter 15 describes a comprehensive methodology framework for the evaluation of the seismic risk and reliability of rural road networks using geographic information system software. Chapter 16 introduces a mathematical model for the oil wave journal bearing and its dynamic behavior is simulated. Chapter 17 presents a mathematical model and solution for workforce capacity planning using zero-one-integer programming.

The editor would like to thank the many authors for their contributions.

Editor

Azah Mohamed
Faculty of Engineering and Built Environment
University Kebangsaan Malaysia, Malaysia

1

Power Quality Disturbance Detection and Source Prediction Using Advanced Signal Processing Techniques

Azah Mohamed, Mohammed Abdol Salem and Mohammad Fuad Faisal
University Kebangsaan Malaysia
Malaysia

1. Introduction

Signal processing techniques have been widely used for analyzing power signals for the purpose of automatic power quality (PQ) disturbance recognition. Among the different signal processing techniques used in extracting features of disturbances from a large number of power signals, the most widely used techniques are the fast Fourier transform (FFT) and the windowed Fourier transform which comprises of the short time Fourier transform (STFT) and the wavelet transform (Moussa et al., 2004). The FFT is ideal for calculating magnitudes of the steady-state sinusoidal signals but it does not have the capability of coping with sharp changes and discontinuities in the signals. Thus, it cannot accurately detect the end of sustained events such as voltage sag, swell, transient and interruption. Although the modified version of the Fourier transform referred to as the STFT can resolve some of the drawbacks of the FFT, it still has some technical problems. In the STFT technique, its resolution is greatly dependent on the width of the window function in which if the window is of finite length, the technique covers only a portion of the signal, thus causing poor frequency resolution. On the other hand, if the length of the window in the STFT is infinite so as to obtain a perfect frequency resolution, then all the time information will be lost. Due to this reason, researchers have switched to wavelet transform from the STFT (Karami et al., 2000).

Some of the well-known wavelet transforms are the continuous wavelet mechanism transform (CWT) and a modification of the CWT which is known as the S-transform. Although CWT based multiresolution analysis monitors the regions of interest closely which is short windows at high frequencies and longer windows at low frequencies, its accuracy is susceptible to noise and if a particular frequency of interest has not been extracted due to octave filter bands, there is a chance of misclassification. To overcome this problem, the S-transform based multiresolution analysis using a variable window (Stockwell, 1996) offers significant advantage with a superior time-frequency localization property and yields amplitude and phase spectrum of the PQ event signals in the presence of noise.

The S-transform is based on a moving and scalable localizing Gaussian window and has characteristics superior to the CWT. It is fully convertible from the time domain to the two-dimensional frequency translation domain and to the familiar Fourier frequency domain.

The amplitude-frequency–time spectrum and the phase–frequency–time spectrum are both useful in defining local spectral characteristics. The superior properties of the S-transform are due to the fact that the modulating sinusoids are fixed with respect to the time axis while the localizing scalable Gaussian window dilates and translates. As a result, the phase spectrum is absolute in the sense that it is always referred to as the origin of the time axis or the fixed reference point (Stockwell et al., 1996). A significant improvement in the detection and localization of PQ disturbances can be obtained from the S-transform (Chilukuri & Dash, 2004).
In this chapter, the background theories of the FFT, CWT and S-transform are first presented. The application of CWT and S-transform for detection of single and multiple PQ disturbances, prediction of incipient faults and prediction of voltage sag sources which may be due to utility or non-utility faults are also described in this chapter.

2. Fourier Transform Theory

2.1 Fourier Series

Most of the signals in practice are time domain signals in their raw format. That is whatever that signal is measuring, it is a function of time. However, the distinguished information is hidden in the frequency content of the signal. The frequency spectrum of a signal is basically the frequency components or spectral components of that signal. The frequency spectrum of a signal indicates what frequencies exist in the signal. By analyzing a signal in time domain using the Fourier transform, the frequency-amplitude of that signal can be obtained (Langton, 2002).
Distorted waveforms can be decomposed into a fundamental component and a set of harmonics using Fourier analysis which is based on the Fourier series principle. A continuous periodic function, x(t) has three parts, namely, a dc component, a fundamental sinusoidal and a series of higher order sinusoidal components and can be expressed as,

$$x(t) = a_0 + \sum_{n=1}^{\infty} a_n \cos(w_n t) + \sum_{n=1}^{\infty} b_n \sin(w_n t) \tag{1}$$

where,

$$a_n = \frac{1}{T} \int_{-T/2}^{T/2} x(t)dt \tag{2}$$

$$b_n = \frac{2}{T} \int_{-T/2}^{T/2} x(t)\cos(nwt)dt \tag{3}$$

w_n : frequency component which is given by $w_n = 2\pi nf$
n : harmonic component

Fourier functions are symmetrical which are odd, even or half way symmetric. When the functions are odd, which means, x (t) is = -x (-t), then a_n and b_n become,

$$a_n = 0 \quad \text{and} \quad b_n = \frac{4}{T}\int_0^{T/2} x(t)\sin(nwt)dt \tag{4}$$

In the case of even functions, that is x (t) = x (-t), a_n and b_n become,

$$b_n = 0 \quad \text{and} \quad a_n = \frac{4}{T}\int_0^{T/2} x(t)\cos(nwt)dt \tag{5}$$

In half wave symmetric function case, in which x (t) = -x (t+T/2), then,

$$a_n = 0 \quad \text{and} \quad b_n = \frac{8}{T}\int_0^{T/4} x(t)\sin(nwt)dt \tag{6}$$

2.2 Fourier Transform

The Fourier transform of a continuous-time signal, x (t) is given by,

$$F(X) = \int_{-\infty}^{\infty} x(t)e^{-j2\pi ft}dt \tag{7}$$

and the inverse of the transform is given by,

$$x(t) = \int_{-\infty}^{\infty} F(X)e^{j2\pi ft}dX \tag{8}$$

The Fourier transform consists of two parts, namely the real and imaginary parts, given as,

$$F(X) = reF(X) + imF(X) \tag{9}$$

where,

$$reF(X) = \int_{-\infty}^{\infty} x(t)\cos(2\pi ft)dt \tag{10}$$

$$imF(X) = -\int_{-\infty}^{\infty} x(t)\sin(2\pi ft)dt \tag{11}$$

The magnitude and phase of the Fourier transform can be expressed as,

$$\left|F(X)\right| = [(reF(X))^2 + (imF(X))^2]^{\frac{1}{2}} \tag{12}$$

$$\phi(X) = tan^{-1}[\frac{imF(X)}{reF(X)}] \tag{13}$$

2.3 Sampling

The sampling theory states that under a certain condition it is possible to recover with full accuracy the values intervening between regularly spaced samples. The condition is that the function should be band limited and have a Fourier transform that is nonzero over a finite range of the transform variable and zero elsewhere (Bracewell, 2000). The Fourier transform F(X) is a summation of discrete signals, $x(nt)$ and it is given by,

$$F(X) = \sum_{n=-\infty}^{\infty} x(nt) e^{-j2\pi fnt} \tag{14}$$

The frequency domain function becomes,

$$x(t) = \frac{1}{f_S} \int_{-f_s/2}^{f_s/2} F(X) e^{j2\pi fnt} dX \tag{15}$$

where,

f_S : sampling frequency used to obtain the samples of a signal.

According to the Nyquist–Shannon sampling theorem, the sampling rate is twice the highest frequency in a signal. This theorem states that perfect reconstruction of a signal is possible when the sampling frequency is greater than twice the maximum frequency of a signal being sampled or equivalently, that the Nyquist frequency exceeds the highest frequency of a signal being sampled. If lower sampling rates are used, the original signal information may not be completely recoverable from the sampled signal (Shannon, 1998).

2.4 Discrete Fourier Transform

Discrete Fourier Transform (DFT) is a special case of the Fourier transform which converts time-domain sequence into an equivalent frequency domain sequence. The inverse DFT performs the reverse operation and converts frequency domain sequence into an equivalent time domain sequence. To find the Fourier transform of sampled and finite length signals, the DFT is used and it is given by,

$$X(f) = \frac{1}{N} \sum_{n=0}^{N-1} x[n] e^{-j2\pi nf/N} \tag{16}$$

where,

$x[n]$: a sequence obtained by sampling the continuous signal $x(n)$

The frequency function, $x[n]$ is then given by,

$$x[n] = \sum_{n=0}^{N-1} X(f) e^{j2\pi f/N} \tag{17}$$

2.5 Fast Fourier Transform

The Fast Fourier Transform (FFT) is a mathematical algorithm used to reduce the calculation time as compared to using the DFT. It is a powerful algorithm used in signal processing analysis made in the twentieth century (Walker, 1996). It is an efficient algorithm that is used for converting a time-domain signal into an equivalent frequency-domain signal, based on the DFT with fewer computations required. The FFT reduces the computational complexity from N in DFT to N log N multiplications. In other words, the results of FFT are the same as DFT, but the only difference is that the FFT algorithm is optimized to remove the redundant calculations. This means that for a 1024-point, the FFT needs just 10,240 operations as compared to 1,048,576 operations for the DFT. FFT is still one of the most commonly used operations in digital signal processor and all modern signal processing to provide a frequency spectrum analysis (Chassaing, 2005).
The FFT can be obtained from the DFT as follows,

$$X(f) = \frac{1}{N} \sum_{n=0}^{N-1} x[n] e^{-j2\pi nf/N} \tag{18}$$

Cooley and Tukey (1948) came up with a computational breakthrough of FFT which allows the computation of N point DFT as a function of only 2N instead of N^2. The FFT can then be written as,

$$FFT(x,f) = \frac{1}{2N} \sum_{n=0}^{2N-1} x[n] e^{-j2\pi nf/2N} \tag{19}$$

By separating $x[n]$ into its odd and even parts, the FFT is expressed as,

$$FFT(x,f) = \frac{1}{2N} \sum_{n=0}^{N-1} x[2n] e^{-j2\pi(2n)f/2N} + \frac{1}{2N} \sum_{n=0}^{N-1} x[2n+1] e^{-j2\pi(2n+1)f/2N} \tag{20}$$

3. Continuous Wavelet Transform

Wavelets are mathematical functions that divide data into different frequency components, and then study each component with a resolution matched to its scale. The fundamental idea behind wavelets is to analyze signal according to scale rather than frequency. The scale is defined as a frequency inverse. Wavelets have advantages over traditional Fourier methods in analyzing physical situations where the signal contains discontinuities and sharp spikes. Wavelet techniques can divide a complicated function into several simpler ones and study them separately. This property, along with fast wavelet algorithms which are comparable in efficiency to the FFT algorithms, makes the wavelet techniques very

attractive in analysis and synthesis problems. Different types of wavelets have been used as tools to solve problems in signal analysis, image analysis, medical diagnostics, geophysical signal processing, statistical analysis, pattern recognition, and many others. By using wavelet multiresolution analysis, a signal can be represented by a finite sum of components at different resolutions so that each component can be adaptively processed based on the objectives of the application. This capability of representing signals compactly and in several levels of resolutions is the major strength of the wavelet analysis.

There are essentially two types of wavelet transforms, continuous wavelet transform (CWT) and discrete wavelet transform (DWT). The CWT type is usually preferred for signal analysis, feature extraction and detection tasks whereas the DWT type is more appropriate for performing some kind of data reduction. CWT uses a time-window function that changes with frequency. This adaptive time window function is derived from a prototype function known as the mother wavelet which is scaled and translated to provide information in the frequency and time domains, respectively (Poularikas, 2000). Thus, a transformed signal is a function of two variables, the translation and scale parameters, respectively. The term wavelet means a small wave in which the smallness refers to the condition that this window or function is of finite length. The term mother implies a function with different region of support that is used in the transformation process. In other words, the mother wavelet is a model for generating the window functions.

The CWT of a continuous signal, is expressed in terms of wavelet coefficients, for different values of scaling factor, s and translation factor, d, as:

$$CWT_x(s,d) = \int_{-\infty}^{\infty} x(t)\psi^*_{s,d}(t)dt \tag{21}$$

where,

x (t) : signal as a function of time

$\psi_{s,d}(t)$: mother wavelet which is given as,

$$\psi_{s,d}(t) = \frac{1}{\sqrt{s}}\psi\left(\frac{t-d}{s}\right) \tag{22}$$

Substituting (22) into (21), the CWT can be written as,

$$CWT_x(s,d) = \frac{1}{\sqrt{s}}\int_{-\infty}^{\infty} x(t)\psi^*\left(\frac{t-d}{s}\right)dt \tag{23}$$

As for the scaling factor, large values of this factor provide a broad time-width windowing function located in the low frequency domain. On the other hand, small values of this factor provide a narrow time-width windowing function in the high frequency domain.

The CWT has a filter-bank interpretation in which each wavelet basis function can be thought of as a filter through which the original signal is passed. Each filter, however, has a fixed relative bandwidth as opposed to the fixed absolute bandwidth in the STFT (Poularikas, 2000).

3.1 Understanding the Mother Wavelet

All the wavelets are generated from a single basic wavelet (t), the so-called mother wavelet, by scaling and translation:

$$\Psi_{s,\tau}(t) = \frac{1}{\sqrt{s}}\Psi\left(\frac{t-\tau}{s}\right) \tag{24}$$

where,
s : scale factor
τ:translation factor

The factor $s^{-1/2}$ is for energy normalization across the different scales. The term mother implies that the functions with different region of support that are used in the transformation process are derived from one main function, or the mother wavelet. In other words, the mother wavelet is a prototype for generating the other window functions.
Each mother wavelet has its own characteristics and will project different types of resolutions. The mother wavelets selected will serve as prototypes for all windows in the process. All the windows that are used are dilated and shifted versions of the mother wavelet functions. The common question commonly asked on WT is, which mother wavelet function generates the best resolution for detection of disturbances. There are different types of mother wavelet, namely, the Morlet, Meyer, Gaussian, Mexican hat, Haar and Daubechies wavelets. The Morlet wavelet is commonly used for signal analysis and therefore it is used for analysis of power disturbances.

4. S-transform Theory

The S-transform produces a time-frequency representation of a time series. It uniquely combines a frequency-dependent resolution that simultaneously localizes the real and imaginary spectra. The basis functions for the S-transform are Gaussian modulated cosinusoids, so that it is possible to use intuitive notions of cosinusoidal frequencies in interpreting and exploiting the resulting time-frequency spectrum. With the advantage of fast lossless invariability from time domain to time-frequency domain, and back to the time domain, the usage of the S-transform is very analogous to the Fourier transform. In the case of non-stationary disturbances with noisy data, the S-transform provides patterns that closely resemble the disturbance type and, thus, requires a simple classification procedure. Furthermore, the S-transform can be derived from the CWT by choosing a specific mother wavelet and multiplying a phase correction factor. Thus, the S-transform can be interpreted as phase-corrected CWT (Lee & Dash, 2003). The S–transform generates contours, which are suitable for classification by simple visual inspection unlike wavelet transform that requires specific methods like Standard-Multi resolution analysis (Jaya et al, 2004).
By using a simple rule base or a neural network along with the features extracted from the S–transform contours, one can easily dispense with the visual inspection procedure of the S-transform. The derivation of S-transform from CWT is described as follows:
The CWT of a function is defined as,

$$W(\tau,d)=\int_{-\infty}^{\infty}h(t)w(t-\tau,d)dt \tag{25}$$

where,
$w(t,d)$: mother wavelet

The S-transform is obtained by multiplying the CWT with a phase factor as expressed below,

$$S(\tau,f)=W(\tau,d)e^{i2\pi f\tau} \tag{26}$$

Substituting (25) into (26), the S-transform can be written as,

$$S(\tau,f)=\int_{-\infty}^{\infty}h(t)w(t-\tau,d)e^{i2\pi f\tau}dt \tag{27}$$

where,
d : scale which can defined as inverse of *f*.

The mother wavelet for this particular case is defined as,

$$w(t,f)=\frac{|f|}{\sqrt{2\pi}}e^{-\frac{t^2f^2}{2}}e^{-i2\pi ft} \tag{28}$$

and can also be written as,

$$w(t,f)=G(\tau,f)e^{-i2\pi ft} \tag{29}$$

where,
$G(\tau,f)$: modulation function

Considering the mother wavelet in (29), the S-transform written in terms of $h(t)$ is defined as:

$$S(\tau,f)=\int_{-\infty}^{\infty}h(t)G(\tau-t,f)e^{-i2\pi ft}dt \tag{30}$$

The modulation function $G(\tau,f)$ is given by,

$$G(\tau,f)=\frac{|f|}{\sqrt{2\pi}}e^{-(t^2/2\sigma^2)} \tag{31}$$

where σ is a Gaussian window width which is given by,

$$\sigma(f) = T = \frac{1}{|f|} \tag{32}$$

Substituting (31) and (32) into (30), the final S-transform equation becomes,

$$S(\tau, f) = \frac{|f|}{\sqrt{2\pi}} \int_{-\infty}^{\infty} h(t) e^{-(t-\tau)^2 f^2 / 2} e^{-i2\pi f t} dt \tag{33}$$

where,
f: frequency
t and τ : time

The S-transform distinguishes itself from the many time-frequency representations available by uniquely combining progressive resolution with absolutely referenced phase information. It is known that progressive resolution gives a fundamentally sounder time-frequency representation (Daubechies, 1990). Referenced phase means that the phase information given by the S-transform is always referenced to time t = 0, which is also true for the phase given by the Fourier transform. This is true for each S-transform sample of the time-frequency space. This is in contrast to the CWT approach, where the phase of the wavelet transform is relative to the center in time of the analyzing wavelet. Thus, as the wavelet translates, the reference point of the phase translates. This is called "locally referenced phase" to distinguish it from the phase properties of the S-transform. From one point of view, local spectral analysis is a generalization of the global Fourier spectrum. However, the fundamental principle of S-transform analysis is that the time average of the local spectral representation should result identically in the complex-valued global Fourier spectrum (Stockwell, 2006). This leads to phase values of the local spectrum that are obvious and significant.
The S-transform has unique properties in which it uniquely combines frequency dependent resolution with absolutely reference phase, so that the time average of the S-transform equals the Fourier spectrum. It simultaneously estimates the local amplitude spectrum and the local phase spectrum, whereas the CWT approach is only capable of probing the local amplitude and power spectrum. It independently probes the positive frequency spectrum and the negative frequency spectrum, whereas many wavelet approaches are incapable of being applied to a complex time series. It is sampled at the discrete Fourier transform frequencies unlike the CWT where sampling is done randomly (Stockwell, 2006).

5. Application of Continuous Wavelet Transform for Power Quality Disturbance Detection

The CWT program was written using the four mother wavelet functions as shown in Figures 1 to 4. Voltage sag signals were obtained from the utility PQ monitoring for the purpose of analysing the signals using the CWT with the four mother wavelets (Faisal, 2009).

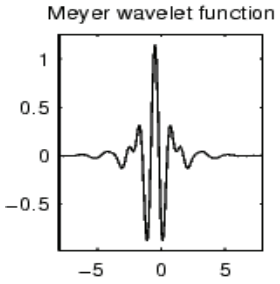


Fig. 1. Meyer Wavelet Function

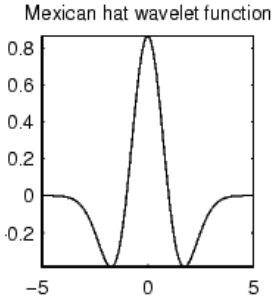


Fig. 2. Mexican Hat Wavelet Function

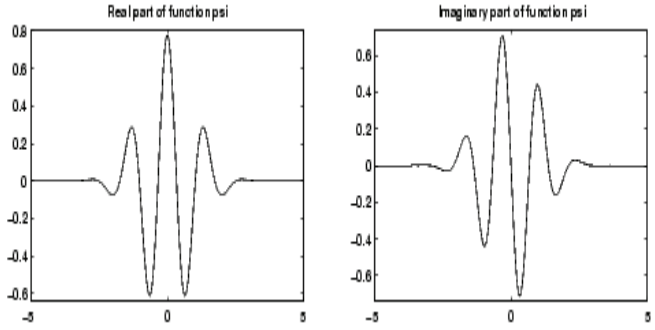


Fig. 3. Gauss Wavelet Function

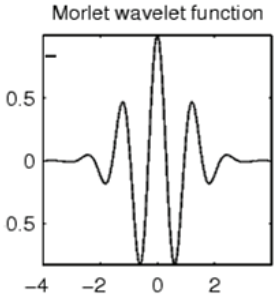


Fig. 4. Morlet Wavelet Function

5.1 CWT Analysis with the Meyer Wavelet Function

Visually, using the Meyer wavelet function (Figure 1) in the CWT analysis, voltage sags can be easily detected as shown in Figure 5. The Meyer wavelet gives high values of CWT coefficients, around 4 to 5. These high value coefficients will enable a clear distinction to be observed in the detection of voltage sags. It can also be used in the extraction of the CWT features for the classification of disturbances. In the feature extraction process, the standard deviation and the mean of the amplitude would be calculated and later used in the classification of the power quality disturbances. The scale axis which is actually the frequency inverse correctly identifies the existence of the system frequency of 50 Hz which is located at a high scale of 20.

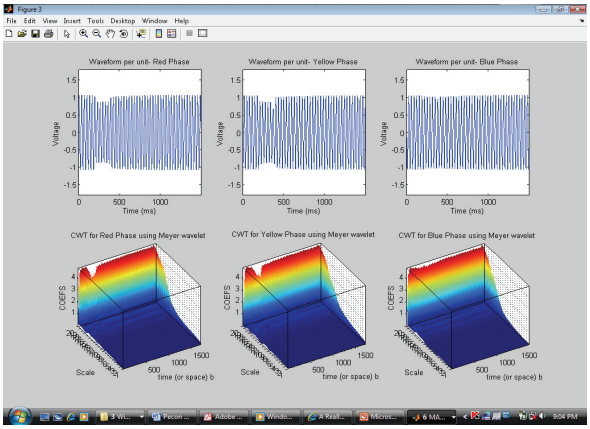


Fig. 5. Analysis of voltage sag using the Meyer wavelet

5.2 CWT Analysis with the Mexican Hat Wavelet Function

The results of CWT analyses using the Mexican Hat wavelet show that the CWT coefficients are with values around 4 as shown in Figure 6. From visual inspection, the signature that depicts a voltage sag event are not obvious by using the Mexican hat wavelet function as compared to using the Meyer wavelet.

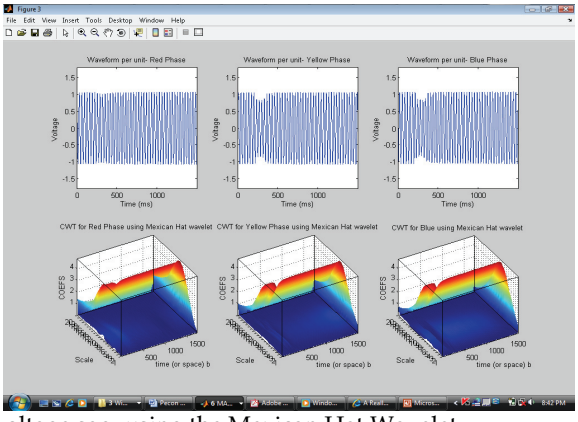


Fig. 6. Analysis of voltage sag using the Mexican Hat Wavelet

5.3 CWT Analysis with Gauss Wavelet Function

The results of CWT analyses using the Gauss wavelet function are as shown in Figure 7. Similar to the Meyer wavelet function, the Gauss wavelet enables one to detect the signature of voltage sag easily. The Gauss wavelet gives the highest values of CWT coefficients, that is, around 6. These high value coefficients will enable a very clear distinction to be observed in the detection of voltage sags.

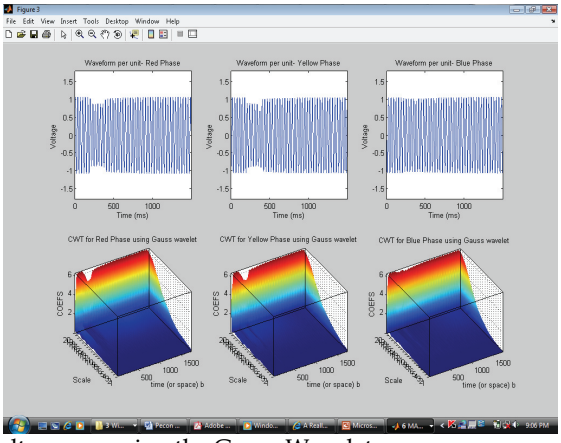


Fig. 7. Analysis of voltage sag using the Gauss Wavelet

5.4 CWT Analysis with Morlet Wavelet Function

Similar to the Meyer and Gauss wavelet functions, the Morlet wavelet enables one to visualize the signature of voltage sag easily. The Morlet wavelet gives lowest values of the CWT coefficients, that is, around 3. However, the low coefficient values will not give a very clear distinction in the extraction of the features for the purpose of disturbance classification. The contours of the CWT are also not very smooth and therefore such information may be misleading because the signature of rugged surface highlights the existence of harmonics. In this study, no harmonics exist in the signal.

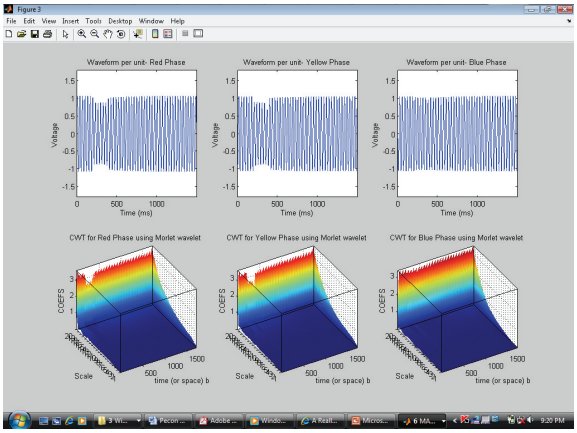


Fig. 8. Analysis of voltage sag using the Morlet Wavelet

From the analyses done on the twenty voltage sag data, it is confirmed that the best result in the detection of voltage sag will depend on the selection of the mother wavelet. The selection of a wavelet that closely matches the signal is important in the detection of the waveform. From the findings, it is noted the Gauss wavelet function gives the most accurate detection of voltage sag. The resolution of the contour is very clear and the high coefficient amplitudes are useful for the extraction of voltage sag features to be used in the power quality disturbance classification.

6. Using the S-transform for Detection of Voltage Disturbances and Incipient Faults in Power Distribution Networks

To illustrate the use of the S-transform for voltage disturbance and incipient fault detection, a case study at an industrial plant in Malaysia is presented (Faisal et al, 2009). The plant had complained of frequent occurrences of nuisance tripping and damages to it's production equipment for the last one year and put the blame on the power utility. To understand the problem, the power utility installed a power quality recorder in the plant for three months and the recorded data were then analyzed by this new technique. From the analysis of results, 3 voltage sags (Figure 9), 3 notches (Figure 10) and 12 unknown events (Figure 11) were detected. In Figure 9, the S-transform contour shows the existence of voltage sags in three phases while in Figure 10, minor notches in the S-transform contour are detected. The causes of the voltage sags were due to lightning strokes on the utility power lines. The cause of the notches (Figure 10) was found out after implementing a thermal scan at the main switch board conductors in which the cause was due to lose connection at the blue phase conductor. The results in Figure 11, showed frequent occurrences of incipient faults at the red phase of a factory power supply. Overall, 12 events were recorded at the red phase.
Locating faults based solely on substation measurements has always been time consuming and difficult, and locating subtle incipient faults is even more challenging because of the low magnitude signals often involved. The signature of the change in the feeder current due to incipient fault is very delicate and difficult to detect. However, based on analysis performed in this study, the S-transform is able to perform efficient detection and isolation for both incipient and abrupt faults in power supply systems. Thus, the existence of the incipient fault is only detectable by using the S-transform.

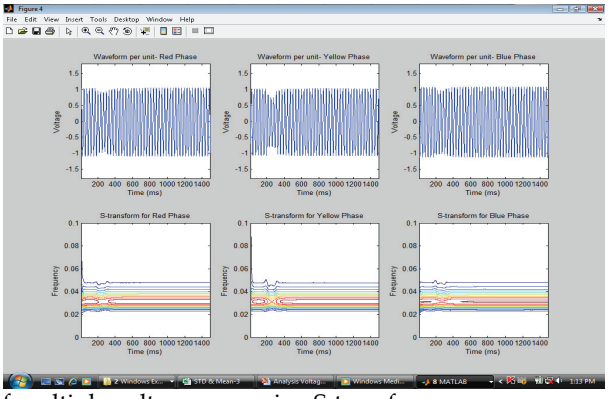


Fig. 9. Detection of multiple voltage sags using S-transform

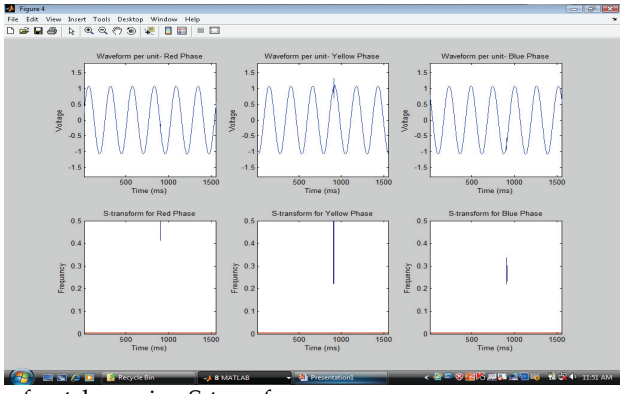


Fig. 10. Detection of notches using S-transform

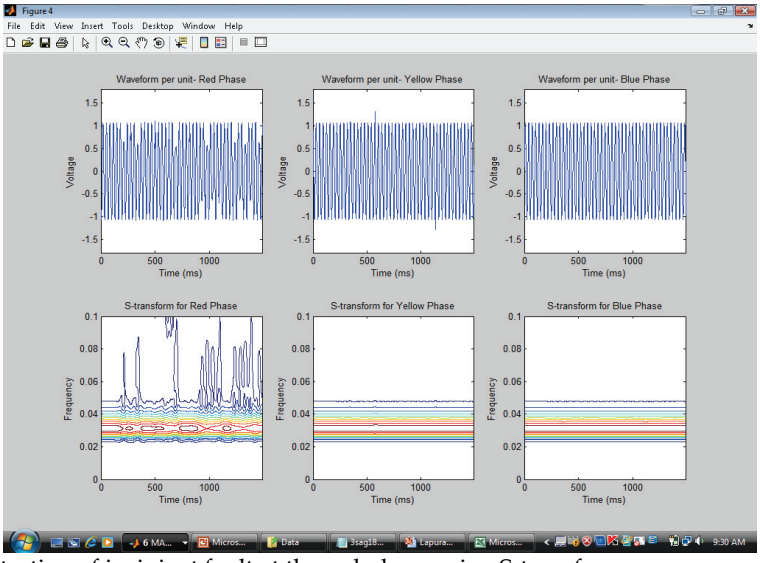


Fig. 11. Detection of incipient fault at the red phase using S-transform

7. Using the S-transform for Voltage Sag Source Prediction

The identification of sources of disturbances is very critical in power quality diagnosis so as to provide the necessary information to customers for resuming back their operations. A disturbance source can originate from either inside a facility or outside in a distribution network. One method of predicting the source of voltage sag at a monitoring point is by determining whether the source is from either upstream or downstream. The concept of automatic sag source prediction is shown in Figure 12. A power quality recorder (PQR) is installed at point M in a power supply network. If voltage sag occurs in the network, the PQR will detect and record the disturbance depending upon its voltage threshold setting which is based on the definition of voltage sags. The source of sag can appear either

upstream or downstream with respect to point M. Upstream side can be defined as the side that supplies the fundamental power into the monitoring point at steady state conditions whereas the downstream side is defined as the side that leaves the fundamental power from the monitoring point.

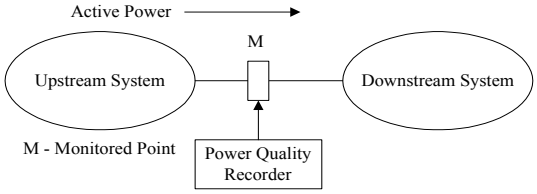


Fig. 12. Concept of upstream and downstream voltage sags

Several methods have been developed for performing sag source prediction (Polajžer, 2007). The obtained results showed that all the existing methods do not work well, particularly in cases of asymmetrical voltage sags due to upstream events and therefore further development is still needed to increase the degree of sensitivity and confidence in the existing techniques for performing automatic sag source prediction. Here, a novel method based on the S-transform is proposed for improving the sensitivity of the sag source prediction (Faisal & Mohamed, 2009). The S-transform is used for producing the time-frequency representation for the voltage and current waveforms. The time-averaged amplitudes and spectral contents for all these signals are extracted from the time-frequency values which are then used to determine the origin of voltage sags.

In the proposed voltage sag source prediction method, the features that characterize both the voltage and current waveforms are extracted from the time and frequency resolutions of the S-transform. The output of the S-transform is an N x M matrix called the S-matrix whose rows pertain to the frequency and columns to time. The S-transform will generate time frequency contours, which clearly display the disturbance patterns for ease of visual inspections. It will also generate the relevant contours for both voltage and current waveforms. From these contours, the values of the disturbance voltages, currents and powers are derived.

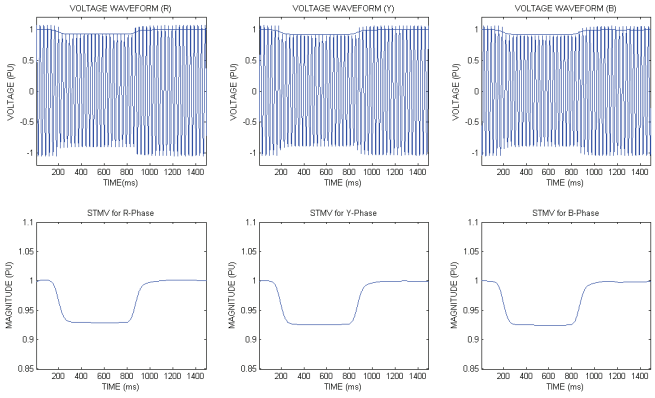


Fig. 13. Plots of the S-matrix locus for voltage values

In this study, the first set of features is based on the maximum values for all the columns in the S-transform contours for voltages. Examples on the graphical representations of the maximum value plots for voltages are shown in Figure 13. The first rows showed the original waveforms for the voltage values. The plots in the second rows are the loci of the maximum values for voltage from the S-transform. Features can be extracted from these maximum value plots to characterize voltage sags and swells.
The second sets of features are selected based on the maximum values for all the columns in the S-transform for currents and disturbance powers. The values are termed as S-transform maximum current (STMI) and S-transform disturbance power (STDP). Examples on the maximum value plots for currents and powers are shown in Figures 14 and 15. In Figure 14, the first rows showed the original waveforms for the current values. The plots in the second rows are the loci of the maximum values for currents from the S-transform. In Figure 15, the plots show the loci of the S-transform disturbance powers for each phase during the voltage disturbances. The disturbance powers are derived from the voltage and current contours of the S-transform. Features can also be extracted from the current and disturbance power plots to identify the origin of the disturbances.

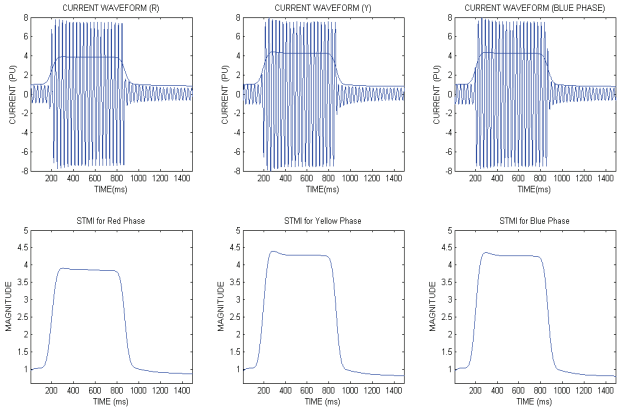


Fig. 14. Plots of the S-matrix locus for current values

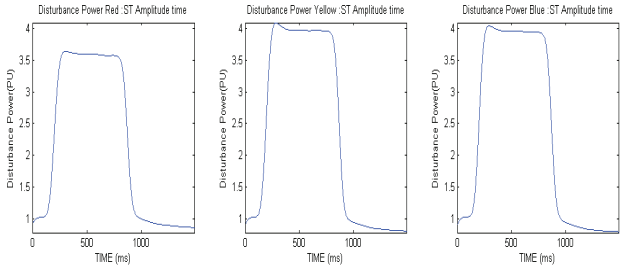


Fig. 15. Plots of the S-matrix locus for disturbance powers (downstream sag source)

When faults occur either at the transmission system or distribution system, they typically draw energy from the power system. Energy is just the value of the power flow multiply with the specified duration. In Figures 14 and 15, the plots of both the current and

disturbance power showed increase of the maximum values for both current and disturbance power during the disturbance. The increase of the maximum value plots for disturbance power during the disturbance will show that the origin of the disturbance is downstream.

If voltage sag originates from upstream, the change in the current and power profile will show either reduction or minor increase. Example on plots for voltages, currents and disturbance powers due to upstream sag sources are shown in Figures 16, 17 and 18.

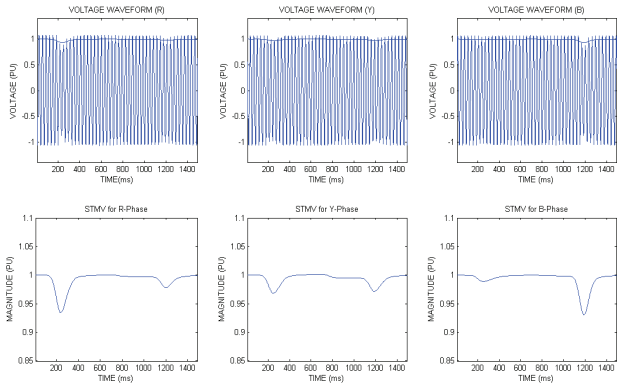


Fig. 16. S-matrix locus for voltages (upstream sag source)

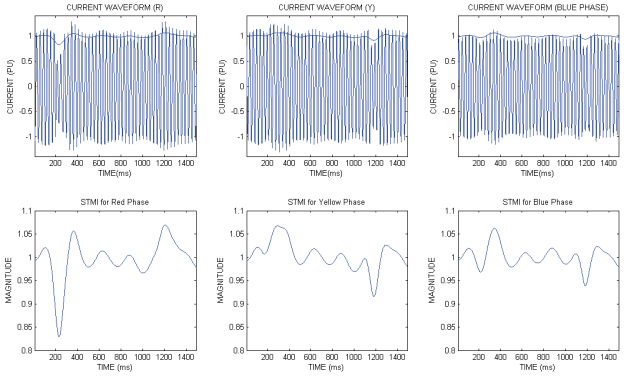


Fig. 17. S-matrix locus for currents (upstream sag source)

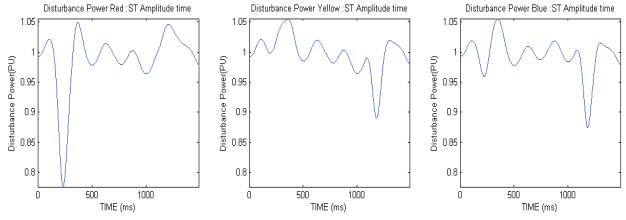


Fig. 18. Plots of the S-matrix locus for disturbance powers (upstream sag source)

The S-transform disturbance power plots are used to indicate the origin of the disturbances from the monitoring point. If during a disturbance, the power flow shows increase in power, then the source of the disturbance is downstream whereas if the power flow indicates decrease in power, then the source of the disturbance is upstream from the monitoring point.

8. Conclusion

The application and performance of CWT and S-transform for detection of multiple PQ disturbances and incipient faults and prediction of voltage sag sources which may originate from upstream or downstream have been presented. From the analyses done to evaluate the effectiveness of using CWT with the various mother wavelet functions in the detection of voltage sags, it is confirmed that the Gauss wavelet function gives the most accurate detection of voltage sag. A novel approach for detecting multiple power quality disturbances and incipient faults using the S-transform techniques has also been presented. The S-transform is proven to be very effective in analyzing and detecting multiple voltage disturbances and incipient faults in underground cable system. The S-transform accurately detects minor voltage and current transients generated from defects in underground cables. The usefulness of the S-transform is further illustrated by applying it for predicting the source of voltage sags. The numerical results obtained with actual power quality data recorded in a power distribution system indicated that the S-transform is effective in locating the sources of voltage sags which originate from either upstream or downstream. The results proved that the S-transform technique has the potential for use in the existing Power Quality Monitoring System for performing diagnosis of real-time power quality measurement data.

9. References

Bracewell, R. N. (2000). *The Fourier Transform and its Applications*. McGraw- Hill Book Co, Singapore.

Chassaing, R. (2005). *Digital Signal Processing and Applications with the C6713 and C6416 DSK*. John Wiley & Sons. Inc., Hoboken, New Jersey.

Chilukuri, M. V. & Dash, P. K. (2004). Multiresolution S-transform based fuzzy recognition system for power quality events, *IEEE Transactions on Power Delivery* 19(1), pp. 323 – 330.

Daubechies, I. (1990). The wavelet transform, time-frequency localization and signal analysis. *IEEE Transactions Information Theory* 36(5),pp. 961-1005.

Faisal, M. F. & Mohamed, A. (2009). Comparing the performance of various mother wavelet functions in the detection of voltage sags, 20th International Conference on Electricity Distribution, Prague, 8-11 June 2009.

Faisal, M. F. ; Mohamed, A.; Hussain, A. (2009). S-transform based support vector regression for identification of incipient faults and voltage disturbances in power distribution networks, Proceedings of the 11th WSEAS Int Conf on Mathematical Methods, Computational Techniques and Intelligent Systems, 1-3 July 2009, Tenerife, Spain,pp. 139-145.

Faisal, M. F. ; Mohamed, A.; (2009). A New Technique to Predict the Sources of Voltage Sags Using Support Vector Regression Based S-Transform, IASTED International Conference Modelling, Simulation and Identification, Beijing, China, 12-14 Oct 2009.

Jaya, B. R.; Dusmanta K. & Karan B. M. (2004). Power system disturbance recognition using wavelet and S-transform techniques. *International Journal of Emerging Electric Power Systems,* 1(2). Article 1007.

Karami, M.; Mokhtari, H. & Iravani M. R. (2000). Wavelet based on-line disturbance detection for power quality applications. *IEEE Transactions on Power Delivery* 15(4), pp. 1212-1220.

Langton, C. (2002). Discrete Fourier Transform (DFT) and the FFT. Signal Processing & Simulation Newsletter. http://www.complextoreal.com/fft3.htm [01 January 2008].

Lee I. W. C. & Dash P. K. (2003). S-Transform-based intelligent system for classification of power quality disturbance signals. *IEEE Transactions on Power Delivery* 18(2), pp. 800-805.

Moussa, A.; El-Gammal, M.; Abdallah, E.N.; & El-SLoud, A.A. (2004). Hardware - software structure for on-line power quality assessment. Proceedings of the 2004 ASME/IEEE Joint, pp. 147 – 152.

Poularikas. (2000). *The Transforms and Applications Handbook,* 2nd Ed. CCR press LLC, New York.

Polajžer, B. ; Štumberger, G. ; Seme, S. ; Dolinar, D. (2007). Impact of asymmetrical disturbance events on voltage sag source detection', International Conference on Renewable Energies and Power Quality (ICREPQ'07), http://www.icrepq.com/icrepq07-papers.htm, Paper No.227

Stockwell, R.G.; Mansinha, L.; & Lowe R. P. (1996). Localization of the complex spectrum: the S-Transform. *IEEE Transactions on Signal Processing* 44(4),pp. 998-1001.

Walker J. S. (1996). *Fast Fourier Transforms*, 2nd Ed, CRC-Press, New York.

2

Voltage Sags and Equipment Sensitivity: A Practical Investigation

Hussain Shareef, Azah Mohamed and Nazri Marzuki
Universiti Kebangsaan Malaysia
Malaysia

1. Introduction

In recent years, interruption of manufacturing processes due to power quality degradation has become a major focal point for many power utilities. The most prominent power quality issue plaguing utility customers is voltage sag or dip. It is a sudden decrease in voltage amplitude followed by a return to its initial level after a short time.

The use of automation and energy efficient equipment with electronic control would greatly improve industrial production. However, since these new devices are more sensitive to supply voltage deviations, characteristics of the power system that were previously ignored are now becoming a nuisance. To evaluate the technical aspects and economic issues related to voltage sags, the process and equipment immunity level has to be known. However, there is little available information related to equipment sensitivity due to voltage sags.

Studies assessing sensitivity of voltage sags on customer loads are divided into practical and theoretical approaches. The practical approaches investigate the effects of voltage sag by monitoring and conducting experiments on customers' sensitive loads, as well as by performing pertinent surveys (Bollen, 2000). Equipment sensitivity to voltage sag can also be considered and presented in the form of power acceptability curves. These curves are plots of bus voltage deviation versus time duration which separate the bus voltage deviation - time duration plane into two regions namely, "acceptable" and "unacceptable" regions. The lower limb of the power acceptability curve relates to voltage sags and momentary outages. The latest power acceptability standards are the SEMI F47 issued by the Semiconductor Equipment and Materials International (SEMI) in the year 2000 (Djokic et al., 2005) and ITIC curve of the Information Technology Industry Council (ITIC) (Kyei et al., 2002). The SEMI F47 specification simply states that semiconductor processing, metrology, and automated test equipment must be designed and built to confirm to the voltage sag ride-through capability as per the defined curve. Equipment must continue to operate without interruption during conditions identified in the area above the defined acceptable region (Institute of Electrical and Electronics Engineers Inc, 2005).

As an effort to understand the voltage immunity level of sensitive equipment, some works have been reported in the past. The categories of sensitive equipment commonly evaluated for voltage sags are personal computers (PCs) that control the on line and off line processes,

lighting systems, and ac contactors that are usually used to control motors and other industrial machineries.

The sensitivity of PCs to voltage sags is addressed in several references in the past. Seven PCs of different ages were investigated for voltage sags (Pohjanheimo & Lehtonen, 2002). The malfunction criterion for the PCs selected was automatic reboot. The authors reported that the PCs tolerate the under voltage level up to 50-60 % of remaining voltage for 100 ms. However, there was no clear correlation between the device age and sensitivity observed. Test results on standard restart/reboot malfunction criterion for computers due to voltage sags can also be found in (Saksena et al., 2005). It was reported that if the depth of voltage sag is larger than 30% and lasts more than 8 cycles, the voltage sag may cause a computer to restart. These tests were only carried out for the 120V/ 60 Hz systems. Similar experiments were conducted by Bok et al. (2008) to identify the effect of rectangular and non-rectangular voltage sags on the same restart/reboot malfunction criteria. It was noted that rectangular sags with loading condition influence most on the susceptibility of PCs. Another comprehensive study on the behavior of PCs during voltage sags and short interruptions was presented in (Djokic et al., 2005). Laboratory experiments were performed with rectangular voltage sags as well as with non-rectangular sags to simulate the starting of the large motors. Results show that all the voltage tolerance curves for different computers have the same rectangular shape with two clearly distinctive vertical and horizontal parts, with a very sharp "knee" between them. In references (Shareef et al., 2009a; Shareef et al., 2009b), the authors conducted laboratory experiments to answer why almost all the PCs have rectangular shaped voltage immunity curves with the flat vertical and horizontal part with a sharp knee between them and developed generic voltage tolerance curves for PCs. Most of these studies also declare that the PC test results can also somewhat extend to microprocessor/ CPU based devices.

Like test findings about PCs during voltage sags, there are published information that gives details on sensitivity of light flicker for different types of lamps. Experiments conducted on most common two categories of lightning loads namely fluorescent lamps (FLs) and helium lamps can be found in (Saksena et al., 2005). For both the fluorescent and helium lamps, it was concluded that the reduction in the intensity of the lamp depends only on sag depth. However, this conclusion was made on the basis of visual inspection. It was also reported that for sag depth of 60%, and 2 cycles, the fluorescent lamps start to switch off but no tests caused helium lamps to malfunction. These tests were conducted only for 120V/60 Hz system. The effects of voltage sags on several 150-W high pressure sodium (HPS) lamps combined with two different types of electronic ballast have been studied by another team of researchers (Díaz et al., 2007). It was notified that the two electronic HPS ballasts allowed the lamp to ride through for at least one cycle of power loss unlike the lamp with electromagnetic ballast. The best immunity level was found to be 57% of nominal rms voltage. Different types of gas discharge lamp namely mercury, HPS and metal halide rated from 70 to 250 watts were exposed to voltage sags by Pohjanheimo and Lehtonen (2002). The study concluded that the mercury and HPS lamps are less sensitive to voltage sags than the metal halide ones. Extensive laboratory tests with FLs having two different types of ballasts were carried out out by Shareef et al. (2009c) to observe the light intensity variation of the FLs during voltage sags and the researchers implemented a method to improve the sensitivity of the electronically ballasted FLs to voltage sags.

Similar to the other categories of sensitive equipment, ac contactors are also susceptible to power system disturbances such as voltage sags. It can disconnect circuits and cause expensive shutdown in industrial processes. Therefore some research work has been conduced to predict the behaviour of ac contactors during voltage disturbances (Pohjanheimo & Lehtonen, 2002), (Djokic´ et al., 2004), (Hasmaini & Khalid, 2004). The experiment described in (Pohjanheimo & Lehtonen, 2002) was performed to show the impact of the point on wave, sag duration, and sag amplitude on the performance of ac contactors. It was shown that voltage sag with a specific magnitude and duration can have different effects on a contactor depending on the point on wave where it originates. Additional research was conducted by Djokic´ et al. (2004) to observe the effect of phase shift during the sag, two-stage sags and sags due to the starting of large motors. It was reported that the threshold voltage that affect tripping does not have a big impact on phase shifts and for two stage sags. Digital simulation procedure for obtaining the contactor susceptibility can be found in (Hasmaini & Khalid, 2004). It concludes that the contactor disengagement initiates for voltage sags that last for 50 ms with 47% remaining voltage. But this result does not very well agree with other aforementioned research findings.
This chapter focuses on investigating the vulnerability of sensitive loads to voltage sags in the 240V/50Hz distribution system. Extensive laboratory tests are conducted for this purpose by analyzing the operation of different equipment during various events of voltage sag. From the analysis of the test findings, it also explains some of the parameters affecting the sensitivity of the tested equipment.

2. Sensitive Equipment

It is not practical to test all sag sensitive devices available in industrial or commercial facilities. Testing an adequate number of devices representing one component category is sufficient to justify the generalization of the acquired results. For this reason, the most sensitive equipment such as PCs, FLs and ac contactors are selected for testing. Functional overviews of these devices are given in the next sub sections.

2.1 Personal computers (PCs)

Personal computers first appeared in the late 1970s. It is a complex electronic computing device designed to be powered by a switch mode power supply (SMPS) which converts incoming single phase ac line voltage into a dc voltage that feeds the electronics components (Fujita & Akagi, 1999). A SMPS can be a fairly complicated circuit with stages such as rectification, filtration, conversion, and protections, as can be seen from the block diagram shown in Fig. 1.
In the first stage, a diode bridge rectifies the incoming voltage. A large capacitor then filters the pulsating dc voltage to create a nearly constant dc voltage. However, under normal operating conditions, over a half-cycle, the capacitor voltage decays to some value. Depending upon the minimum voltage value set by the design of the SMPS, the dc-dc converter in the conversion stage will deliver rated dc output voltage until the capacitor voltage reaches the designed minimum value. The time to reach this voltage at rated load is defined as the holdup time, T_h, which is represented mathematically as (Fernandez et al., 2005):

$$T_h = \frac{C_{dc}(V_{norm}^2 - V_{min}^2)}{2P} \quad (1)$$

where
C_{dc} is the capacitance of the filter capacitor.
V_{norm} is the peak nominal voltage.
V_{min} is the peak minimum voltage set by the SMPS design.
P is the rated power of the SMPS.

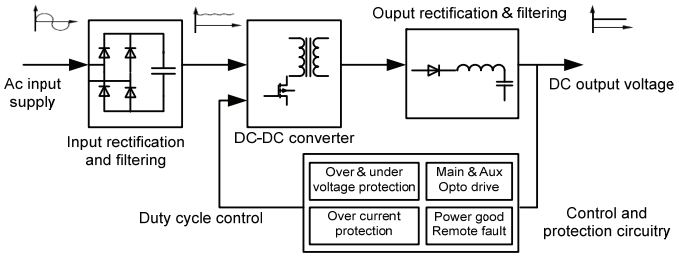


Fig. 1. Block diagram of a SMPS used in a PC

Another component related to sensitivity of PCs is the protection circuit of SMPS. These circuits monitor critical circuit conditions and report any violations of prescribed limits. Protection circuit provides over voltage and over current protection for 3.3V, ±5V and ±12V, generates power good logic output signal, programmable timing for power good signal, stable internal voltage reference and voltage reference for main and auxiliary regulation. In addition, there is a special under voltage detection input for sensing the input voltage to the power supply. This input causes the power good signal to toggle if there is insufficient voltage to run the power supply unit outputs. A high power good logic output indicates that the power from the mains is good for PC operation.

2.2 Fluorescent Lamps (FLs)

The operating principle of FLs is the same whether the form is a straight tube, circular, or convoluted as in compact fixtures. When a voltage is applied across the ends of a sealed glass tube containing mercury vapour, it causes the vapour to ionize. This vapour radiates light in the ultra violet region of the spectrum, which is converted to visible light by a fluorescent coating on the inside of the lamp. However, it requires a high voltage pulse across the tube to start the process and some form of limiter to prevent the current increasing to a level where the lamp can be destroyed. The current limiter is commonly known as ballast.
The traditional ballast contains an inductor connected in series with the lamp, and a starter (Vitanza et al., 1999). The starter triggers the tube when it is first turned on, by easing the current flow through the inductor and the filaments of the tube in the first place. When the starter bimetal strip reopens, the high circuit impedance and consequent sharp reduction in inductor current causes enough overvoltage to ionize the gas in the tube. However, this solution has significant weaknesses which include a high power loss in the inductor core, light flickering, and a very low power factor due to high inductive reactance.

Electronic ballasts replace the starting and bulk inductive elements of the conventional electromagnetic ballasts. The electronic ballast improves the performance of the lamp by operating at a higher frequency above the 50Hz determined by the mains supply. This eliminates lamp flickering because the gas in the tube does not have time to de-ionize between current cycles which also leads to lower power consumption, and longer tube life. Moreover, since the inductor required to ionize the tube is smaller, resistive loss and the system size is reduced (Vitanza et al., 1999).
Fig. 2 depicts a block diagram of an electronic ballast. The first block contains the protection, filtering, and current peak limiting components. Block 2 is the full diode bridge rectifier to convert the ac line into a dc stage. Block 3 is the smoothing capacitor. It provides the dc link voltage of the resonant inverter for the tube in Block 4. The resonant inverter normally runs at 10-40 kHz. The most commonly used resonant inverter circuits for low-wattage FLs are voltage fed half-bridge quasi-resonant circuits and current fed half-bridge resonant circuits (Vitanza et al., 1999).

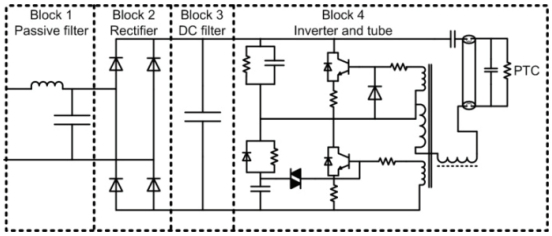


Fig. 2. Block diagram of an electronic ballast

2.3 AC contactors

An ac contactor utilizes a solenoid to cause one or more pairs of electrical contacts to engage when an appropriate voltage is applied to solenoid's coil as shown in Fig. 3. The solenoid consists of an electromagnet that attracts a moveable bar. The moveable bar is spring loaded so as to cause the bar to move away from the electromagnet when the electrical signal is not present on the coil. Electrical contacts are attached to the moveable bar and the movement causes the contacts to close or open depending on the strength of the magnetic field. The instantaneous flux, φ, and the force, F, that tends to close the air gap in the contactor can be expressed respectively as (Hasmaini & Khalid, 2004):

$$\phi = \frac{NI}{l/\mu A}\cos(\omega t) \tag{2}$$

$$F = \frac{\phi^2}{2\mu_0 A} \tag{3}$$

where
N = number of winding in the coil.
I = current flow through the coil.
l = the length of magnetic path.
μ_0 = absolute permeability.
μ = permeability of the coil substances.
A = the cross sectional area of the air gap.
ω = steady state frequency in radians.

Then, assuming the coil self inductance is constant and dominant, the minimum voltage, V_{hold}, required to keep the contactor from dropping out is given as (International Electrotechnical Commission, 2009):

$$V_{hold} = \frac{N\omega\phi}{\sqrt{2}} \tag{4}$$

According to (4) the hold in voltage, V_{hold}, depends very much on the flux which is directly proportional to the instantaneous current applied to the coil. Since the solenoid coil is assumed to be a pure inductor, the phase difference between the coil's current and voltage have a dramatic effect on the point on wave of the voltage sag event. For instance, a voltage sag initiates at the peak of the voltage waveform may leave a very small amount of current to produce enough force to hold the contacts while an event at zero crossing of voltage waveform will leave a much higher current to hold the contacts.

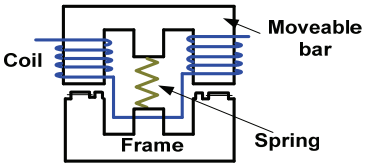


Fig. 3. Basic structure of an ac contactor

3. Methods and Materials

This section illustrates the design of the experiment for equipment testing and the procedures followed to obtain the results on the performance of the equipment during voltage sags.

3.1 PC testing

The methodology that is used in the testing is generally based on the guideline given in International Electrotechnical Commission (1994). Five PCs with different specifications are tested to study the effect of voltage sags on the performance of the computers. The specifications of the tested PCs are shown in Table 1. The specifications of the test PCs listed in Table1are assumed to cover some old and new models of PCs that are commonly in use at the time of the experiments.

The experimental set up consists of four components namely, sag generator, equipment under test (EUT), data acquisition system, and a computer to analyze the signals. In this case, an industrial power corruptor (IPC) from the Power Standards Lab is used, which is a voltage sag generator combined with built-in data acquisition system that is capable of producing and interrupting voltages up to 480V and current at 50A in single or three phase systems.

A series of test results on PCs are obtained by following the pre-defined procedure given below.

I. Using the terminal blocks available at the back of IPC, the conductors from mains panel and conductors to the PC under test are connected and the IPC is powered on.
II. The PC with all input/output (I/O) and pointing devices connected is switched on, allowing it to boot and load the operating system.
III. Allow Disk Defragmenter program to scan and defragment system discs.
IV. Starting from nominal voltage, voltage sags are initiated in steps of 2.5% down to zero volts. The sag initiation angle and the duration are kept constant. The initial sag duration and phase angle are set to 1 cycle and 0° respectively. The critical sag depth for the pre-defined malfunction criteria is determined by repeated testing for at least 3 times for a particular sag magnitude and duration. If reboot malfunction condition is observed, a quick inspection for proper operation of PC under test is conducted before initiating the next sag. For each triggered sag event, different voltage and current waveforms supplying and controlling the PC under test are recorded. Observations such as visible or audible influence on the PC are also noted.
V. The duration of sag is adjusted in steps of 1 cycle and measurements outlined in Step 4 are repeated.

PC no.	Specifications
PC1	CPU: Pentium III, 450MHz Processor, 128 MB SDRAM
	OS: Windows Me
	Power Supply: 100-127V/200-240V, 5/2.5A, 60/50Hz, 145W, Model: PS-5141-2D1
PC2	CPU: Pentium 4, 2.40GHz Processor, 261 MB RAM
	OS: Windows 2000 Professional
	Power Supply: 100-120V/200-240V, 5/3A, 60/50Hz, 180W, Model: NPS-180BBA
PC3	CPU: Core 2 Duo, 2.00GHz Processor, 1 GB RAM
	OS: Windows XP Professional
	Power Supply: 100-127V/200-240V, 8/4A, 60/50Hz, 250W, Model: PS-5251-08T
PC4	CPU: Pentium III, 933 MHz Processor, 256 MB RAM
	OS: Windows 2000 Professional
	Power Supply: 100-127V/200-240V, 9/4.5A, 60/50Hz, 300W, Model: SA-320-35005
PC5	CPU: Pentium 4, CPU 1.90GHz, 504 MB RAM
	OS: Windows XP Home Edition
	Power Supply: 100-127V/200-240V, 6/3A, 60/50Hz, 250W, Model: ATX-480W

Table 1. Specifications of tested PCs

3.2 FL testing

Many FLs with different ballast types are tested to study the effect of voltage sags on the performance of the lamps. The specifications of these FLs are shown in Table 2. The selected lamps are commonly found in residential and commercial applications. Since the main objective of lamp testing is to detect and determine light output variations of the FLs during voltage sags, it is important that the design of the test system must be fast enough to capture the light intensity variation of the test lamps accurately. Therefore, in addition to the materials used for PC testing an advanced photometer which is fast enough to capture the light intensity variation of the test lamps during sag events is used.

FL no.	Ballast type	Power rating	Lamp Type
FL1	Electronic	8W	CFL, Convoluted
FL2	Electronic	8W	CFL, Convoluted
FL3	Electronic	8W	CFL, Convoluted
FL4	Electronic	8W	CFL, Convoluted
FL5	Electronic	14W	CFL, Convoluted
FL6	Electromagnetic	18 W	Straight tube
FL7	Electronic	18 W	Straight tube
FL8	Electronic	18 W	CFL, Convoluted
FL9	Electronic	32 W	CFL, Convoluted
FL10	Electromagnetic	36 W	Straight tube

Table 2. Specifications of tested FLs

The test system shown in Fig. 4 has been built to perform the voltage sag disturbances and evaluate the resultant light output levels from the lighting source. The lamp under test is enclosed in a prefabricated lighting chamber which eliminates stray light and reduces reflections by its internal matt black surface. This point source method measures the light directly produced by the lamp with a light detector at the opposite end of the chamber. The detector head photocurrent is converted to a voltage and it is more than capable of detecting flicker in the human visible range of 0-35Hz (Frater & Watson, 2007). The conversion process of light detector current into an appropriate level of voltage is performed by the processor in the photometer. However, since the photometer does not have its own built in data acquisition system, the converted voltage waveform is therefore fed to the data acquisition system channels available in the IPC for post processing and analysis.

Similar to the series of tests conducted on PCs, test results for FLs are obtained for predefined malfunction conditions known as zero illuminance condition. At transition points where zero illuminance condition is start to observe, the procedure is repeated for at least three times to avoid probable errors that may occur during the experiments.

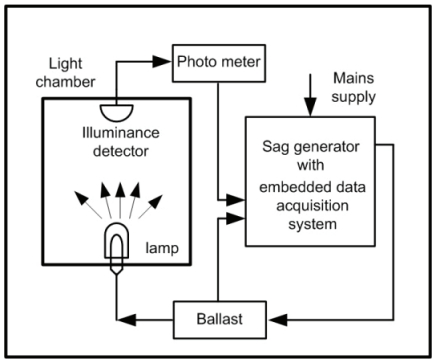


Fig. 4. FL test setup

3.3 AC contactor testing

In order to find the impact of voltage sags on ac contactors, different contactors listed in Table 3 are tested. All contactors are tested with a 2000 Watt spot light load attached to their main electrical contacts. First the contactor is warmed up to its normal operating temperature by applying nominal coil voltage for a couple of minutes before initiating the sag event. In the case of ac contactor testing, the malfunction condition was defined as the disengagement of the main contacts. The disengagement of contacts can be guaranteed with the test setup shown in Fig. 5 where one of the normally open contacts is used to energize the ac coil of the contactor.

Contactor no.	Manufacturer/ Model	Current rating
C1	LG Industrial System / GMC-18	18 A
C2	LG Industrial System / GMC-40	40 A
C2	FUJI / SC-N2 S	50 A

Table 3. Specifications of tested AC contactors

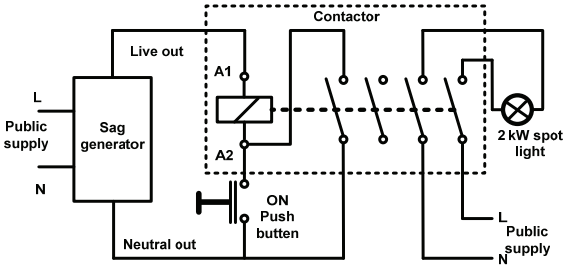


Fig. 5. AC contactor test setup

Here again, the test procedure follows the same basic steps illustrated in PC testing where the main variables are sag depth and duration. However, since point on wave of sag initiation affects contactor performance, point on wave of sag is also added as a test variable

in this case. Sag initiation angle was varied in steps of 15° at a specific sag magnitude and duration.

4. Results and Analysis

The test findings of different equipment to voltage sags are initially presented as typical voltage tolerance curves. The upper region of these curves represents proper operation region while the lower region indicates unacceptable voltage conditions for equipment operation. Based on the findings, a generic voltage tolerance curve for each equipment category is then constructed.

4.1 Analysis of PCs' voltage tolerance level

Effect of voltage sag on all the tested PCs is shown in Fig. 6 along with the standard SEMI F47 and ITIC voltage acceptability curve. Like in previous research findings on sensitivity of PCs to voltage sags, the obtained curves have the same rectangular shape with two clearly distinctive vertical and horizontal parts, with a very sharp "knee" between them. From Fig. 6, it can be seen that for PC1 to PC5, the knee points are 47.5% - 14 cycles, 25% - 8 cycles, 40% - 12cycles, 50% - 11 cycles and 45% - 14 cycles, respectively. If one compares these individual voltage tolerance curves, it can be observed that PC4 is the most sensitive PC to voltage magnitude while PC2 is the least. When the sensitivity of the PCs in terms of duration is considered, PC2 starts to malfunction at 8 cycles. One final observation that can be obtained from Fig. 6 is that all tested PCs can ride through indefinitely if the magnitude of the sag is less than 50 % nominal voltage and satisfy the design goals of SEMI F47 and ITIC standard.

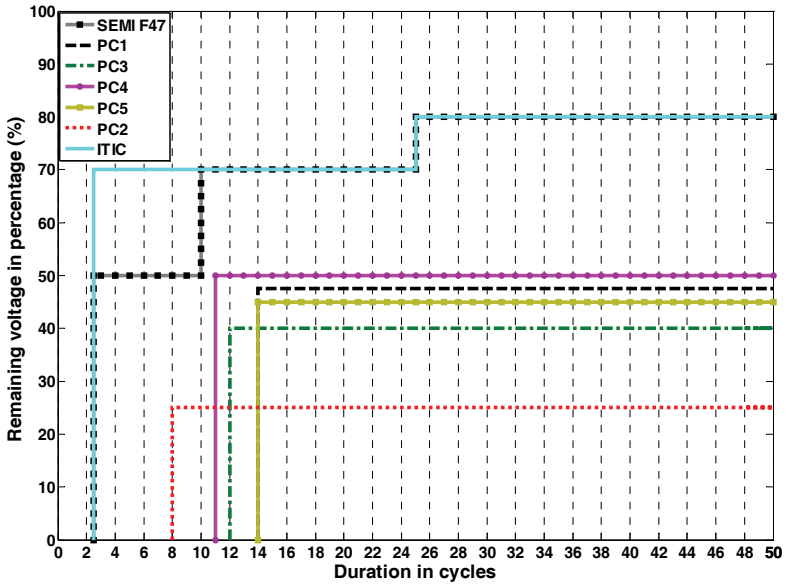


Fig. 6. Voltage tolerance curves of various PCs

As illustrated in Fig. 6, each personal computer potentially has its own standard of power acceptability. An approach to define the overall acceptability region is to apply intersection to the individual voltage tolerance curves (Kyei et al., 2002) as shown in Fig. 7.

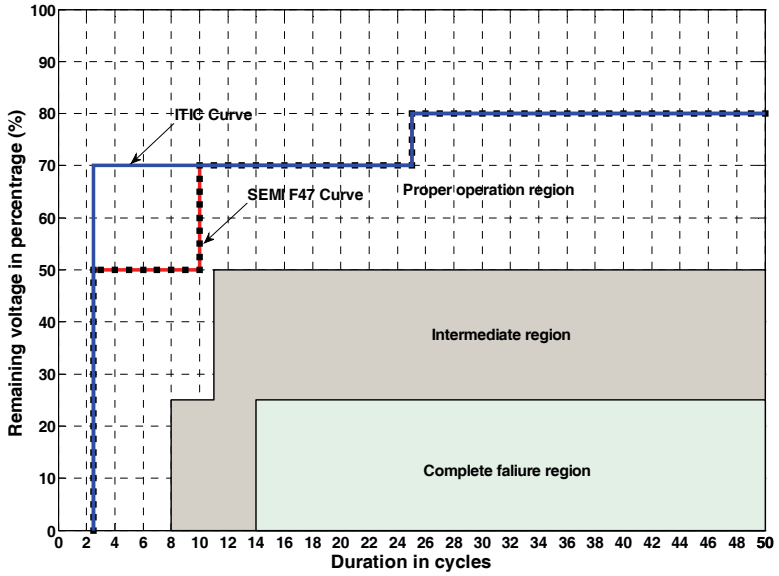


Fig. 7. Generic voltage tolerance curves of various PCs

In Fig. 7, the upper acceptable region is the region that all PC loads properly operate, the lower region indicates that all PCs fail, and the intermediate region corresponds to some PC failures and some 'ride-throughs'.

To further understand the reason why almost all the PCs have rectangular shaped voltage sensitivity curves, signals obtained at different points of the SMPSs are analyzed. Figs. 8 and 9 illustrate the waveforms obtained at the rectifier dc output and power good logic output of the PC4 SMPS during different magnitude of voltage sags. From Fig. 8, it can be observed that by varying sag depth from 52.5% to 30 % remaining voltage for 10 cycles, the voltage decay at the rectifier dc output remains almost unchanged even for very deep sags. At 10 cycles, the energy stored in the dc link capacitor does not allow the rectifier dc output to decrease up to its minimum voltage as defined in (1). For this reason, the under voltage detection section of the housekeeping circuit does not toggle the power good signal as shown in Fig. 8. This indicates that PC4 will continue to operate normally for 10 cycles even if there is no mains supply for 200 ms.

Fig. 9 shows the variation of the rectifier dc output voltage and power good signal where PC4 starts to malfunction at 11 cycle. Since the sag duration is 20 ms longer at 11 cycles, the rectifier dc output voltage decays further. From Fig. 9, it is clear that deeper sags cause the power good signal to toggle and PC4 to reboot. This is due to the fact that the deeper sags starting from 50% remaining voltage cause the rectifier dc output to fall below the set minimum voltage of 154 Volts for PC4. Almost similar waveforms are obtained for the series of tests conducted on other PCs. In order to observe the effect of voltage sag duration,

waveforms at the rectifier dc link are also observed for constant sag magnitude. It was noted that the dc link voltage decreases at a constant rate no matter what the sag duration is. So voltage sag duration does not have an effect on the time to reach the set minimum voltage of the SMPS design.

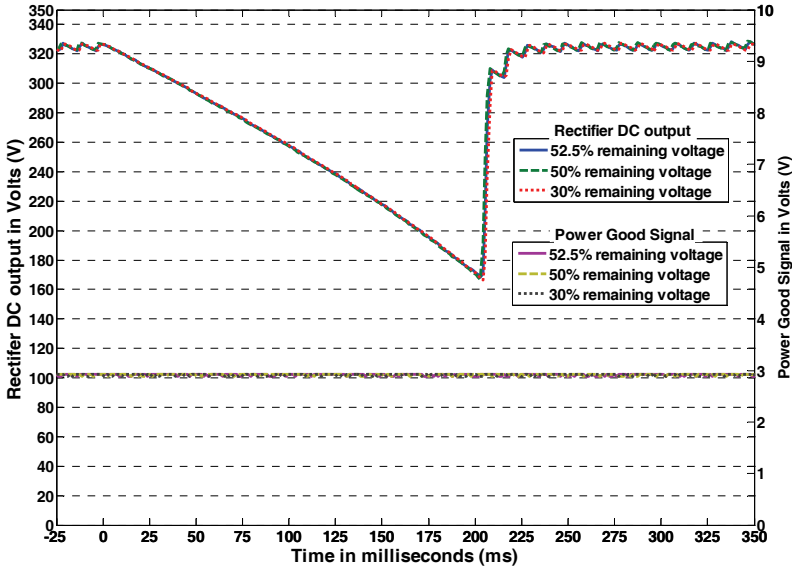


Fig. 8. Effect of sag depth on the rectifier dc output at 10 cycles for PC4

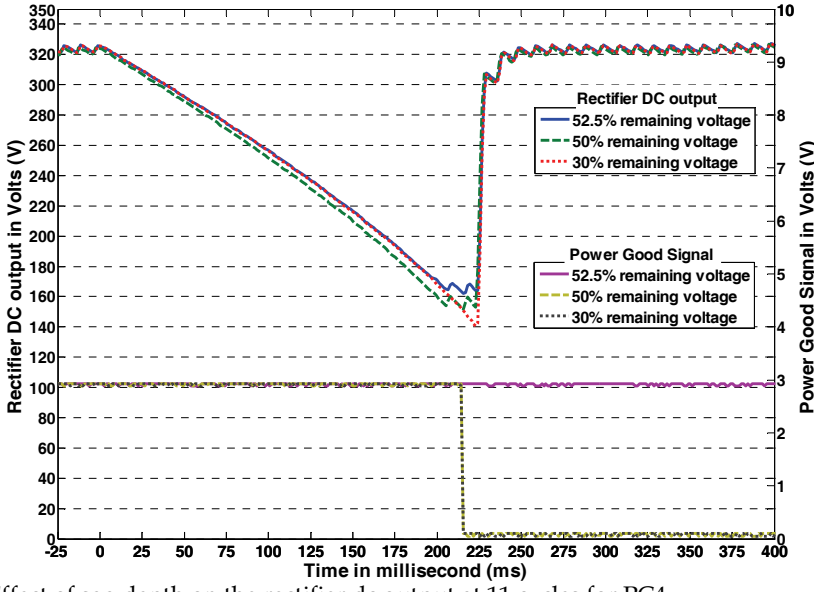


Fig. 9. Effect of sag depth on the rectifier dc output at 11 cycles for PC4

From the extensive tests and analysis, it can be concluded that the standard reboot/restart malfunction in the event of a voltage sag depends upon the energy stored in the dc link capacitor and the minimum voltage for which PC is designed to trigger the under voltage protection circuit embedded in the SMPS of the PC. Moreover, studies conducted to analyze the effect of sag depths and duration help to understand the rectangular nature of voltage tolerance curves of PCs.

4.2 Analysis of FLs' voltage tolerance level

Numerous test results of FLs are analyzed in this section. It is done by investigating the signals obtained from the photo sensor, lamp current, supply voltage and current. In the case of FLs with electronic ballast, the voltage variation at the dc bus which feeds the resonant inverter circuit shown in Fig. 2 is also investigated.

The overall immunity level of all 10 FLs to voltage sags are presented in Fig. 10 as typical voltage tolerance curves along with the SEMI F47 and ITIC standard. The upper region of these curves represents proper operation region while the lower region indicates zero illuminance conditions for FLs' operation. From Fig. 10, the FLs with electromagnetic ballasts are found to be the most sensitive lamps for short duration sag events. The lamp turn off condition for FLs with electromagnetic ballast generally initiated for voltage sags as short as 1 cycle. CFLs and conventional FLs with electronic ballasts are also sensitive to voltage sag. The main difference in the case of electronically ballasted FLs is that it is a little more immune to sags in terms of duration. However they are generally more sensitive of voltage sag magnitude as shown in Fig 10. FL7 with electronic ballast happened to be the most immune lamp to voltage sags. It is found to malfunction for sag magnitude beginning from 5% and for all durations greater than 5 cycles as shown in Fig. 10.

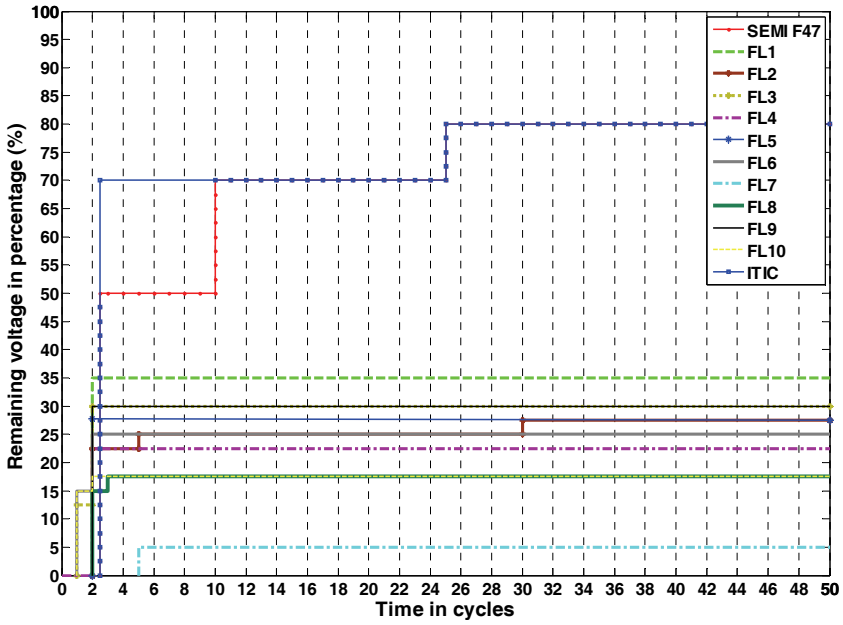


Fig. 10. Voltage tolerance curves of various FLs

Fig. 11 shows the generic voltage tolerance curve generated for FLs using intersection method. By comparing immunity curve of FLs shown in Fig. 11 it can be said that many FLs do not satisfy the design goals of SEMI F47 and ITIC standard as most of the FLs fail to deliver light for voltage sags lasting more than 2 cycles.

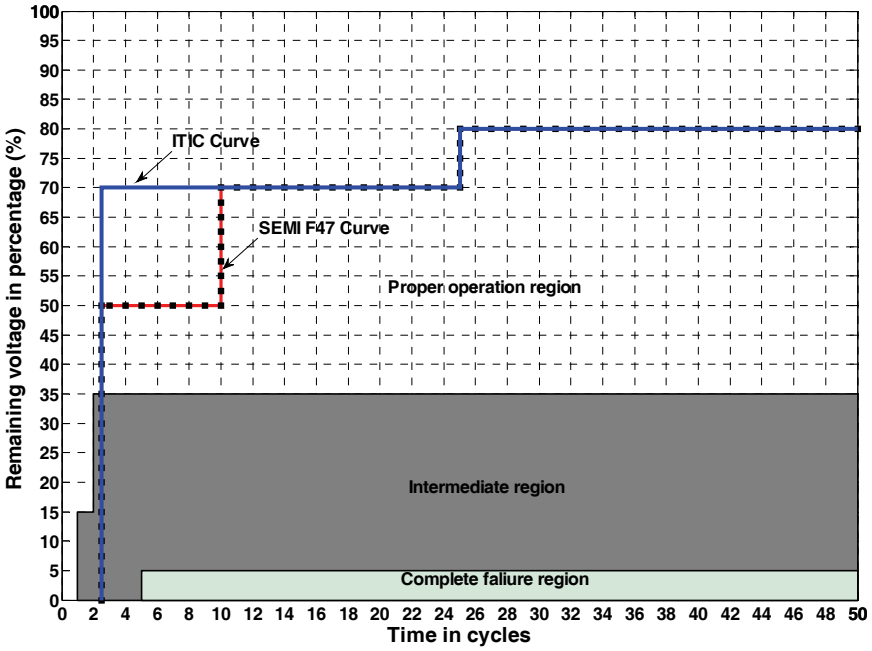


Fig. 11. Generic voltage tolerance curves of various FLs

Since the above voltage tolerances do not show how the lamp current, voltage and light output variation changes during voltage sag events, some observations obtained for FL6 and FL7 are illustrated below.

Figs. 12 and 13 illustrate the waveforms obtained from the photo sensor and the lamp current for the 18 Watt conventional FL (FL6) with electromagnetic ballast listed in Table 2, respectively. The effect of varying the sag depth starting from 25% to 15% remaining voltage for 1 cycle, on light output variation of the lamp is shown in Fig. 12. It shows very important information about the behavior of light output in conventional FL with electromagnetic ballast during voltage sag. The first information that can be derived from Fig. 12 is that the lamp turn off condition starts to occur for sag having 15% remaining voltage. At this point the lamp cannot reignite itself and requires the starter circuit to initiate ionization again. Furthermore, for different depths of voltage sags, the decay time of light output variation remains almost constant between 0 ms and 20 ms which represent the starting and end point of the sag respectively. Although FL6 starts to malfunction at 15% remaining voltage for voltage sag that last for 1 cycle, it is different for longer duration sags. For 2 cycle sags, the lamp becomes more sensitive to the depth of the sag. It starts to extinguish for all sags that is deeper than 25% remaining voltage.

Fig. 13 shows the variation of lamp current for different depths of voltage sag that last for 1 cycle. Observe that the lamp current, during all compared events of sag depth, reduce to a very low value. However, except in case of sag event that leaves 15% remaining voltage for 1 cycle, the lamp currents returns to normal as soon as the supply voltage recovers from the sag event.

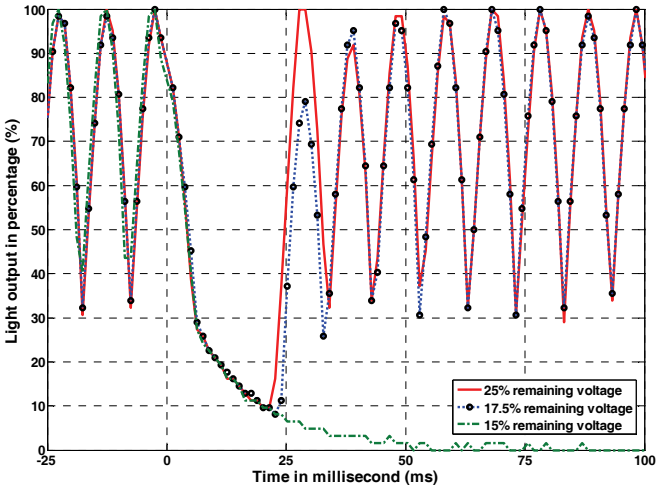


Fig. 12. Effect of sag depth on the light output at 1 cycle for FL with electromagnetic ballast

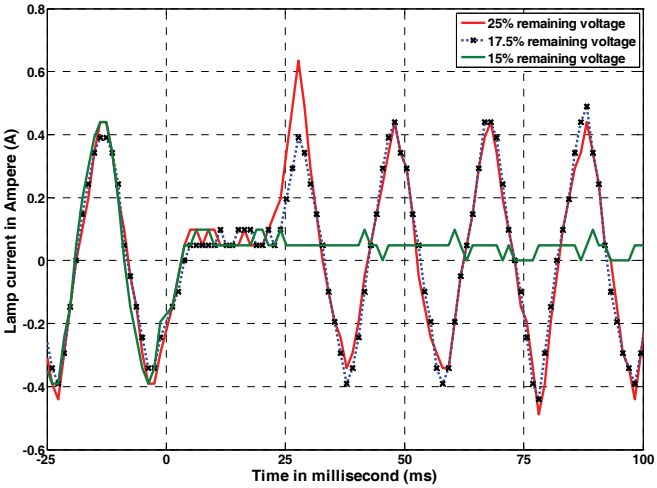


Fig. 13. Effect of sag depth on the lamp current at 1 cycle for FL with electromagnetic ballast

Similar to the FLs with electromagnetic ballasts, FLs with electronic ballasts also experienced zero illuminace condition due to voltage sag disturbances. Fig. 14 shows the variation in light output where FL7 with electronic ballast first starts to malfunction for voltage sag lasting for 5 cycles. From Fig. 14, it can be clearly seen that the FL7 is much more immune to voltage sag when compared to FL6 with electromagnetic ballast. Note from Fig.

14 that FL7 just reaches zero illuminance malfunction condition at a sag depth of 5% remaining voltage just at the end of the 5 cycles. Moreover, observe that the light output fluctuation in the steady state operation varies at a higher frequency within a narrow band of 90% to 100% of the nominal light output. This reduces the flicker effect that is obvious in the case of FL6 as seen in Fig. 12.

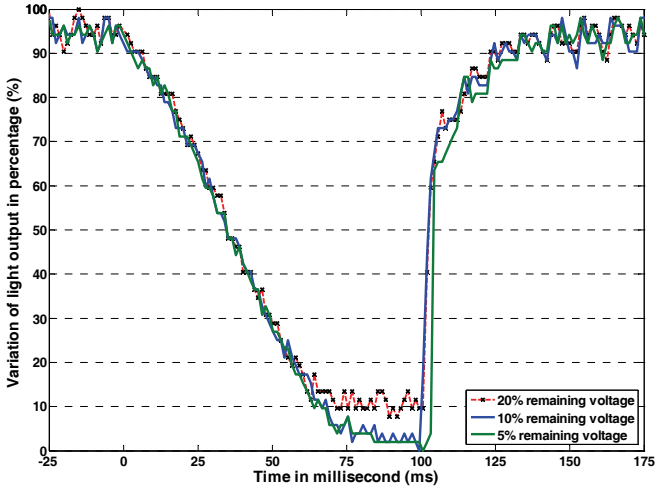


Fig. 14. Effect of sag depth on the light output at 5 cycles for FL with electronic ballast

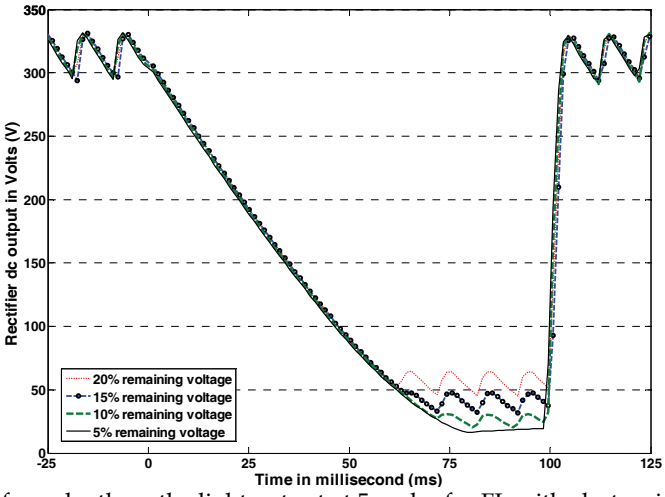


Fig. 15. Effect of sag depth on the light output at 5 cycles for FL with electronic ballast

Another way to confirm the malfunction condition of FLs that uses electronic ballast is to analyze the variations in the rectifier dc output or dc bus voltage and lamp current. These waveforms for FL7 are shown in Fig. 15 for different depths of voltage sag lasting for 5 cycles. Observe that for sag depth of 5% remaining voltage, the dc bus voltage maintains

almost at a constant voltage level just after 80 ms unlike for voltage sags that are shallower. This indicates that the lamp does not draw sufficient current for its proper operation.
To analyze the effect of variation of sag duration on the performance of FL7 at sag depth of 10% remaining voltage, Fig.16 is plotted. From Fig. 16 it can be noted that the light output variation does not drop down to zero completely even if the sag duration is varied between 3 to 6 cycles and therefore the lamp is considered to operate properly. Here again it can be observed that the decay rate of light output variation during the sag, remains almost the same. For example, all sag events cause the light output of this FL to drop up to 18% of full brightness at 60 ms as shown in Fig. 16.

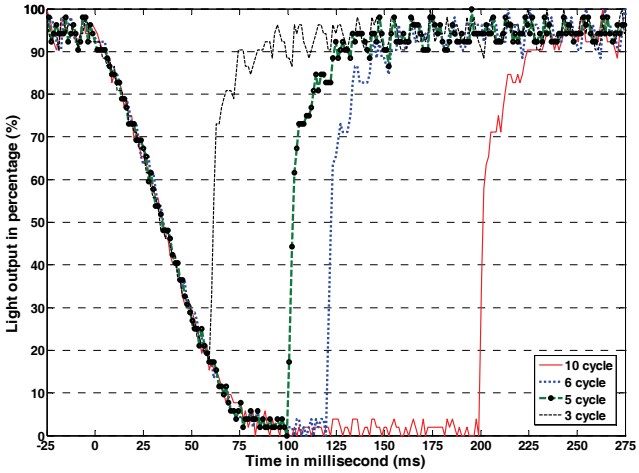


Fig. 16. Effect of sag duration on the light output at 10% remaining voltage for FL with electronic ballast

From the extensive tests and analysis, it can be concluded that the malfunction of FLs with electronic ballast, in the event of voltage sag, depends upon the energy stored in the dc link capacitor and the minimum voltage for which the ballast is designed to function properly unlike the conventional electromagnetic blasted lamps. However, this conclusion is not true for FLs with electromagnetic ballast.
Although it has not been highlighted in the FL test procedure, it is found that 0° sag initiation angle influence most on the sensitivity of electromagnetically ballasted FL compared to tests conducted to observe the effect on initial point on wave of the sag.

4.3 Analysis of AC contactors' voltage tolerance level

Test results of testing of the contactor with rectangular voltage sag do not produce a single voltage tolerance curve, but the families of curves corresponding to different point on wave initiation. Typical effect of point on wave for the contactor C1 listed in Table 1 is shown in Fig. 17. The contactor C1 tolerates for very deep sags and interruptions up to 4 cycles in cases where the sag initiation occurs at 0° or 45° on the point of voltage waveform. However, when the initiation angle is 90°, the contactor disengages for sags that last only for 1 cycle. This result also agrees with the theory highlighted in Section 2.3. For other tested contactors, the effect on point of wave shows similar behavior. Fig. 18 shows the generic

voltage tolerance curve obtained from individual immunity curves of the contactors. From this generic curve it can be seen that SEMI F47 and ITIC curves are not a suitable standards to compare the tolerance levels of ac contactors. A more suitable standard to compare the performance of contactors could be IEC standard 60947-4-1 (International Electrotechnical Commission, 2009). According to this standard the limits between which the contactors should drop out and open fully are 75% to 20% of their rated control supply voltage for ac contactors. If one compares the tolerance levels of tested contactors with IEC standard 60947-4-1, all falls within the specified limits.

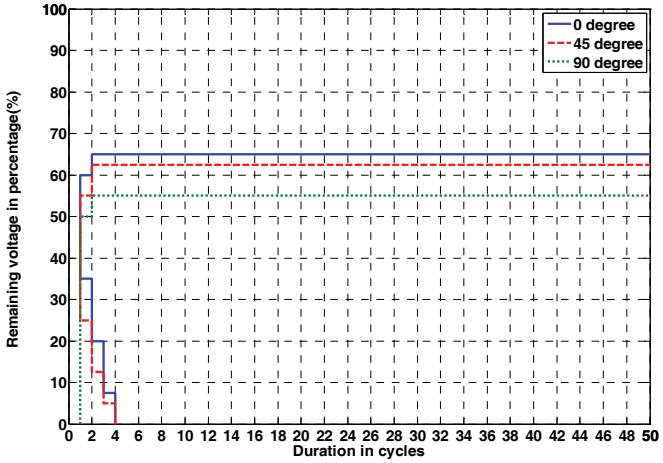


Fig. 17. Voltage tolerance curves for contactor C1

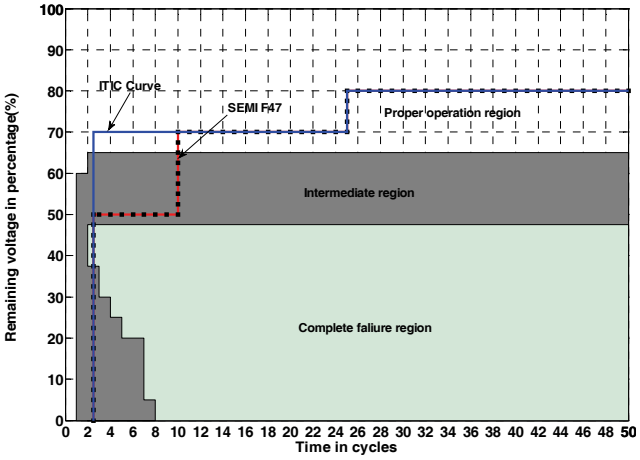


Fig. 18. Generic voltage tolerance curves for contactors

5. Conclusions

An extensive experimental study has been performed to determine the effect of voltage sag on sensitive loads such as PCs, FLs and ac contactors. Based on the experimental results it is

possible to construct a generic voltage tolerance curve for sensitive equipment which can clearly show acceptable and unacceptable regions for different voltage sag disturbances. The curve provides a quick overview about the immunity level of these devices in a particular power distribution network.

From the experimental results, it can be concluded that the voltage tolerance level of the PCs used in the tests vary over a wide range. All the voltage immunity curves obtained appear to have similar shape with distinctive vertical and horizontal parts. When the voltage immunity levels of the tested PCs are compared with the SEMI F47 and ITIC standards, all the tested PCs satisfy their design goal. By observing different waveforms at the SMPS of PCs, the reason behind the rectangular nature of PCs voltage tolerance curves is revealed. Moreover, investigation on the rectifier dc output and power good signal of SMPS shows that malfunction of a PC occurs at a specific time for a predefined minimum voltage defined by the design of the PC SMPS. Therefore, deeper and longer duration sags do not have any correlation with the initiation of PC restart malfunction condition. It only appears to rely on the hold-up time and the set minimum voltage of the SMPS.

From the results of the tested FLs, the voltage immunity level of the lamps with electromagnetic ballasts is more sensitive to voltage sags than the FLs equipped with electronic ballasts. By observing variations in the light output of the lamps, it is possible to conclude that the light intensity of the lamp not only depends on the voltage sag depth but also on the duration of the sag event depending upon the design of the ballast used in the lamp. Moreover, investigations of rectifier dc output and lamp current of FLs with electronic ballasts show the exact timing where the ballast stops functioning properly.

The test results on ac contactors clearly show that the magnitude and the duration of voltage sag are not the only parameters that influence the sensitivity of a contactor to voltage disturbances. The point on wave initiation has significant influence on the behaviour of ac coil contactors.

6. References

Bollen, M.H.J. (2000). Understanding Power Quality Problems: Voltage Sags and Interruptions, IEEE Press, ISBN 0-7803-4713-7, New York

Bok, J.; Drápela, J. & Toman, P. (2008). Personal computers immunity to short voltage dips and interruptions, Proceeding of 13th international conference on harmonics and quality of power, pp. 1-6, ISBN 978-1-4244-1771-1, Australia, September-October 2008, IEEE, Wollongong

Djokic, S.Z.; Milanovic, J.V. & Kirschen, D.S. (2004). Sensitivity of AC coil contactors to voltage sags, short interruptions, and undervoltage transients, IEEE Transaction on Power Delivery, Vol.19, No. 3, July 2004, pp1299-1307, ISSN 0885-8977

Djokic, S.Z.; Desmet, J. Vanalme, G. Milanovic, J.V. & Stockman, K. (2005). Sensitivity of personal computers to voltage sags and short interruptions, IEEE Transaction on Power Delivery, Vol.20, No.1, January 2005, pp. 375-383, ISSN 0885-8977

Díaz, F.J.; Azcondo, F.J. Ortiz, F. Ortiz, A. Mañana, M. & Renedo, C. (2007). Effects of voltage sags on different types of ballasts for 150-W HPS lamps, Proceeding of 9th international conference on power quality and utilization, pp. 1-6, ISBN 978-84-690-9441-9, Spain, October 2007, Barcelona

Fujita, H. & Akagi, H. (1999). Control and performance of a pulse-density-modulated series-resonant inverter for corona discharge processes, IEEE Transaction on Industrial Application, Vol. 35, No.3, May-June 1999, pp 621-627, ISSN 0093-9994

Fernandez, J.; Sebastian, M. Hernando, M. Villegas, P. & Garcia, J. (2005). Helpful hints to select a power-factor-correction solution for low- and medium-power single-phase power supplies, IEEE Transaction on Industrial Electronics, Vol. 52, No.1, February 2005, pp 46-55, ISSN 0278-0046

Frater, L.P. & Watson, N.R. (2007). Light flicker sensitivity of high efficiency compact fluorescent lamps, Proceedings of australasian universities power engineering conference, pp. 1-6, ISBN 978-0-646-49488-3, Australia, December 2007, IEEE, Perth

Hasmaini, M. & Khalid, M.N. (2004). Evaluation on Sensitivity of AC Contactor During Voltage Sag, Proceeding of IEEE TENCON 2004 region 10 conference, pp. 295-298, ISBN 0-7695-2378-1, Thailand, November 2004, IEEE, Chiang Mai

International Electrotechnical Commission. (1994). Electromagnetic Compatibility (EMC), Part 4: Testing and Measurement Techniques, Section 11: Voltage Dips, Short Interruptions and Voltage Variations Immunity Tests, Std. IEC 61000-4-11

International Electrotechnical Commission. (2009). Low-Voltage Switchgear and Controlgear – Part 4-1: Contactors and Motor-Starters – Electromechanical Contactors and Motor-starters, IEC 60947-4-1.

Institute of Electrical and Electronics Engineers Inc. (2005). Recommended Practice for Powering and Grounding Electronic Equipment, IEEE Press, ISBN 0-7381-4979-9 SH9551, New York

Kyei, J. Ayyanar, R. Heydt, G.T. Thallam, R. & Blevins, J. (2002). The design of power acceptability curves, IEEE Transaction on Power Delivery, Vol.17, No.3, July 2002, pp.828–833, ISSN 0885-8977

Pohjanheimo, P. & Lehtonen, M. (2002). Equipment sensitivity to voltage sags – test results for contactors, PCs and gas discharge lamps, Proceeding 10th international conference on harmonics and quality of power, pp. 559- 564, ISBN 0-7803-7671-4, Brazil, October 2002, IEEE, Rio de Janeiro

Saksena, S.; Shi, B. & Karady, G. (2005). Effects of voltage sags on household loads, Proceeding of power and energy society general meeting - conversion and delivery of electrical energy in the 21st century, pp. 2456- 2461, ISBN 0-7803-9157-8, USA, June 2005, IEEE, USA

Shareef, H.; Mohamed, A. & Marzuki, M. (2009a). Immunity level of personal computers to voltage sags in the 240 v/50 Hz distribution systems, Journal of Applied Sciences, Vol.9, No.5, January 2009, pp. 931-937, ISSN 1812-5654

Shareef, H.; Mohamed, A. & Marzuki, M. (2009b). Analysis of personal computers ride through capability during voltage sags, Electric Power Systems Research, Vol.79, No.1, December 2009, pp. 1615-1624, ISSN 0378-7796

Shareef, H.; Mohamed, A. & Marzuki, M. (2009c). Analysis of ride through capability of low-wattage fluorescent lamps during voltage sags, International Review of Electrical Engineering, Vol.4, No.5, September- October 2009, pp. 1093-1101, ISSN 1827-6660

Vitanza, A.; Scollo, R. & Hayes, A. (1999). Electronic fluorescent lamp ballast, STMicroelectronics Application Note AN527/1294

3

Applications of the Parallel-LN-FDTD Method for Calculating Transient EM Field in Complex Power Systems

Rodrigo M. S. de Oliveira, Reinaldo C. Leite, Ricardo H. Chamié Filho, Yuri C. Salame and Carlos Leonidas S.S. Sobrinho
Federal University of Pará (UFPA), Centrais Elétricas do Norte do Brasil S/A (Eletronorte)
Brazil

Abstract

The present work shows the results of a R&D project carried out by ELETRONORTE and by the Federal University of Pará (UFPA). Its core objective is the development of a computational system, called LANE-MAXWELL, for performing analysis and synthesis involving electromagnetic interference (EMI) in a power system substation environment. For the analysis stage, the numerical solution of the problem is obtained by numerically solving the Maxwell's equations written in a local non-orthogonal coordinate system by employing the Non-orthogonal Curvilinear Finite-Difference Time-Domain Method (LN-FDTD). The truncation of the analysis region is done by a new formulation called LN-UPML which involves the solution of the Maxwell Equations for lossy anisotropic media in a general coordinate system. For the synthesis stage, techniques such as neural networks, genetic algorithms and particle swarm optimization are used. The computational environment conceived has a friendly data input/output interface for the user, which permits a better understanding of the electromagnetic phenomena involved. For illustrate the versatility of the computational environment some practical applications involving complex power systems structures are presented.

Keywords
Complex Power Systems Structures, Parallel Computational Environment, LN-FDTD method, LN-UPML technique.

1. Introduction

It is widely known that the Finite-Difference Time-Domain (FDTD) method is adequate for solving numerically Maxwell's Equations in order to obtain accurate transient solutions regarding lighting discharges occurring in the vicinity or on power systems (Tanabe, 2001; Oliveira & Sobrinho 2009]. These lighting pulses can produce considerable power in bands

around frequencies much higher than 60 Hertz, what demands full-wave solutions for treating and modeling this kind of problem if high accuracy is desired for solutions. In fact, frequencies can reach dozens of megahertz. Besides that, in order to consider realistic situations, complex structures which constitute the power systems must be modeled, such as wires, dielectric materials, grounding systems (and the ground itself), transmission lines and circuit elements, which are distributed in the tridimensional space in highly complex geometric arranges, such as it happens in a power substation.

This way, a computational tool was built in order to allow the modeling and simulation of such electromagnetic environments. The software, based on the FDTD method, was implemented in such way that coordinates and electromagnetic parameters of the objects are defined by the user as input data, and an equivalent electromagnetic scenario is graphically displayed as an interactive tridimensional representation. The same data is treated by the FDTD simulator, which automatically divide the analysis domain among several computers connected as a Beowulf cluster (a distributed memory computer system). This way, problems in electrodynamics are treated in a friendly way so that the users do not need to have a strong background on physics or mathematics involved in order to construct the structures to be simulated. It is also important to mention that this tool also reduces the possibility of human mistakes while modeling the scenarios, because geometric information is represented graphically.

In order to show the simulations possibilities provided by this computational approach, the following problems were simulated: 1) lightning discharge in a curved grounding system; 2) lightning discharge in a grounded transmission line tower, in which voltages and currents across its insulators were calculated; 3) transient analysis of induced voltages in transmission lines in an urban environment, in which buildings, transmission lines, grounding systems and a radio base station are considered and 4) analysis of a power system substation behavior during an atmospheric discharge, in which high-voltage switches, circuit breakers, isolators, protection inductances, lightning arresters, protection fences, grit layer, etc are some of the elements considered in the numerical model.

The obtained results were consistent with those available in the literature (for simpler cases) and with the physics related to the problems. It is worth to say that the developed methodology can be used not only on power systems problems, but it can also solve a great range of different applications related to electromagnetic transient analysis, as it is based on full-wave solutions of Maxwell's equations and on automated parallel processing. The developed software represents considerable advance in this study field, as very complex problems can now be solved.

The computational environment conceived was called LANE MAXWELL (*Synthesis and Analysis of Grounding Systems*). The accomplishment of several tasks related to the project was possible by using the developed software. The mentioned tasks are the following:

1. Development of sensors for partial discharge (PD) detection in power system environment;
2. Characterization of EMI effects in a pilot substation defined by Eletronorte;
3. Development of practices for PD sources identification;
4. Development of practices to solve EMI problems;
5. Development of an adequate mathematical formulation for the EMI analysis, capable to provide useful information for various problems;
6. Electromagnetic Shielding projects;

7. Location of PD sources on high voltage (HV) coaxial cables from the obtained transitory signals by using optimization techniques.

The necessity of characterization of the electromagnetic phenomena with a greater accuracy led ELETRONORTE to establish an R&D project in cooperation with UFPA (Federal University of Para). This level of accuracy makes the analysis and synthesis of the problems more complex making it unviable by analytic means, in such way the numerical solution techniques became the most adequate choice for the treatment of practical problems. Nowadays, numerical solutions are considered to be as important as experimental measurements, as long as they are complementary to each other. The main objective is to make the practical measurements to have a more economic and safer implementation and procedures.

In this context, this work contributes in several ways, based on the development of a computational environment that permits the realization of the analysis and synthesis of problems in electrodynamics in a friendly way. The methodology used in the analyses involves the solution of the Maxwell's equations in Time Domain, written in curvilinear coordinates, by the Finite-Difference Time-Domain method (Taflove & Hagness, 2005). The analysis scenarios are truncated by the uniaxial perfect matching layer technique (UPML): a specific mathematic formalism had to be developed in order to adequate this technique to the curvilinear system (Oliveira & Sobrinho, 2007); for the synthesis several optimization techniques are available, such as: genetic algorithm (Rahmat-Sami & Michielsssen, 1999), particle swarm optimization (Lazinica, 2009) and artificial neural networks (Hagan, & Menhaj, 1994). The software is complemented by input/output interfaces that facilitate the creation of scenarios, visualization of electromagnetic field distribution, video generation, visualization of scenarios, generation of voltage and current graphics, etc.

2. Related Theory

A. The General Coordinate System

Considering a general or curvilinear coordinate system, a position vector $\vec{r}$ can be obtained as a function of the general coordinates, and its differential length vector $d\vec{r}$ is given by

$$d\vec{r} = \sum_{l=1}^{3} \frac{\partial \vec{r}}{\partial u^l} du^l = \sum_{l=1}^{3} \vec{a}_l \, du^l, \tag{1}$$

in which the vectors $\vec{a}_l$ are called unit vectors and form a unit base; which defines the axes of a general curvilinear space (each point in space has its own coordinate system). Figure 1 shows the vectors $\vec{a}_l$ (l = 1, 2, 3) which are tangent to the u^l axes and can be written as functions of the Cartesian Coordinates: x, y and z. An alternative set of three complementary vectors $\vec{a}^l$, reciprocal vectors, can be defined in such a way that each one is normal to two unit vectors with different indexes, forming a reciprocal base. This set can be mathematically calculated by the expressions

$$a^1 = \frac{\vec{a}_2 \times \vec{a}_3}{\sqrt{g}}, a^2 = \frac{\vec{a}_1 \times \vec{a}_3}{\sqrt{g}}, a^3 = \frac{\vec{a}_1 \times \vec{a}_2}{\sqrt{g}}, \tag{2}$$

in which $\sqrt{g}$ is the volume of the hexahedra formed by the vectors $\vec{a}_1, \vec{a}_2, \vec{a}_3$ (Fig.1). In (2), $\sqrt{g} = \vec{a}_1 \cdot (\vec{a}_2 \times \vec{a}_3)$, in which g is the determinant of the metric matrix or the covariant tensor (Taflove & Hagness, 2005).

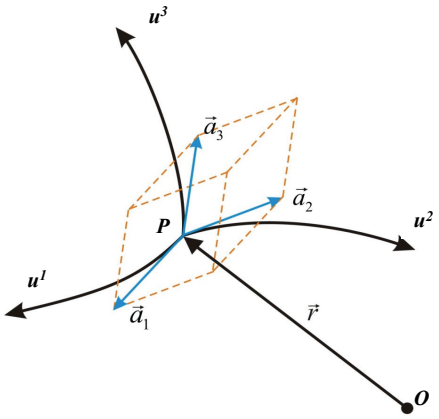


Fig. 1. Curvilinear coordinate system at point P and the unit vectors in a non-orthogonal cell.

Based on the previously presented idea, let a fixed vector $\vec{F}$ to be represented by the unit system or by the reciprocal system as shown by (3)

$$\vec{F} = \sum_{l=1}^{3} f^l \cdot \vec{a}_l = \sum_{l=1}^{3} f_l \cdot \vec{a}^l . \tag{3}$$

In (3), f_l and f^l are called covariant and contravariant components of the vector $\vec{F}$, respectively. The components f^l and f_l can be calculated by means of the dot product of the previous equation by $\vec{a}^l$ and $\vec{a}_l$ respectively, as it is shown by (4).

$$f^l = \vec{F} \cdot \vec{a}^l , f_l = \vec{F} \cdot \vec{a}_l . \tag{4}$$

These components (covariant and contravariant) can be related by means of the following expressions:

$$f_m = \sum_{l=1}^{3} g_{lm} \cdot f^l , f^m = \sum_{l=1}^{3} g^{lm} \cdot f^l . \tag{5}$$

It is worth mentioning that the unit and reciprocal vectors do not necessarily have unitary amplitudes because they depend on the nature of the curvilinear coordinate system used (cells' edge lengths). Hexahedrons similar to that in Fig. 1 are used as Yee's cells. This way,

the appropriate set of unit vectors and their respective unitary lengths are defined by the elementary equations

$$\vec{i}_1 = \frac{\vec{a}_1}{\sqrt{\vec{a}_1 \cdot \vec{a}_1}} = \frac{\vec{a}_1}{\sqrt{g_{11}}}, \vec{i}_2 = \frac{\vec{a}_2}{\sqrt{g_{22}}}, \vec{i}_3 = \frac{\vec{a}_3}{\sqrt{g_{33}}}. \tag{6}$$

From (3), we can write

$$\vec{F} = F^1\vec{i}_1 + F^2\vec{i}_2 + F^3\vec{i}_3, \tag{7}$$

in which F^l represents the physical components in the base system and their values are thus calculated by (8)

$$F^l = f^l \cdot \sqrt{g_{ll}}, F_l = \sqrt{g^{ll}}. \tag{8}$$

This representation of vectors can be used to describe electric and magnetic fields' components, as shown in the next topic.

B. The LN-FDTD Method applied for solving Maxwell's Equations

The Maxwell's equations in differential form can be written for lossy, isotropic, non-dispersive media as:

$$\nabla \times \vec{E} = -\mu \frac{\partial \vec{H}}{\partial t} \quad and \quad \nabla \times \vec{H} = \sigma\vec{E} + \varepsilon \frac{\partial \vec{E}}{\partial t}, \tag{9}$$

in which $\vec{E}$ is the electric field strength vector, $\vec{H}$ is the magnetic field strength vector, ε is the electrical permittivity of the media, μ is the magnetic permeability of the media and σ is the electrical conductivity. By calculating the contravariant components of the fields $\vec{E}$ and $\vec{H}$, by using central differences approximations for the derivative terms and by observing the primary and secondary cells in Fig. 2, the equations, for updating the components of the fields, can be obtained as follows:

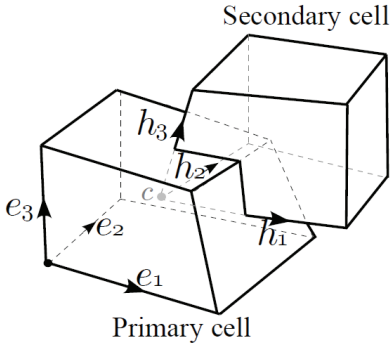


Fig. 2. Distribution of the covariant components in Yee non-orthogonal cells: primary cell for the electric field and secondary cell for the magnetic field.

For obtaining a field component (for example, the first contravariant component of $\vec{H}$, h^1), (4) (and (2)) can be used for l=1. In this case, the dot product operation is performed on Faraday's law. That is

$$\left(\nabla \times \vec{E}\right).(\vec{a}_2 \times \vec{a}_3)/\sqrt{g} = -\mu\left(\frac{\partial \vec{H}}{\partial t}\right).\vec{a}_1, \tag{10}$$

from which it is easy to see that

$$\frac{\partial H^1}{\partial t} = -\frac{1}{\mu\sqrt{g}}\left(\frac{\partial e_3}{\partial u_2} - \frac{\partial e_2}{\partial u_3}\right), \tag{11}$$

by observing that, from (2), it is possible to say that $\vec{a}^m.\vec{a}_n = \delta_{m,n}$, in which $\delta_{m,n}$ is the Konecker delta function. Equation (11) and the equations for the other five field components can be represented by finite differences by using the standard Yee's algorithm.

C. Stability and Precision of the LN-FDTD Method

The numerical stability of the method is related to the increment Δt, which follows the condition stated by (12),

$$\Delta t \leq \frac{1}{c \cdot \max\left(\sqrt{\sum_{l=1}^{3}\sum_{m=1}^{3} g^{lm}}\right)}. \tag{12}$$

To reduce the effects of numerical dispersion to adequate levels and in order to ensure the precision of the calculations, the dimensions of each non-orthogonal cell in every direction should be less than λ/10 (Taflove & Hagness, 2005), where λ is the smaller wavelength involved in the problem.

D. UPML for Conductive Media – Truncation of the LN-FDTD Method

The Maxwell's equations in frequency domain for an unaxial anisotropic media, can be written as follows:

$$\nabla \times \vec{E}^* = -j\varpi\mu_0\mu_r\overline{\overline{S}}\vec{H}^* \quad and \quad \nabla \times \vec{H}^* = (j\varpi\varepsilon_0\varepsilon_r + \sigma)\overline{\overline{S}}\vec{E}^*, \tag{13}$$

in which ω defines the angular frequency of the electromagnetic wave; $\vec{E}^*$ and $\vec{H}^*$ are the Fourier transforms of the electric field strength and magnetic field strength, respectively, μ_r is the relative magnetic permeability and ε_r is the relative electrical permittivity, $\overline{\overline{S}}$ is the tensor that promotes the attenuation along the coordinate axes normal to the layers of the absorbing region and σ is the media conductivity. $\overline{\overline{S}}$ has the following form

$$\overline{\overline{S}} = \begin{bmatrix} \frac{S_2 \cdot S_3}{S_1} & 0 & 0 \\ 0 & \frac{S_1 \cdot S_3}{S_2} & 0 \\ 0 & 0 & \frac{S_1 \cdot S_2}{S_3} \end{bmatrix}, \tag{14}$$

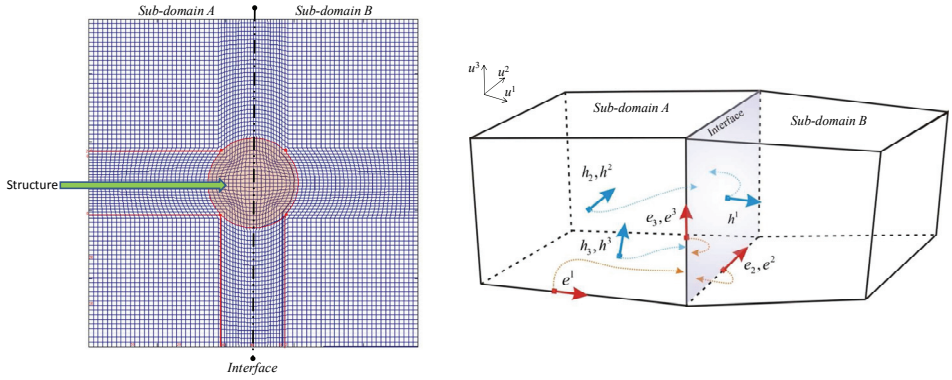
where S_α (α=1, 2, 3) are the parameters that characterize the UPML attenuation along the mentioned general coordinate axes, for which they assume the form (Taflove & Hagness, 2005):

$$S_\alpha = K_\alpha + \frac{\sigma_\alpha}{j\varpi\varepsilon_0}. \tag{15}$$

By following the procedures similarly to that was indicated in item B (and by introducing auxiliary variables for calculating transformations of fields to time domain), the updating equations for the field components are obtained (Oliveira & Sobrinho, 2007).

E. Parallel Processing for the LN-FDTD Method

The main idea of using parallel computation for solving electromagnetic problems by the LN-FDTD is based in a division of the analysis domain into sub-domains A and B (Fig. 3a). Each sub-domain is a fraction of the whole numerical volume that will be treated by a single processing core. Each core executes essentially the same FDTD code, but with particular boundary conditions. Field information must be exchanged at the sub-domains' interfaces, as illustrated by Fig. 3b.



(a) (b)

Fig. 3. (a) Domain decomposition into two sub-domains and cross section of a 3-D non-orthogonal grid; (b) the field components exchanged between two neighbor sub-domains.

As it is also illustrated by Fig. 3(a), some structures can be divided among two (or possibly more) sub-domains. This issue was treated by implementing a subroutine, which automatically distributes the media parameters among the processing units in such way that

the electromagnetic behavior of the divided scenario is identical to the behavior of the full numerical domain processed by a single core. This way, the software can generate complex electromagnetic scenarios by automatically defining the parameters for each CPU by using the data defined by the user in a graphical user interface (GUI), which is graphically displayed by a specific visualization tool developed by using the OpenGL library. This tool, associated to the automatic domain division subroutine, significantly reduces the probability of human errors when constructing and simulating the scenarios. Many figures on this work are direct outputs of the 3D-viewer.
The developed software also supports the insertion of thin-wires (Baba et al., 2005) and thin-planes (Maloney & Smith, 1992) (considering the Cartesian coordinate system), and ground ionization due lighting discharges (Ala et al., 2008).

3. Software Applications

This section, results obtained by using the developed computational environment are presented. Such applications range from the analysis of simple grounding systems to highly complex electromagnetic structures, such as the analysis of the effects of lighting surges on power substations. This way, initially a validation of the implementation is shown, by comparing experimental and numerical results for a grounding structure. Then the importance of using the local coordinate system is evidenced for a grounding structure which geometry is not compatible to the Cartesian system of coordinates. Finally, the effects of lighting currents on three complexes are analyzed for three cases: transmission tower, induced voltages on transmission lines for an urban block of buildings and for the structural part of a power substation.

A. Analysis of a rectangular grounding system.
In order to validate the implementation of the parallel numerical algorithm previously described, the transient responses of a rectangular electrode due the application of a artificial lighting current, originally proposed and analyzed by (Tanabe, 2001) were calculated. The geometry is illustrated by Fig. 4, which was reproduced in the computational environment.
Basically, this problem consists on a rectangular grounding electrode, with dimensions of 0.5×0.5×3.0 m, buried in a soil with electromagnetic parameters $\sigma = 2.28$ mS/m, $\mu = \mu_0$ and $\varepsilon = 50\ \varepsilon_0$. Additionally, the problem includes two circuits: one for current injection and other for voltage measurement, which dimensions are defined by Fig. 4. This figure also shows the identification of the sub-domains assigned to each processing core of a Beowulf cluster of PCs (regions numbered from 1 to 10), defined automatically by the automatic domain division subroutine.
Fig. 5 shows the numerical results obtained by the simulator developed in this work and the results obtained in (Tanabe, 2001) by conducting physical experiments. It is possible to observe that experimental and numerical transient responses for current, voltage (and consequently for the Voltage/Current ratio) are agreeing for most of the time. The main propose of this test is to certify that the parallel simulator generates trustable and accurate results.

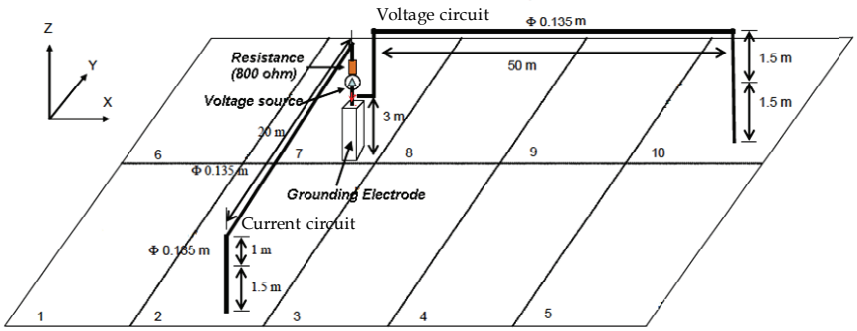


Fig. 4. Representation of the grounding electrode and of the measuring system originally proposed and analyzed by (Tanabe, 2001) and the automatic domain division schema defined automatically for this problem.

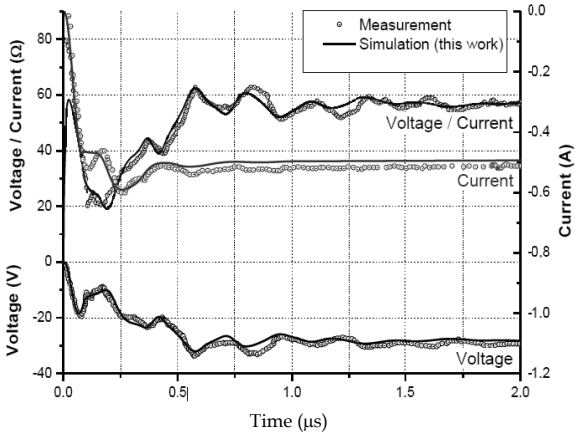


Fig. 5. Comparison of experimental measurements obtained by (Tanabe, 2001) to the numerical results obtained in this work.

B. Simulation of a circular grounding system by using Cartesian and curvilinear coordinates.

Figure 6 shows four vertical grounding rods connected by a circular conducting ring, modeled by the classical FDTD formulation. The rods are four meter long and the ring's diameter is eight meters, which is positioned 0.5 m under the ground surface The soil electrical parameters are $\sigma = 2$ mS/m and $\varepsilon = 50\ \varepsilon_0$. The idea is to compare results generated by employing the classical (orthogonal) FDTD method and the curvilinear FDTD methodology (LN-FDTD) and to investigate the influence of the staircase approximation effects for non-rectangular grounding structures and to evidence the need for using non-orthogonal modeling for such cases. The computational mesh illustrated by Fig. 3a was used for simulating this structure by using the LN-FDTD method. The circumference defined by the mesh of Fig. 3a was used to represent, in a more accurate way, the metallic grounding ring analyzed in this example.

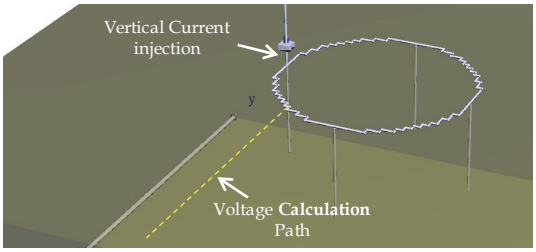


Fig. 6. A circular grounding structure modeled by using the Cartesian coordinate system.

According to (Mattos, 2004), the influence of the diameter of the conductors is negligible for the steady grounding resistance R (0.0 Hz), which is given by the equation

$$R = \left(\frac{1}{4L} + \frac{1}{4\pi D}\right)/\sigma \, . \tag{15}$$

In (15), L is the rods' length and D is the diameter of the considered circumference. For this case, the expression (15) provides R = 36.224 Ω, which agrees very well to the result obtained by using the LN-FDTD method (Fig. 7) at 2.5 μs (R = 36.973). This Figure also presents the results obtained for simulations performed by the classical FDTD algorithm, by approximating the conducting circumference by staircase. It is observed that, as the FDTD spatial step Δ is reduced, the results tend to converge to the result obtained by the LN-FDTD method (for both, the steady state and transient periods). In Fig. 7, it is also possible to observe that the ohmic values calculated by the FDTD method are higher than the values obtained by the LN-FDTD simulator. This is consistent, as the stair case approximation for the circular path for the electrical current acts as an obstruction for the movement of the electrical charges, thus increasing the voltage/current ratio. This effect is greatly reduced by the non-orthogonal coordinate system and it clearly shows that there is the necessity of avoiding this kind of geometric approximation for performing such analysis especially for higher frequencies.

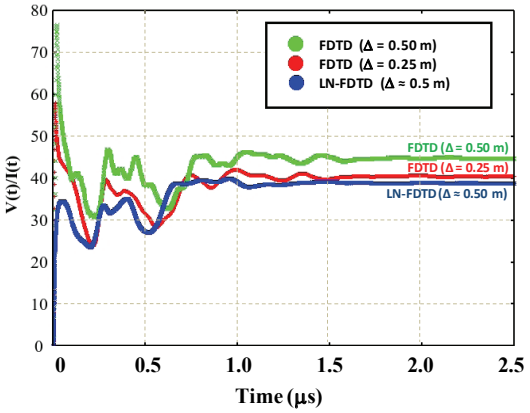
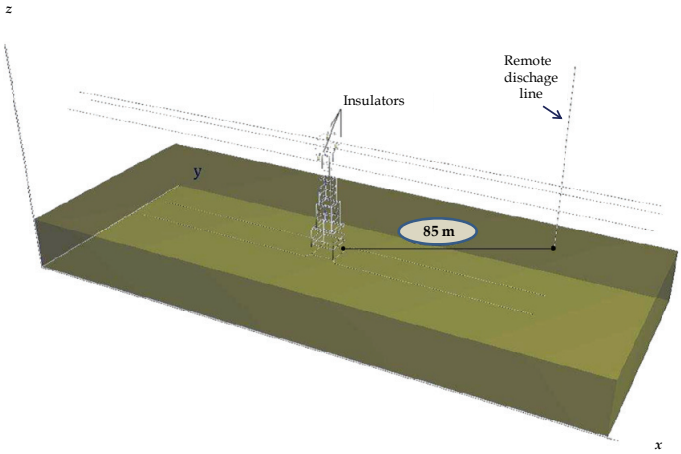


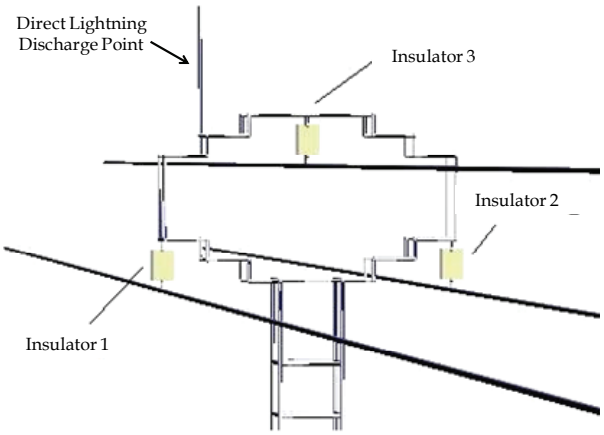
Fig. 7. Voltage / Current ratios obtained by using the FDTD method (Δ = 0.50 m and Δ = 0.25 m) and by the LN-FDTD method (Δ around 0.50 m).

C. Transmission Line Tower and Induced voltages on Line Insulators due to Remote Lighting Discharge

One of the several applications of the developed computational environment is the simulation of atmospheric discharges and the calculation of induced voltages at specific parts of a structure of interest. For this problem, this structure is a transmission line tower and its line insulators.



(a)



(b)

Fig 8. (a) overview of the scenario created for the calculation of the induced voltages due to atmospheric discharges and (b) detail of the positioning of the insulators, transmission lines and the location of the point of occurrence of the direct lighting stroke.

Fig. 8 shows the simulation scenario created. The goal here is to calculate the transient induced voltages at the insulators due to a lighting current reaching the ground surface 85

meters away from the tower. The simulations were carried out by considering the following parameters: numerical volume: 250x70x100 m by using cubic Yee cells with edges of 0.5 m; ground parameters: $\varepsilon = 50\ \varepsilon 0$ and $\sigma = 2$ mS/m; insulator parameters: the insulator modeled has 1.0 m of height with the electrical conductivity of 10^{-11} S/m and relative electrical permittivity of 7.5; excitation source: the same used previously (Tanabe, 2001). The lighting discharge was included into the analysis domain by creating a vertical conductor, which penetrates the field absorbing boundary region (UPML) at its top (with impedance matching), representing an infinitely long conductor. By forcing the magnetic field around this conductor to follow the function defined by (Tanabe, 2001), a punctual transient current (52 meters from ground surface) with maximum value of 10 kA was created and used as source. The transmission lines and the grounding rods also penetrate the UPML region.
The basis of insulators 1 and 2 (Fig. 8b) were placed at 43.5 meters from the ground surface and the basis of insulator 3 was placed 49.5 meters from the ground plane. It was also simulated the case in which the lighting current occurs at the tower structure, at the point indicated by Fig. 8b and the obtained results for induced voltage on insulator 1 are compared in Fig. 9.
Fig. 9 shows the behavior of the voltage induced on the insulator 1 (*z*-oriented path). In Figure 19, it is shown the induced voltage during the transient period and in Figure 19b it is shown the induced voltage for the stationary period. For this insulator, induced voltage reaches no more than 450 kV during the transient time evolution. As expected, by considering the direct lightning current injection at the tower top, produces a much higher peak, which reaches approximately 1.8 MV at 0.5 μs, as shown by Fig. 9a. As shown by Fig. 9b, the difference between the steady induced voltages is approximately 24 kV. For each 10 kA of lighting current peak, the insulator can be subjected to 50 kV (which is added to the nominal line voltage) during considerable amounts of time.

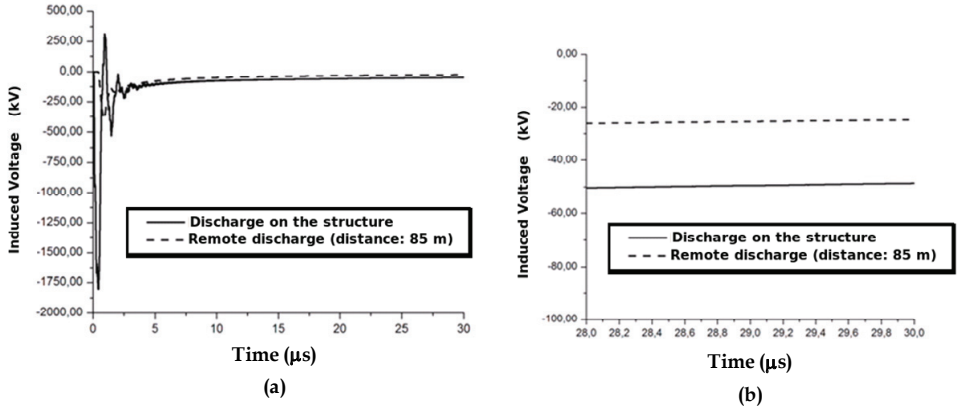


Fig. 9. (a) induced transient voltages and (b) the induced voltages at the stationary period

D. Distribution Line in Urban Environment: lighting discharge at a Radio Base Station (RBS).

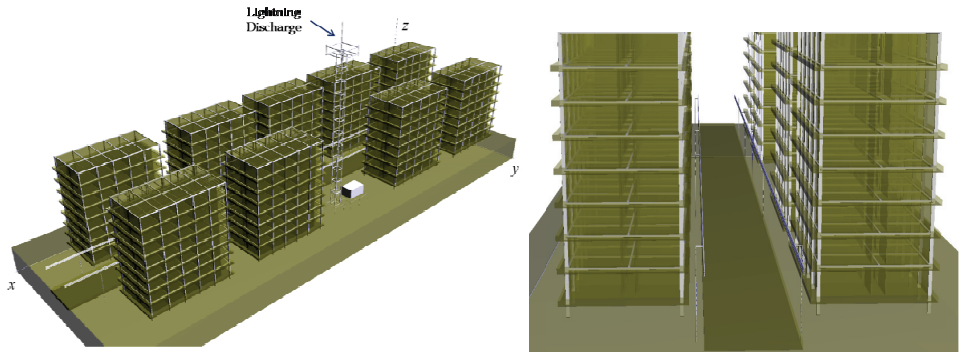
In recent years, due to world-wide expansion of cell phone systems in urban areas, the raising of tower-like structures operating as radio-base stations (RBSs) became common in

cities. With considerable heights, around 50 meters, these structures are preferential points for atmospheric discharge strokes. The discharges can occur at the tower's lightning rod or at its metallic body. The occurrences of discharges on telephone towers cause an undesirable set of effects, either in the ground (such as increase of potentials and return currents in grounding systems) or at lines for energy transmission (causing outbreaks due to electromagnetic induction). Such outbreaks are characterized by high-voltage peaks, followed by sudden potential differences between pairs of conductors of the low voltage lines, affecting the consumer units with possible serious damage to electric and electronic devices.

The goal of this application of the LANE-MAXWELL software is to study the induced voltages on low-voltage systems and to estimate how the electromagnetic field, generated by atmospheric discharges on high metallic structures might behave by taking into account the complex structures with different materials around the transmission lines. The two low voltage lines in this scenario can be represented by four metallic cables suspended in the air lined up and separated vertically by a specific distance. On both modeled low voltage lines, two connections between the neutral cable and the ground are established. These low voltage cables were positioned in such a way that their extremities penetrate the UPML of the analyzed region. This set of complex structures modeled in this simulation is represented by a realistic Radio Base Station with its metallic container of equipments and its grounding structure, a set of realistic modeled buildings located next to the RBS, and a realistic road, which was positioned parallel to the *x* direction of the analysis region as shown by Fig 10. The conception of a scenario was carried out by using the following parameters:

1. Nine eight-floor buildings of 24 m height each, with the following dimensions: 24x12.8x24 m;
2. Analysis region with 168x60x64 m;
3. Yee cells with edges measuring Δ = 0.2 m (840x300x320 cells);
4. Lateral and frontal separation spaces for Buildings: 6 m and 14 m, respectively;
5. The bars were considered as perfect metal conductors;
6. Wall and street material: reinforced concrete (σ=0.2 S/m, ε_r = 7.5 and μ_r = 1);
7. Low voltage lines with three cables as phase and one as neutral conductors, located at 10 m and 20 m away from the geometric center of a cell phone radio base station (RBS);
8. Cable heights of 6.6 m, 6.8 m, 7.0 m (the three phase lines) and 7.2 m (the neutral line) respectively from the ground. Cables' diameter: 15 mm.
9. RBS is 50.0.m-high and the container has dimensions of 6.0x4.0x3.0 m ;
10. Neutral grounding with ½"x 3 m length cooper rods, with a grounding resistance of 80 Ω.
11. Conductive earth with electrical parameters σ= 0.0040836 S/m and ε_r = 10.
12. Excitation source: see Fig. 12a.

Figure 11 shows the scenario built in the LANE-MAXWELL environment, detailing the above specifications.



(a) (b)

Fig. 10. (a) the virtual urban environment built for analysis of the electromagnetic induced voltage on electrical distribution lines (b) a longitudinal view of the street and distribution line.

The obtained transient induced voltages on the cables due to an atmospheric discharge on the RBS (Fig. 12a) are shown by Figure 12b. In this case, the ground system of the buildings and the line were not connected together trough wires (current flows trough ground only). Fig. 12b shows that, for the three live cables located 10 meters away from the tower, the peak value of the induced voltage is around 8.0 kV when the simulation time is about 1.2 μs. The induced voltages on both energy lines are caused partially by the electric current that propagates trough ground and circulates trough the lines' grounding electrodes and partially by the direct and indirect incidences of the electric field on the lines. It is important to mention that the smooth oscillations present on Fig. 12b can be caused by reflections of this propagating electromagnetic field on the buildings' walls, on the ground surface and on the lines themselves, and these oscillations tend to disappear as time progresses.

Moreover, results show that induced voltage at a line takes more than a half microsecond to start decaying from its maximum value. However, it is important to mention that the voltage induced at the neutral cable was also increased proportionally, so that the live-live and live-neutral voltages did not suffer any major changes. In the final moments of the simulation, all four curves tend to get closer quickly. Similar behavior is observed for the line located 20 m away from the tower.

Figure 13 shows the electric field spatial distributions and its time evolution at the ground surface (Figs. 13a-c) and at the vertical *yz*-plane which contains the current source (Figs. 13 d-e).

In order to facilitate the visualization of the electric field distribution inside the analysis region, the behavior of the electric field was registered at two planes inside the structure, generating hi-resolution colored images, in which the blue tones represent lower values of electric field, and the red tones or represent higher intensities of electric field. Fig. 13 shows the electric field distribution on the *xy* plane, with z = 7.8 m (ground surface) at three different steps of simulation. These figures show that the electric field is more intense at adjacent regions of the RBS' grounding mesh, and these values decays abruptly as the observation point departs from it, generating significant amounts of step potentials towards the mesh neighborhood. Such electric field behavior generates a potentially dangerous region near these structures during a lighting stroke. In Fig. 13, it is still possible to visualize

the *yz*-plane of the structure at x = 84 m (midpoint of the energy line) after 3 μs (12982 iterations) of simulation. It is clear that the electric field which penetrates the buildings is less intense than those present in regions outside this structure. This fact is due to the metallic and concreted constitution of the building structure, which acts as a Faraday's Cage the field, keeping most of the energy of the scattered electromagnetic wave outside the buildings' structure.

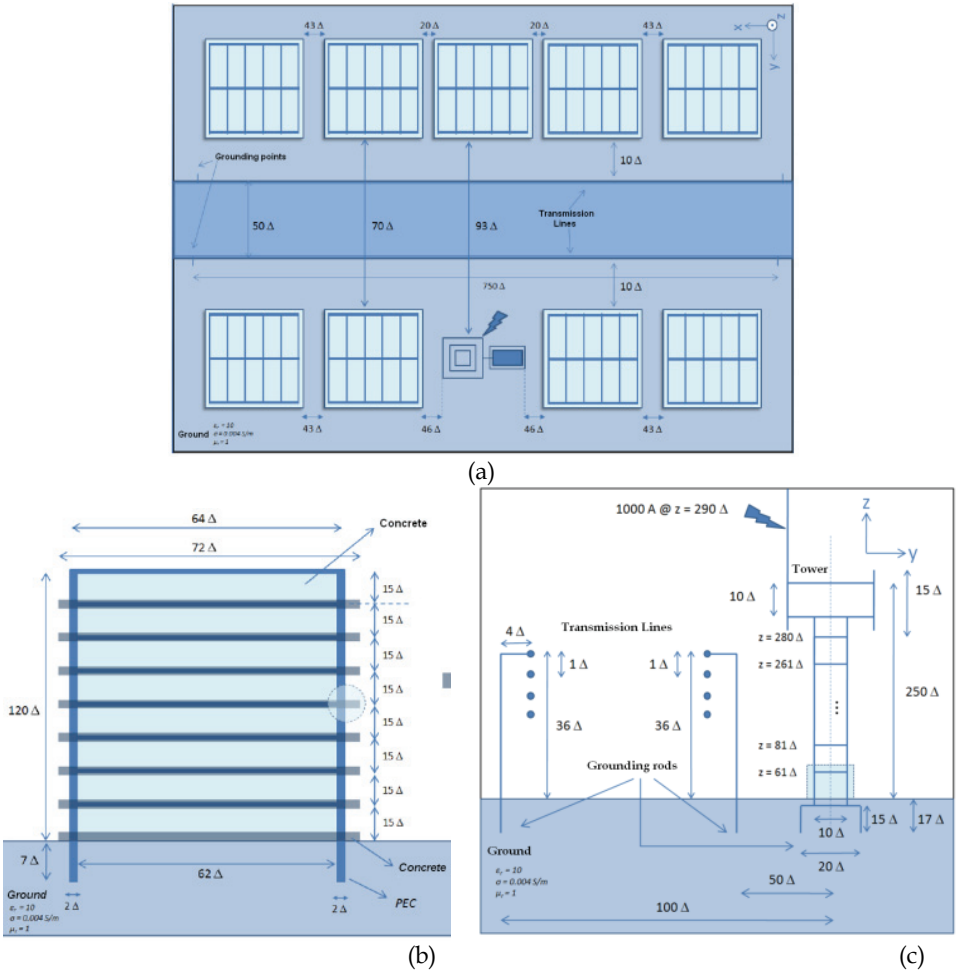


Fig. 11. (a) top overview of the urban block, (b) dimensions of one of the buildings and (c) tower, grounding system and transmission lines geometric configuration.

The analysis of the obtained results show high levels of overvoltage and potential differences between live to live and live to neutral conductors on electrical distribution systems, during lightning strokes on a RBS. The results computed with the FDTD method, include full wave solutions for complex structures, involving dielectric materials, a complex

tower model, transmission lines, grounding systems, and the consequent complex electromagnetic interactions involving surface waves, diffractions, refractions, and reflections. It has been shown that all those aspects substantially affect results, which, in this work, are close to realistic situations.

The problems observed here (relative to transitory high voltage induced in transmission lines) represent high risk when considering the final consumer unit, providing both material and human health damage.

Besides that, it was possible to verify from the electric field spatial distribution that high potential differences induced on the soil surface can be found in regions near the tower and its grounding mesh. Such structures are usually located near residences, or in places with constant circulation of people. Telecommunication towers require special attention by the operators which are responsible for the service and by the local energy company as well, keeping safe areas located near those installations, as well as the energy consumer unit.

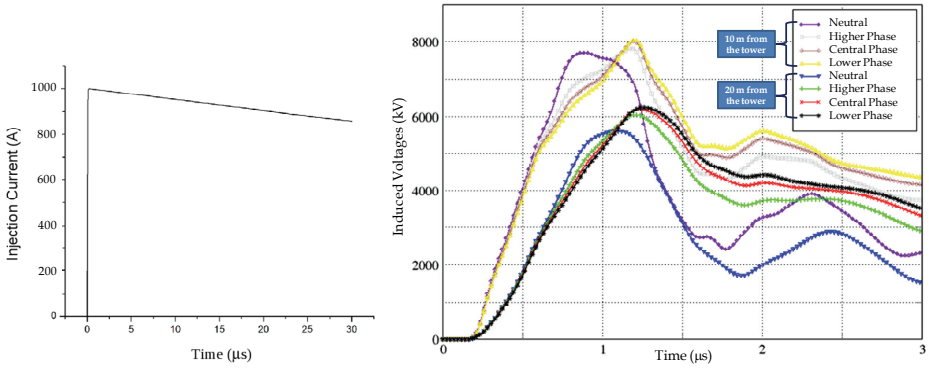


Fig. 12. (a) waveform of the injected current, used as excitation source and (b) the transient voltages obtained for each transmission line.

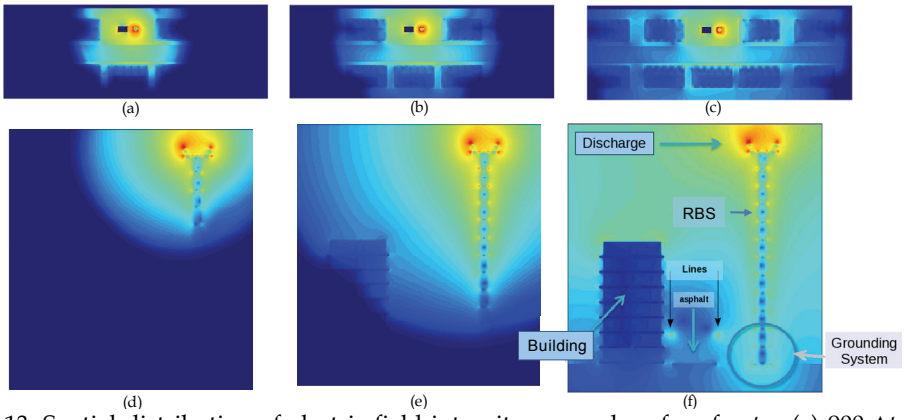


Fig. 13. Spatial distribution of electric field intensity: ground surface for t = (a) 900 Δt, (b) 1050 Δt, (c) 1440 Δt; and at the plane x = 420 (the plane containing the source) for t = (d) 330 Δt, (e) 420 Δt and (f) 1410 Δt.

E. Analysis of Induced voltages on Transmission Lines of a Power Substation due to Atmospheric Discharge.

Figure 14 shows an overview of a computational model of the power substation created by using the LANE-MAXWELL software. It is composed by the structural elements of high-voltage switches, insulators, transmission lines, towers, fence, grit layer, grounding grid, current Transformers, voltage transformers, lightning arrester, etc.

The analyzed region's volume is 1000x195x45 m^3, which is represented by cubic Yee cells with edges measuring 0.5 m. As indicated by Fig. 14, the lighting current was injected at a lighting arrester placed near the 230 kV line, with a maximum value of 1.0 kA (Tanabe, 2001). The wave propagation phenomenon was analyzed up to 12 μs after the beginning of the lighting process. The electrical parameters used for representing the ground region were $\varepsilon = 50\ \varepsilon_0$ and $\sigma = 2$ m S/m. The grit layer was modeled as an isotropic media with thickness of 0.5 m and with $\sigma = 0.333$ m S/m $\varepsilon = 50\ \varepsilon_0$. This model follows the specifications of a real substation, which is under responsibility of Eletronorte, installed at the city of Belém (Brazil).

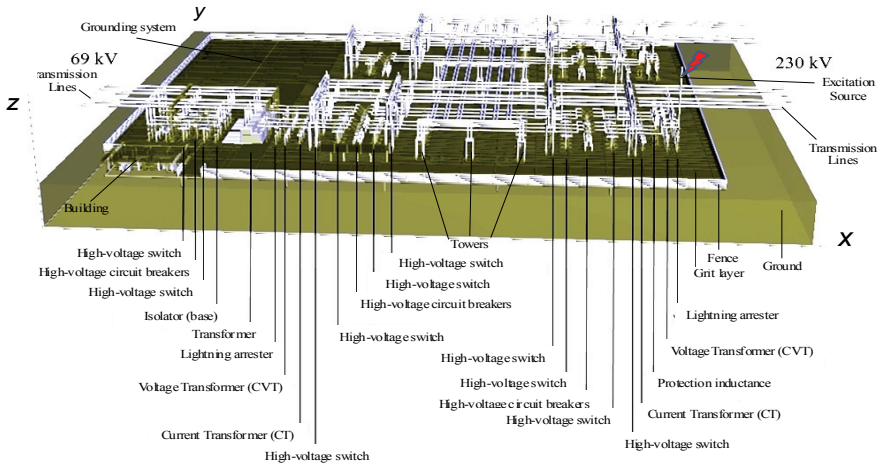
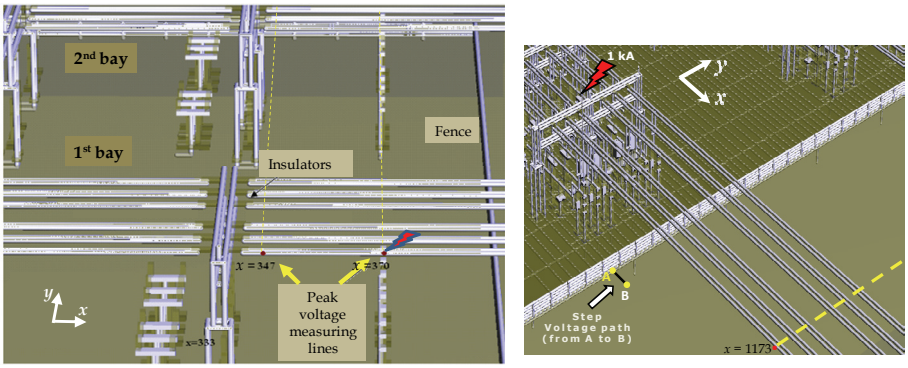


Fig. 14. Structure of the Power Substation represented in LANE-MAXWELL.

In order to create the scenario of Fig.14 with precision, specific database software has been developed. The idea is to create prototypes of the objects of the substation and to clone them at proper spatial coordinates. These objects (towers, switches, etc.) are composed by basic elements of the simulation software (metal and dielectric blocks and thin wires), which are grouped forming a prototype. A text file containing the positions of each clone of each prototype defines the electromagnetic environment. Then, the domain decomposition subroutine distributes the basic elements among the processing sub-domains, which quantity of cores to be used is defined by the user (Fig.3a). Here, sixteen 64-bits processors were used as a Beowulf cluster for solving this problem, which requires approximately 12 hours for concluding the computer simulation. Each core processed a volume of 62.5x195x45 m^3 each time step.

Figure 15a shows two *y*-aligned lines (in doted yellow style) over which induced transient voltages, calculated from the transmission lines to ground, were registered. The red points numbered as 370 and 347 indicate the discrete Yee's indexes related to the *x* direction, where the voltages are measured. Figure 15b defines of the path used for integrating electric field,

at the ground surface, in order to calculate the transient step voltage outside the substation, in the region as close as possible to the discharge point. The points A and B are separated by one meter (average separation between the legs of a human victim).



(a) (b)

Fig. 15. (a) Definition of two of the lines (dashed yellow style) were induced voltage peaks were registered. Voltages were measured from transmission lines to ground and (b) definition of the path used for integrating electric field and determining the transient step voltage (from A to B) ; definition of the line in which induced voltages were registered at the transmission lines (x = 1173). Coordinates are given in cells.

Figure 16 presents the obtained maximum induced voltages on the transmission lines, measured from the ground surface to those lines, considering (a) x = 370, (b) x = 347, (c) x = 316 and (d) x = 1173, along the y direction. The voltage values obtained for each transmission line are indicated by the black dots on the curves (lines were obtained by interpolation).

As shown by Fig. 15a, the line in which x = 370 is the closest to the discharge point. For the first line, the voltage peak induced reaches about 110 kV (Figs. 16a and 16b). For its neighbor parallel line, the induced voltage is radically reduced to 35 kV (x = 370) and to about 15 kV for x = 347. For the case of Fig. 16a, the induced voltage on the remaining transmission lines decreases exponentially, reaching about 10 kV at the second bay's lines. For the case of Fig. 16b (12 m from the line considered in Fig. 16a), it is possible to see that the decay pattern is close to an exponential function with oscillations for the first bay lines. The voltages for the second bay are also close to 10 kV for each kA injected.

Figure 16c shows similar results obtained for x = 316. It is possible to see in Fig. 15 (a) that between this point and the discharge coordinates there are several structures, like towers and switches, which act as reflective objects to the electromagnetic field, what justifies the considerable reduction of the peak voltage for the first line (to about 50 kV). The induced voltage to the neighbor line drops to 25 kV and decays almost exponentially to about 6.5 kV at the second bay. The reduction due the reflection of the fields are confirmed when results of Fig. 16d are analyzed. The reference point is x = 1173 (Fig. 15b), which is outside the substation (401 m from the discharge point). The maximum induced voltage is close to 72 kV, decreasing exponentially to 5 kV as distant lines are analyzed. It is also confirmed by analyzing Fig. 17, which shows the electric field distributions for (a) t = 0.225 μs, (b) t = 0.365

μs and (c) t = 1.00 μs. In those figures, the blue color represents small amplitudes, green represents intermediate strengths and red is associated to the regions of greater intensity for electric field. For all cases, it is possible to see that the electric field presents smaller intensities as one moves from the fist bay (where the discharge occurs) to the second bay (in y-direction). However, when one moves from the first bay to outside of the substation (in –y-direction), the field attenuation is less intense.

Fig. 18 (a) shows the obtained step voltage between the points A and B indicated at Fig. 15b. It is possible to see that after approximately 1 μs of the beginning of the lighting discharge, this voltage reaches its maximum value of about 807 V for each kA of peak current. However, its time duration is relatively short. However, the person can be exposed to voltages around 200 V for each kA of lighting current for a much longer time, depending on the time and amplitude characteristics of the injected current. Finally, Fig. 18b shows the electric field distribution at the ground grid plane after 38.5 ms of simulations, illuminating the grid. It is possible to observe that there are red dots, indicating higher intensities of the field, which are associated to the electrical connections of the substation devices to the ground grid. This indicates that current is returning or it is being injected to the grounding mesh.

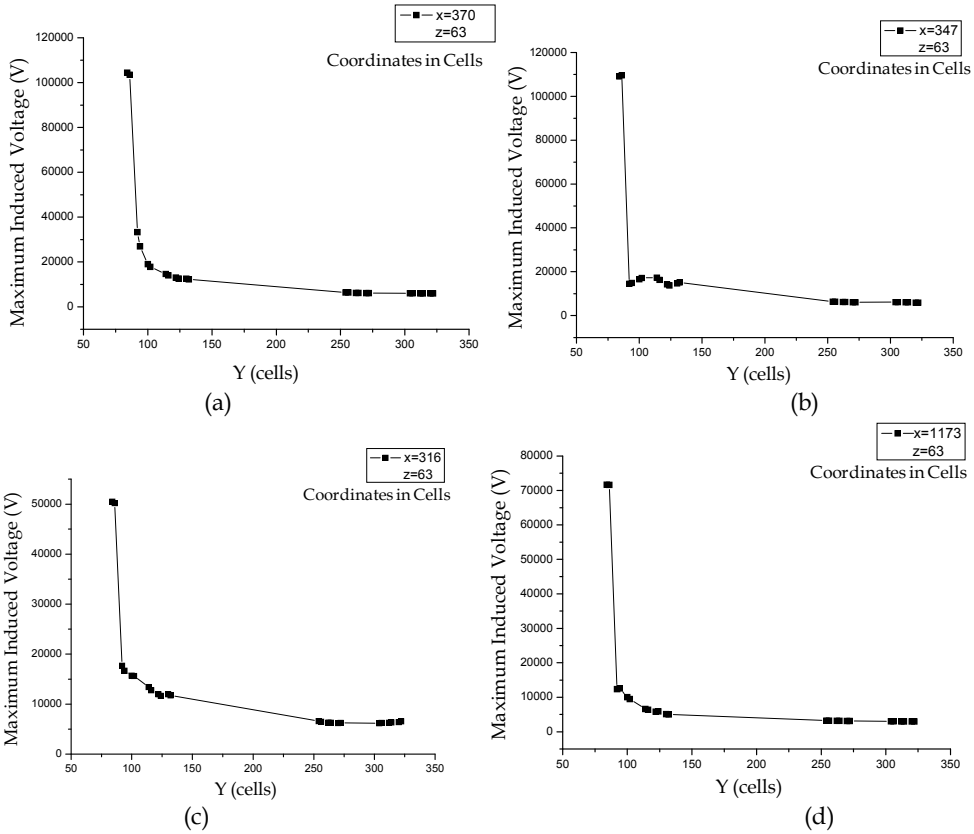
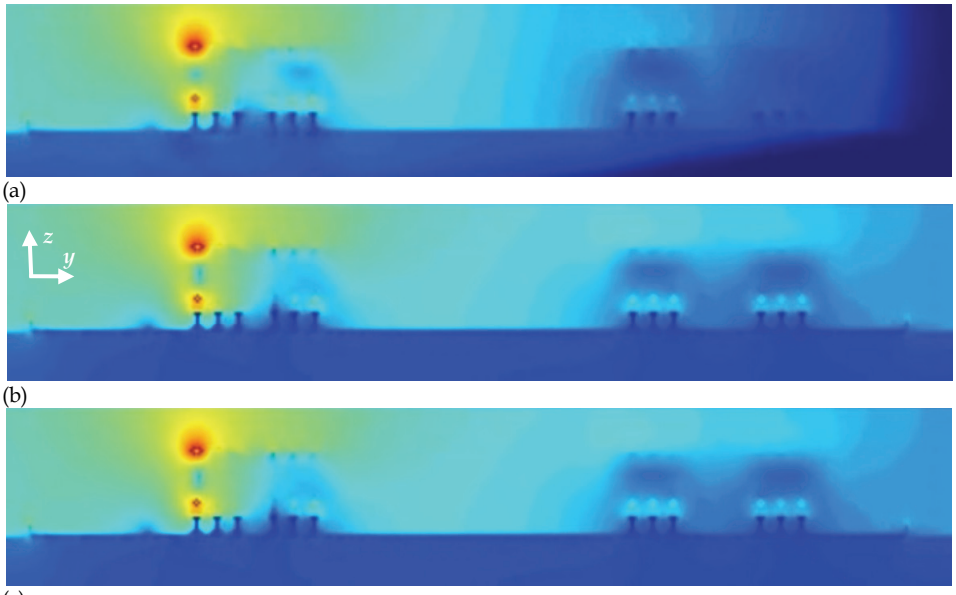


Fig. 16. Maximum induced voltages induced on transmission lines for (a) x = 370, (b) x = 347, (c) x = 316 and (d) x = 1173.

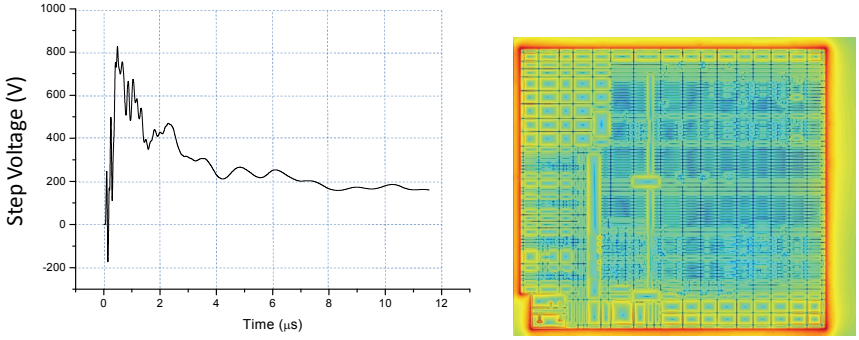


(a)

(b)

(c)

Fig. 17. Spatial electric field distribution for the plane parallel to the *yz*-plane which contains the discharge source, for: (a) t = 0.225 μs, (b) t = 0.365 μs and (c) t = 1.00 μs.



(a) (b)

Fig. 18. (a) Step voltage as function of time and (b) E_z electric field component distribution for t = 38.5 μs at the plane of the ground grid.

4. Conclusion

This R&D project generated a computational environment for the analysis and synthesis of problems on EMC. The initially obtained results were consistent with those available in the literature and with the physics that the problems involve. The used methodology represents a full wave solution and can be used in any frequency range. The developed software can be used in a great range of different applications, as Maxwell's equations are solved by an automated parallel processing environment.

The association of the computational environment with the laboratory equipment acquired represents a desirable infra-structure for the realization of high level works.

Acknowledgements

This work was supported by ELETRONORTE – Centrais Elétricas do Norte do Brasil S/A and by the CNPq – Conselho Nacional de Desenvolvimento Cientifico e Tecnológico.

5. References

Ala, G., Buccheri, P. L., Romano, P., Viola, F. (2008). Finite Difference Time Domain Simulation of Earth Electrodes Soil Ionisation Under Lightning Surge Condition. S*cience, Measurement & Technology, IET,* Vol.2, No.3, May 2008, 134 - 145, ISSN: 1751-8822.

Baba, Y., Nagaoka, N. and Ametani, A. (2005). Modeling of thin wires in a lossy medium for FDTD simulations. *IEEE Transactions on Electromagnetic Compatibility,* Vol. 47, No. 1, Feb. 2005, pp. 54-60, ISSN: 0018-9375.

Hagan, M. T. & Menhaj, M. B. (1994). Training Feedforward Networks with the Marquardt Algorithm, *IEEE Transactions on Neural Networks,* Vol. 5, No. 6, Nov. 1994, 989–993, ISSN: 1045-9227.

Lazinica, A. (2009). *Particle Swarm Optimization,* InTech Education, ISBN: 978-953-7619-48-0.

Maloney, J. G. & Smith, G. S., R. (1992). The efficient modelling of thin material sheets in the finite-difference time-domain (FDTD) method. *IEEE Transactions on Antennas and Propagation,* Vol. 40, No. 3, March 1992, pp. 323-330, ISSN: 0018-926X.

Mattos, M. A. (2004). *Técnicas de Aterramento,* Okime, ISBN:8598294012, Brazil.

Oliveira, R. M. S. & Sobrinho, C. L. S. S. (2007), UPML Formulation for Truncating Conductive Media in Curvilinear Coordinates, *Numerical Algorithms*, vol. 46, No. 4, Dec. 2007, pp. 295-319, ISSN: 1572-9265.

Oliveira, R. M. S. & Sobrinho, C. L. S. S. (2009). Computational Environment for Simulating Lightning Strokes in a Power Substation by Finite-Difference Time-Domain Method. *IEEE Transactions on Electromagnetic Compatibility*, Vol. 51, No. 4, Vov. 2009, pp. 995-1000, ISSN: 0018-9375.

Rahmat-Sami, Y. & Michielsssen, E. (1999). *Electromagnetic Optimization by Genetic Algorithms,* John Wiley & Sons, ISBN: 0471295450, Canada.

Taflove, A. & Hagness, S. C. (2005), *Computational Electrodynamics: The Finite-Difference Time-Domain Method*, Artech House, ISBN: 1580538320, Boston-London.

Tanabe. K. (2001). Novel method for analyzing the transient behavior of grounding systems based on the finite-difference time-domain method, *Proceedings of Power Engineering Society Winter Meeting*, pp. 1128-1132, OH USA, Feb. 2001, IEEE, Columbus.

Yee, K.(1996), Numerical solution of initial boundary value problems involving Maxwell´s equations in isotropic media, *IEEE Trans. Antennas and Propagation*, vol. 14, No. 3, May 1966, pp. 302-307, ISSN: 0018-926X.

4

Study on Oscillation Damping Effects of Power System Stabilizer with Eigenvalue Analysis Method for the Stability of Power Systems

Fang Liu[1,3], Ryuichi Yokoyama[1], Yicheng Zhou[2], Min Wu[3]
[1]*Waseda University*
Japan
[2]*Tepco Systems Corporation*
Japan
[3]*Central South University*
China

1. Introduction

With the increasing of the scale and the complexity of the interconnected power networks, the problems on the various potential power oscillations, which have the nervous damage against the system stability and the security operation, have been drawn more and more attention (Kunder, 1994; Anderson & Fouad, 2003; Bikash & Balarko, 2005). Power system oscillations were first reported in northern American power network in 1964 during a trial interconnection of the Northwest Power Pool and the Southwest Power Pool (Schleif et al., 1966). Up to now, generally speaking, power oscillations could be divided into three kinds of types, that is, local mode, inter-area mode, and global mode. Local oscillations lie in the upper part of that range and consist of the oscillation of a single generator or a group of generators against the rest of the system. In contrast, inter-area oscillations and global oscillations are in the lower part of the frequency range and comprise the oscillations among groups of generators. As a classic oscillation mode, there are relative mature technologies and devices such as kinds of power system stabilizers equipped as a part of the additional excitation system of machine unit to provide the efficient damping ratio to suppress the local oscillation. Nevertheless, as for the inter-area and the global oscillation mode, the classic stabilizer cannot play an important role to damp such oscillation very well. The leaded result is that the line power transmitted from one area to another will form the instable oscillation with the unease attenuation characteristic.

As a result, if there is no the effective solution to suppress these power system oscillations, the instability could lead the machine unit cut even the networks breakout. Nowadays, severe consequences have been coursed by large-scale blackouts, such as blackouts in the USA, Europe and many other countries in recent years. Moreover, blackouts not only lead to financial losses, but also lead to potential dangers to society and humanity. So it is necessary to pay attention to keep the stability and security of the electrical power systems. Up to now, many authors are trying to develop new methods to enhance the various types of

oscillations in power system. Various theories and technologies are introduced to against such power oscillations, such as wide area measurement systems (Ray &Venayagamoorthy, 2008; Kawma & Grondin, 2002), FACTS devices (Pourbeik & Gibbard, 1996; Pourbeik & Gibbard, 1998; Zhang et al., 2006), robust controllers and the design technologies (De Oliveira et al., 2007; Pal et al. 1999), and so on, to enhance the stability and the security operating ability of the close-loop systems.

In this chapter, we will deal with the application of power system stabilizer to improve the power system damping oscillation by using eigenvalue analysis method. This paper is organized as follows: In Section II, the operating principle and main structure types of power system stabilizer (PSS) will be described briefly. In Section III, the eigenvalue analysis method based on small single model will be introudeced. In section Section IV, the detail nonlinear simulations on two typical test systems will be performed to evaluate the performance with installing powes system stabilizer (PSS). In Section V, we will give some conclusions.

2. Power System Stabilizers

2.1 Operating Principle

The basic function of power system stabilizer (PSS) is to add damping to the generator rotor oscillations by controlling its excitation by using auxiliary stabilizing signal(s). Based on the automatic voltage regulator (AVR) and using speed deviation, power deviation or frequency deviation as additional control signals, PSS is designed to introduce an additional torque coaxial with the rotational speed deviation, so that it can increase low-frequency oscillation damping and enhance the dynamic stability of power system. Fig.2.1 shows the torque analysis between AVR and PSS.

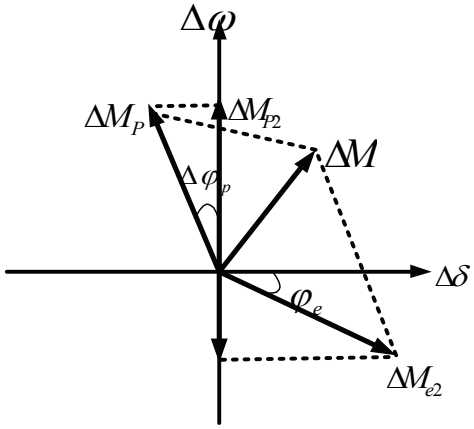


Fig. 2.1. Torque analysis between AVR and PSS

As shown in Fig.2.1, under some conditions, such as much impedance, heavy load need etc., the additional torque ΔM_{e2} provided by the AVR lags the negative feedback voltage ($-\Delta V_t$) by one angle φx , which can generate the positive synchronizing torque and the negative damping torque component to reduce the low frequency oscillations damping. On the other

hand, the power system stabilizer, using the speed signal ($\Delta\omega$) as input signal, will have a positive damping torque component ΔM_{p2}. So, the synthesis torque with positive synchronous torque and the damping torque can enhance the capacity of the damping oscillation. Fig.2.2 shows the structure diagram of power system stabilizer (PSS).

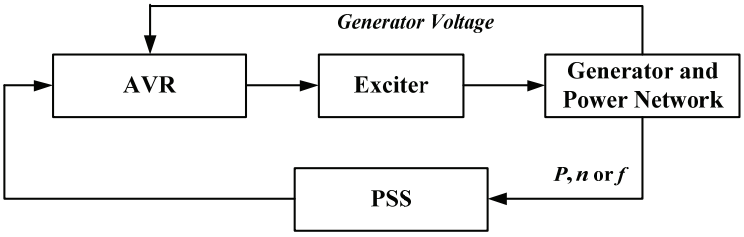


Fig. 2.2. The structure diagram of power system stabilizer(PSS)

2.2 Structure Types (IEEE Power Engineering Society (1992))

Power system stabilizers (PSS) are added to excitation systems to enhance the damping of power system during low frequency oscillation. For the potential power oscillation problem in the interconnected power networks, the power system stabilizers solution is usually selected as the relative practical method, which can provide the additional oscillations damping enhancement through excitation control of the synchronous machines.

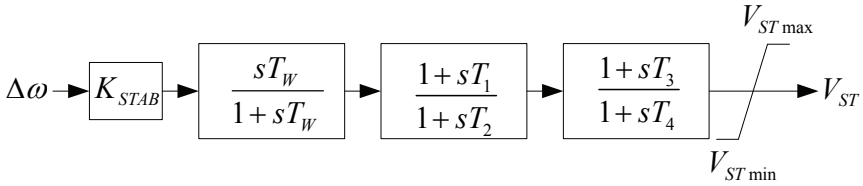


Fig. 2.3. General power system stabilizer model

Fig.2.3 shows the general power system stabilizer model with a single input, and from which, it can be seen that as for the additional damping control of the excitation system of the synchronous machines, basically the general input signal is the rotor speed deviation. The damping amount is mostly determined by the gain K_{STAB}, and the following sub-block has the high-pass filtering function to ensure the stabilizer has the relative better response effect on the speed deviation. There are also two first-order lead-lag transfer functions to compensate the phase lag between the excitation model and the synchronous machine.
Fig2.4 shows the power system stabilizer mode with dual-input singles, which is designed by using combinations of power and speed or frequency as stabilizing singles. From it, it can be seen this model can be used to represent two distinct types of dual-input stabilizer implementations. One hand, as for electrical power input stabilizers in the frequency range of system oscillations, they can use the speed or frequency input for the generation of an equivalent mechanical power signal, to make the total signal insensitive to mechanical power change. On the other hand, by combining the speed /frequency and electrical power,

they can use the speed directly (i.e., without phase-lead compensation) and add a signal proportional to electrical power to achieve the desired stabilizing signal shaping.

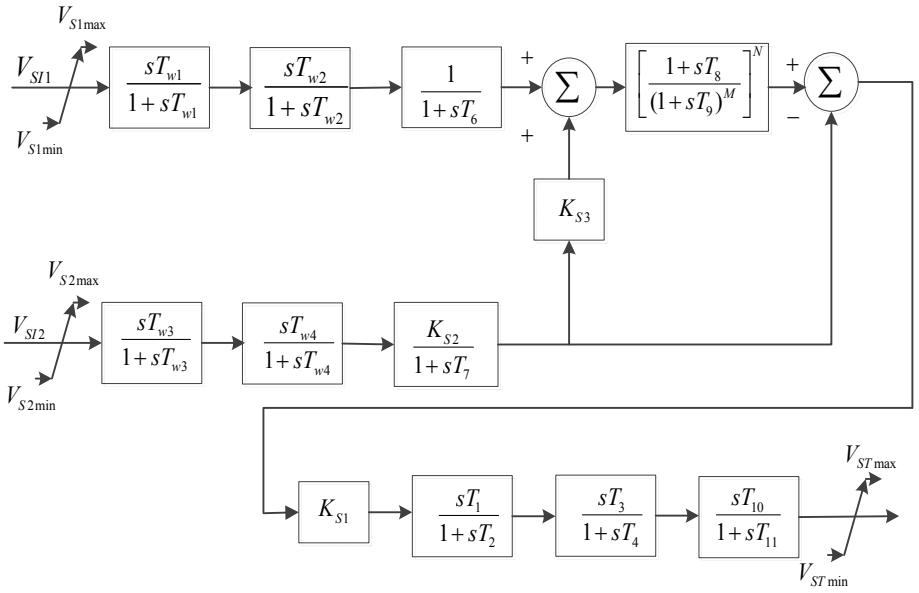


Fig. 2.4. Power system models with dual inputs

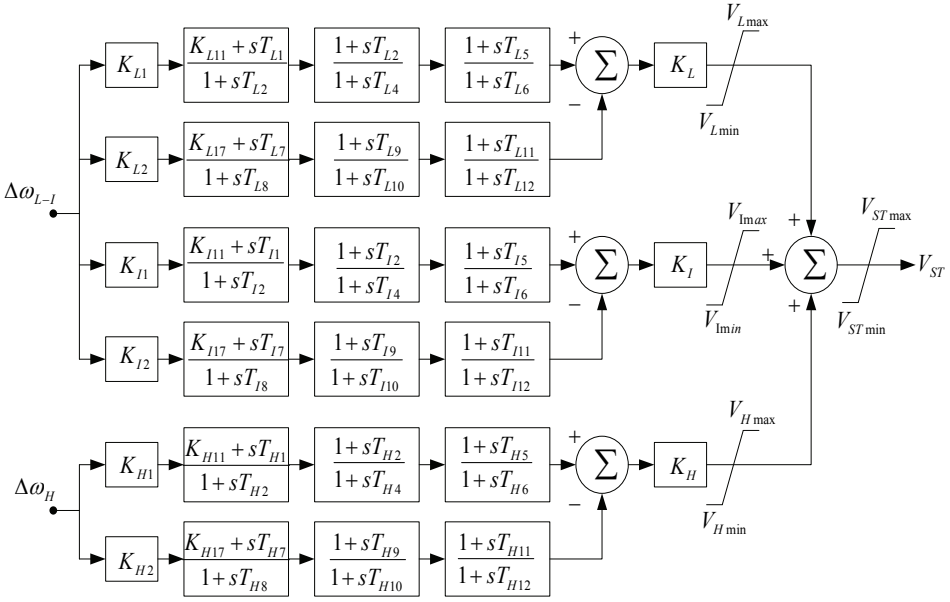


Fig. 2.5. Multi-band power system stabilizer model

Although the conventional stabilizer model has a certain damping effect on the active power oscillation, the action on the special oscillation such as inter-area or global oscillation cannot be considered very well. To solve such oscillation problems, various methods have been provided with the special consideration of the inter-area or global oscillation. Here, the multi-band power system stabilize shown in Fig.2.5 is studied in detail to the inter-area oscillation environment.

In essence, the standardized multi-band power system stabilizer is the multi-structure of the general stabilizer with three kinds of frequency bands action function to consider the mostly potential power system oscillations. In that case, the measured input signal, which has been transferred through high-pass filter sub-block, can be used by the related gain, phase compensation block, and limiter to generate the special output control signal for the local oscillation damping mode. Similarly, the measured input signal, and the related transfer function blocks are used to damp the impossible inter-area and global power oscillation.

3. Eigenvalue Analysis Method

3.1 Small Signal Modelling

The behaviour of a normal power system can be described by a set of first order nonlinear ordinary differential equations and a group of nonlinear algebraic equations. It can be written in the following form by using vector-matrix notation:

$$\begin{aligned} \dot{x} &= f(x,\ w,\ u) \\ 0 &= g(x,\ w,\ u) \\ y &= h(x,\ w,\ u) \end{aligned} \tag{1}$$

In which, x is vector of state variables, such as rotor angle and speed of generators. The column vector w is the vector of bus voltages. u, y is the input and output vector of variables respectively.

Although power system is a nonlinear, it can be linearized by small signal stability at a certain operating point. $(x_0,\ w_0,\ u_0)$ is supposed to be a equilibrium point of this power system, then based on direct feedback, it can be expressed as the following standard form (Zhang et al., 2006):

$$\begin{cases} \Delta\dot{x} = A\Delta x + B\Delta u \\ \Delta y = C\Delta x + D\Delta u \end{cases} \tag{2}$$

where Δx , Δy , and Δu express state, output, and input vector, respectively; A, B, C, and D expresses the state, control or input, output, and feed forward matrices, respectively.

3.2 Damping Ratio and Linear Frequency

The eigenvalues λ of A matrices can be obtained by solving the root of the following characteristic equation:

$$\det(\lambda I - A) = 0 \tag{3}$$

As for any obtained eigenvalues $\lambda_i = \sigma_i \pm j\omega_i$, the damping ratio ρ and oscillation frequency f can be defined as follows:

$$\rho_i = -\frac{\sigma_i}{\sqrt{\sigma_i^{\,2} + \omega_i^{\,2}}} \tag{5}$$

$$f_i = -\frac{\omega_i}{2\pi} \tag{6}$$

The above parameters ρ_i and ω_i can be used to evaluate the damping effects of the power system stabilizers on the power oscillation. It is obvious that the higher damping ratio and the lower oscillation frequency, the better damping effects to enhance the stability of the power system, so as for the solution with power system stabilizers to damp the power oscillation, the best scheme is that install the power system stabilizer for every machine in the power networks, in that case, it can inevitably obtain the best damping effects. Nevertheless, such installation scheme must increase the investment cost, which may be not the economical solution scheme. So, with the precondition of demand damping effects within the specific limits, the optimal arrangement for stabilizers in the areas and the machines of the power networks could be valuably performed with the consideration of economical factor, which will be discussed in the case study.

3.3 Participation Factor

if λ_i is an eigenvalue of A, v_i and w_i are non zero column and row vectors respectively such that the following relations hold:

$$Av_i = \lambda_i v_i,\ i = 1, 2, \cdots, n \tag{7}$$

$$w_i A = \lambda_i w_i,\ i = 1, 2, \cdots, n \tag{8}$$

where, the vectors v_i and w_i are known as right and left eigenvectors of matrix A. And they are henceforth considered normalised such that

$$w_i \cdot v_i = 1 \tag{9}$$

Then the participation factor p_{ki} (the kth state variable x_k in the ith eigenvalue λ_i) can be given as

$$p_{ki} = |v_{ik}||w_{ki}| \tag{10}$$

where w_{ki} and v_{ki} are the ith elements of w_k and v_k, respectively.

4. Cases Study

4.1 Four-Machine Two-Area Test System

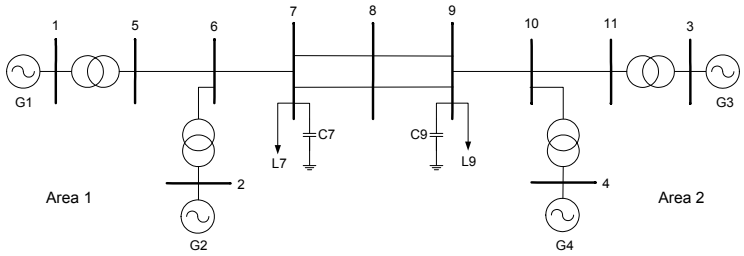
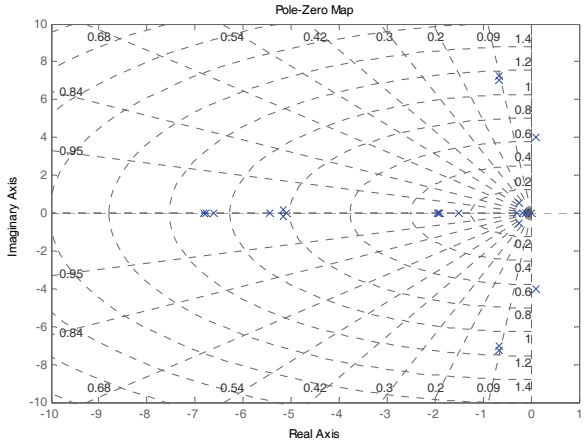
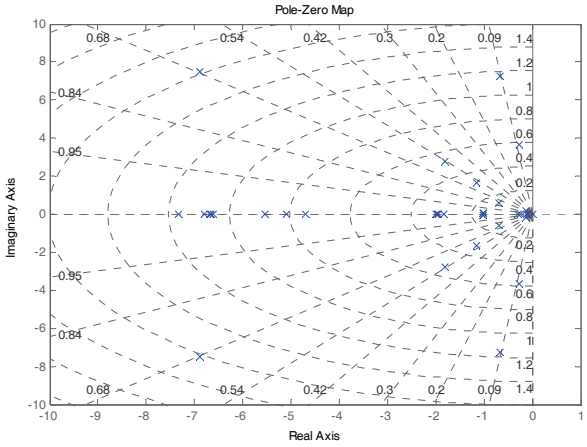


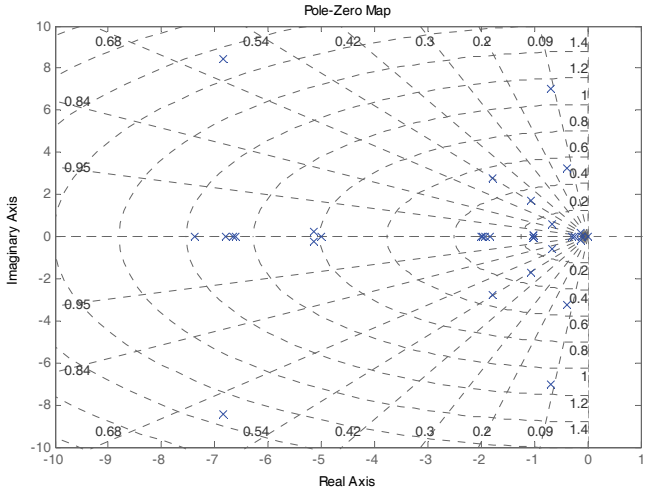
Fig. 4.1. Two-area test system



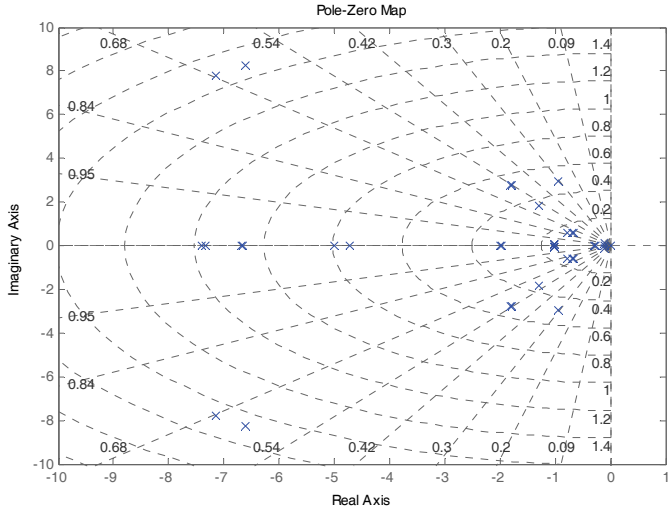
(a)



(b)



(c)



(d)

Fig. 4.2. Dominant eigenvalues of the two-area test system, (a) no stabilizer; (b) with stabilizers in area-1; (c) with stabilizers in area-2; (d) with stabilizers in both areas.

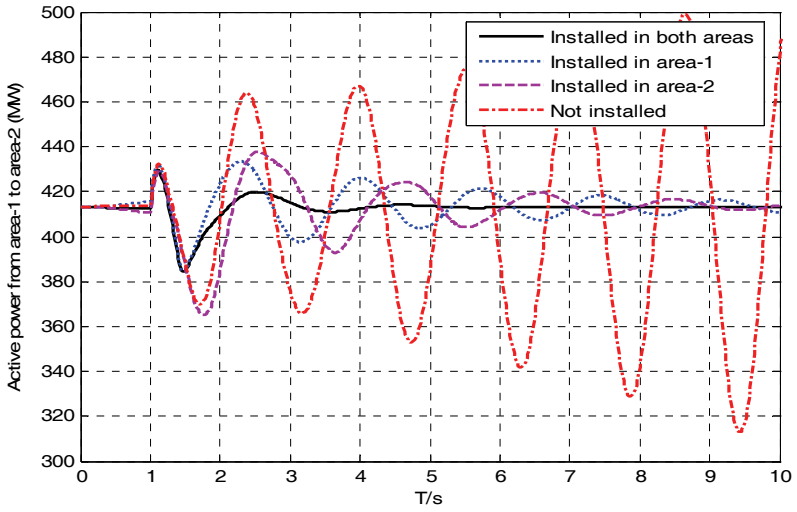
Fig. 4.1 shows the two-area benchmark power system(Kunder,1994) for inter-area oscillation studies. From this, it can be seen that there are two machines in each area, and two-parallel 220km transmission lines are used to interconnect the both areas. In order to discuss the impacts of different stabilizer arrangement on the power oscillation damping, as for the area arrangement scheme, the following tests have been performed: (a) install stabilizers in both areas; (b) install stabilizers in area-1; (c) install stabilizers in area-2; (d) not install stabilizers.

As for the machine arrangement scheme, the following tests have been performed: (a) install stabilizers for G1~G4; (b) install stabilizers for G1 and G3; (c) install stabilizer for G1; (d) no machine installed stabilizer.

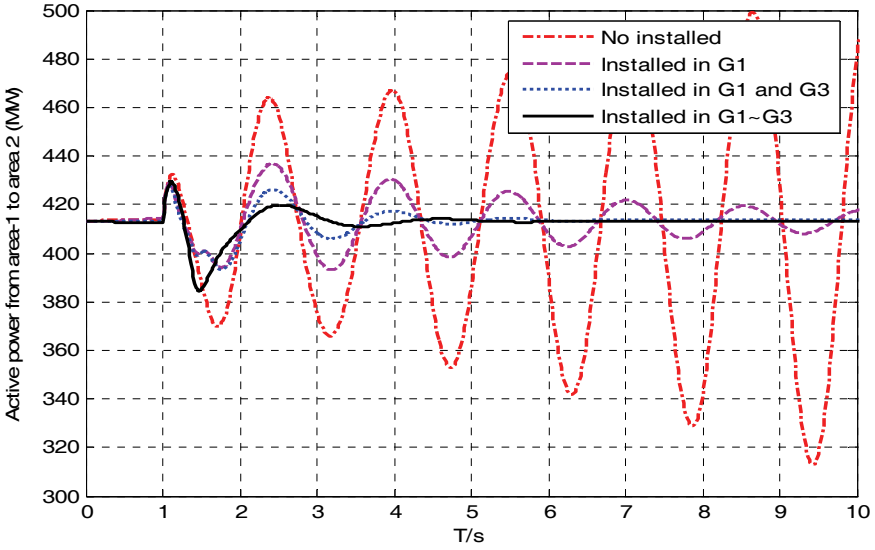
It is worth to remark that such testes mentioned above are achieved by small disturbance for the G1's reference voltage step from 1.0pu to 1.02pu with the duration time of 0.2s. In the corresponding situations, the small signal stability for the inter-area oscillation has been analyzed in detail with the eigenvalues analysis method.

Fig. 4.2 shows dominant eigenvalues analysis results for the two-area test system with different area stabilizer arrangement. From Fig. 4.2(a), it can be seen that as for the open-loop system without any installed stabilizer, there is some instability for the inter-area mode. By installing the stabilizers in area-1, the inter-area oscillation mode has been suppressed, and meanwhile the local mode in area-1 between G1 and G2 is also enhanced greatly shown in Fig.4.2 (b). Such similar damping effect shown in Fig.4.2 (c) is also achieved by installing stabilizers in area-2. If we stall the stabilizers in both area-1 and -2, both the inter-area mode and two local modes can be obtained the high damping ratio and lower oscillation frequency shown in Fig. 4.2(d).

In order to represent the related oscillation effects, the time domain for the test system has been performed. Fig. 4.3 shows the simulation results on the line power flow from area 1 to area 2. From this, it can be found that the arrangement on stabilizer installation for every machine in both areas has the best damping effects on inter-area oscillation, which is in unison with the above dominant eigenvalues analysis results. If there is no any stabilizer for machine in both areas, the inter-area oscillation cannot be avoided. The other arrangement schemes exists a certain difference. By comparative analysis, it can be found that the arrangement scheme on installing the stabilizer for G1 in area-1 and G3 in area-2 is the relative optimal solution to damp the inter-area oscillation between area-1 and 2.



(a)



(b)

Fig. 4.3. Oscillation damping effects of installed stabilizers, (a) different areas; (b) different machines

4.2 Sixteen-Machine Five-Area Test System

In order to indicate the stabilization effects of multi-PSSs for large-scale power system, the 16-machine 5-area test system (Rogers, 2000) shown in Fig.4.4 is simulated in this section. This is in fact the simplified New England and New York interconnected system. The first nine machines (G1-G9) and the second four machines (G10-G13) are belonged to the New England Test System (NETS) and the New York Power System (NYPS), respectively. In addition, there are other three machines (G14-G16) used as the dynamical equivalent of the three neighbour areas connected with NYPS area. It should be remarked that all the machines are described by the sixth-order dynamical model.

The eigenvalue analysis mentioned in the above Section 3 has been performed on the linearized system model of the multi-machine test system without any PSS. The calculated dominate oscillation modes are shown in Table 4.1. From this, it can be seen that as for the system without PSSs, there are kinds of low frequency oscillations (LFOs) with the very weak damping ratios, which is disadvantage to the normal operation of the multi-machine test system.

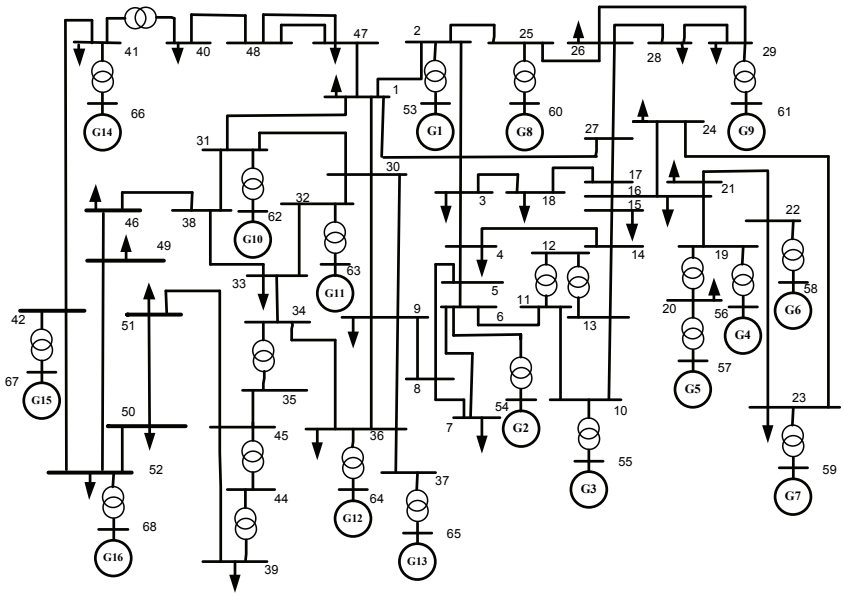


Fig. 4.4. The 16-machine 5-area test system

Mode	Eigenvalues	Frequency (Hz)	Damping Ratio
1	-0.064±2.756i	0.439	0.023
2	-0.032±3.590i	0.571	0.009
3	-0.003±4.408i	0.702	0.001
4	-0.131±5.170i	0.823	0.025
5	0.408±7.673i	1.221	-0.053
6	0.240±7.722i	1.229	-0.031
7	0.647±7.790i	1.240	-0.083
8	-0.157±8.357i	1.330	0.019
9	0.291±8.462i	1.347	-0.034
10	0.477±8.615i	1.371	-0.055
11	0.167±8.690i	1.383	-0.019
12	-0.116±10.095i	1.607	0.012
13	0.092±10.188i	1.621	-0.009
14	-0.383±10.207i	1.625	0.037
15	0.516±12.543i	1.996	-0.041

Table 4.1. Dominant oscillation modes (without PSS)

Furthermore, according to the calculation results shown in Table 4.2 about the participation factor of each machine to the corresponding operation mode, it can be observed that under the normal operation condition, the system mainly has four inter-area oscillation modes and eleven local oscillation modes. Combined to Table 4.1, we can obviously obtain the common results about the LFO characteristics. That is to say, as for the inter-area modes, the oscillation frequency is less 1.0Hz, and as for the low-frequency local modes, the oscillation frequency is between 1.0Hz and 2.0Hz.

Mode	Participation factor (from G1 to G16)	Oscillation mode
1	0.0147,0.0110,0.0137,0.0127,0.0130,0.0167,0.0120,0.0093, 0.0142,0.0049,0.0046,0.0253,0.1400,0.0848,0.1006,0.0366	G1-G9 vs G10-G16
2	0.0022,0.0012,0.0017,0.0021,0.0023,0.0029,0.0020,0.0014, 0.0024,0.0001,0.0000,0.0001,0.0002,0.2065,0.0063,0.2730	G1,G4-G9,G14 vs G2,G3, G10-G13,G15,G16
3	0.0309,0.0141,0.0214,0.0383,0.0461,0.0546,0.0366,0.0191, 0.0392,0.0000,0.0006,0.0205,0.1746,0.0029,0.0003,0.0056	G1,G4-G8 vs G2,G3, G9-G16
4	0.0000,0.0000,0.0000,0.0001,0.0001,0.0001,0.0001,0.0000, 0.0000,0.0000,0.0000,0.0003,0.0029,0.1322,0.3144,0.0498	G1-G9,G12,G13,G15 vs G10,G11,G14,G16
5	0.0001,0.0059,0.0038,0.0037,0.0048,0.0102,0.0039,0.0008, 0.0140,0.0013,0.0006,0.4178,0.0681,0.0000,0.0000,0.0001	Local oscillation mode
6	0.0176,0.0979,0.0734,0.0429,0.0565,0.1126,0.0381,0.0040, 0.0758,0.0103,0.0010,0.0422,0.0017,0.0001,0.0000,0.0000	Local Oscillation mode
7	0.0123,0.0898,0.0728,0.0015,0.0014,0.0038,0.0029,0.0111, 0.3152,0.0000,0.0000,0.0019,0.0004,0.0000,0.0000,0.0000	Local Oscillation mode
8	0.0055,0.0005,0.0007,0.0167,0.2878,0.1593,0.0536,0.0031, 0.0005,0.0019,0.0001,0.0001,0.0001,0.0000,0.0000,0.0000	Local Oscillation mode
9	0.1377,0.0427,0.0108,0.0007,0.0023,0.0102,0.0082,0.0783, 0.0166,0.1799,0.0043,0.0077,0.0013,0.0002,0.0000,0.0002	Local Oscillation mode
10	0.0027,0.1864,0.2582,0.0001,0.0002,0.0015,0.0007,0.0014, 0.0006,0.0018,0.0000,0.0000,0.0000,0.0000,0.0000,0.0000	Local Oscillation mode
11	0.1213,0.0008,0.0035,0.0001,0.0004,0.0084,0.0012,0.0710, 0.0047,0.2569,0.0032,0.0029,0.0021,0.0001,0.0000,0.0002	Local Oscillation mode
12	0.0025,0.0001,0.0005,0.1471,0.0727,0.1056,0.1381,0.0007, 0.0003,0.0000,0.0000,0.0000,0.0000,0.0000,0.0000,0.0000	Local Oscillation mode
13	0.0001,0.0000,0.0000,0.1957,0.0345,0.0365,0.1852,0.0001, 0.0000,0.0000,0.0000,0.0000,0.0000,0.0000,0.0000,0.0000	Local Oscillation mode
14	0.1793,0.0000,0.0001,0.0001,0.0001,0.0002,0.0001,0.3086, 0.0000,0.0000,0.0000,0.0000,0.0000,0.0000,0.0000,0.0000	Local Oscillation mode
15	0.0007,0.0002,0.0001,0.0001,0.0000,0.0000,0.0000,0.0002, 0.0000,0.0100,0.4323,0.0021,0.0043,0.0001,0.0000,0.0002	Local Oscillation mode

Table 4.2. Participation factor and oscillation modes

To describe the existing oscillation modes vividly, the angle eigenvectors for mode 1-8 have been drawn as shown in Figure 4.5 and 4.6. By comparing there two figures, we can obviously find the difference between inter-area mode and local mode. In practice, with the demand of competitive power markets and the large-scale transmission and distribution of electric energy, more and more regional electric networks are interconnected to gradually form the relative bigger scale electric power systems. In that case, the dynamic performance changes more complex, which lead to various instability problems such as voltage instability, power oscillations, and so on. Especial for the inter-area oscillation mode, it could be the typical LFO modes existing in the modern power system, which should be considered carefully. Generally, as for the typical common selection for stabilization of power system, PSS can provide a certain damping for the LFO mode especial for the local mode. Also, as for the inter-area oscillation damping, the multi-band PSS mentioned in the above section should be a better alternative.

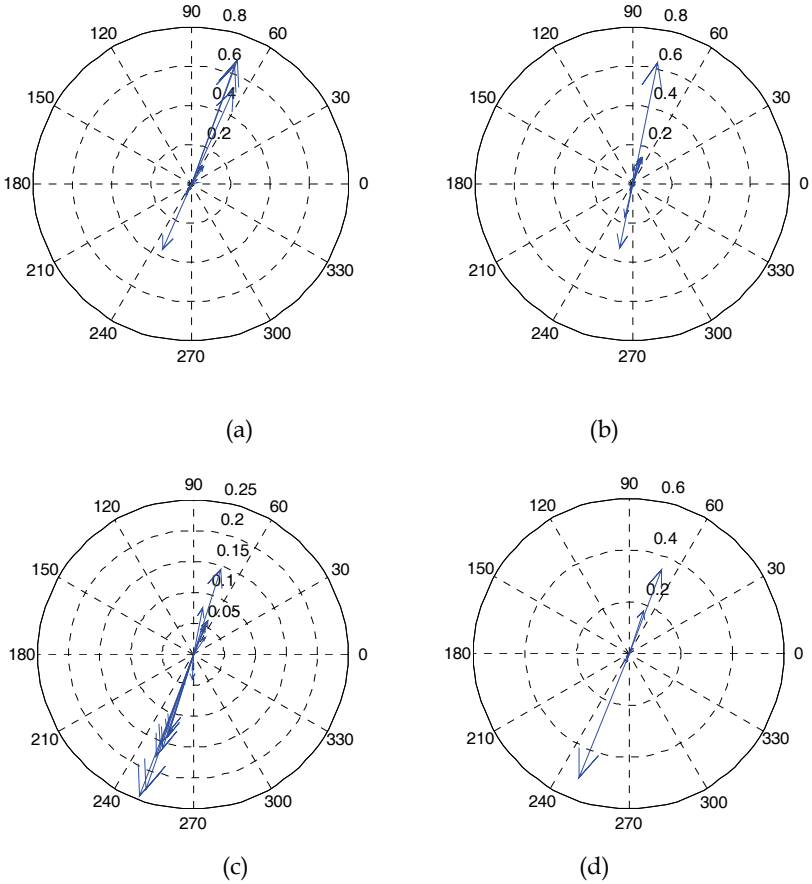


Fig. 4.5. Inter-area oscillation modes. (a) mode-1, (b) mode-2, (c) mode-3, (d) mode-4

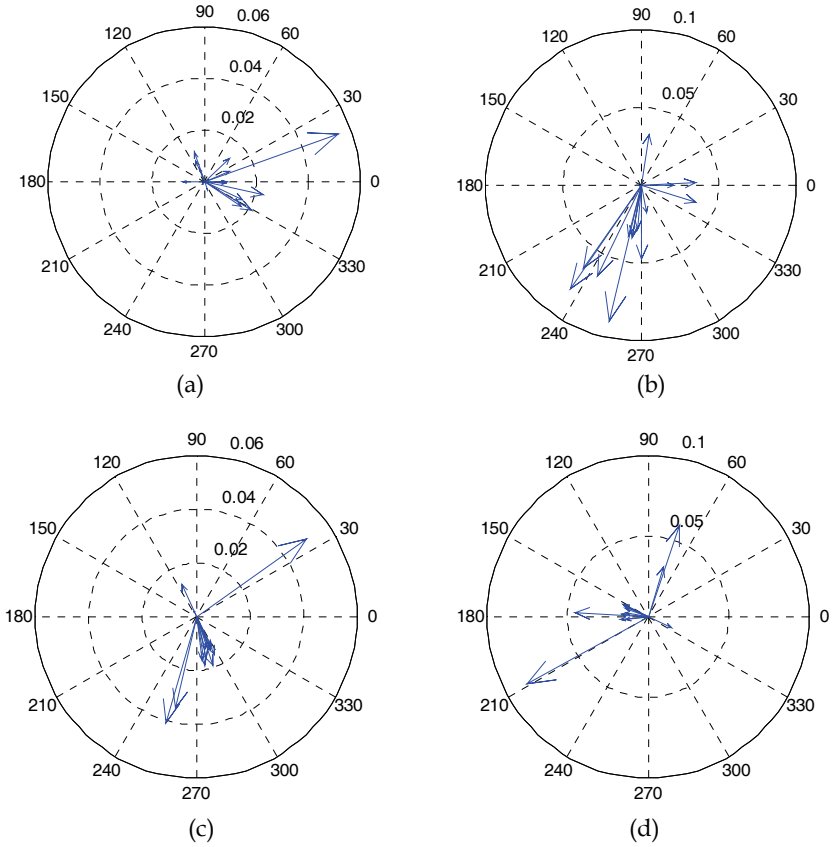


Fig. 4.6. Local oscillation modes. (a) mode-5, (b) mode-6, (c) mode-7, (d) mode-8

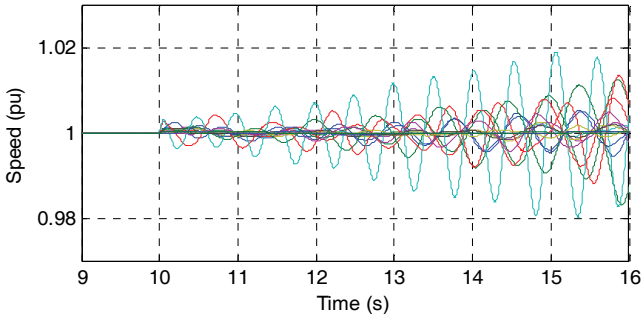
In order to reveal the stabilizing effects of the general PSS on the LFO modes, the PSS with the classical structure shown in Figure 2.3 has been installed to each machine in the multi-machine test system. As for the lead-lag time constants for the phase lag compensation, they can be determined using the method given in (Kunder, 1994; Rogers, 2000).

The eigenvalues analysis for the linearized model of the multi-machine test system with 16-PSSs has been performed. The calculated dominate oscillation modes are shown in Table 4.3. By comparing with Table 4.1, it can be seen that with the implement of PSSs installation, the damping ratios for both the inter-area and the local modes are greater than 0.1, which indicates the very well stabilization effects of PSS on LFO mode.

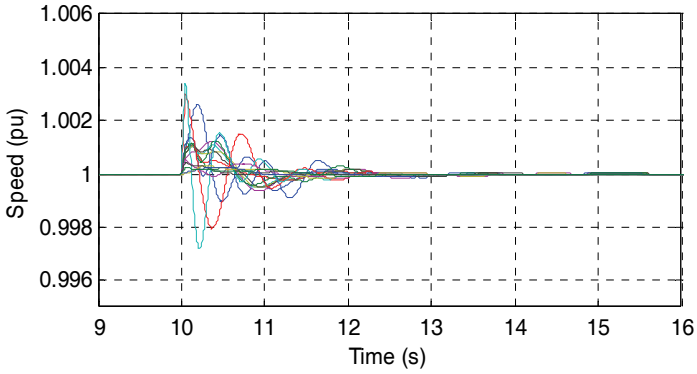
Mode	Eigenvalues	Frequency (Hz)	Damping Ratio
1	-0.598±2.667i	0.424	0.219
2	-0.690±3.489i	0.555	0.194
3	-0.676±4.209i	0.670	0.159

4	-0.674±5.037i	0.802	0.133
5	-1.205±7.434i	1.183	0.160
6	-1.494±7.543i	1.201	0.194
7	-1.560±8.152i	1.297	0.188
8	-1.810±8.337i	1.327	0.212
9	-1.738±8.540i	1.359	0.199
10	-1.271±8.622i	1.372	0.146
11	-2.767±8.879i	1.413	0.297
12	-1.724±11.114i	1.769	0.153
13	-2.913±11.414i	1.817	0.247
14	-3.024±11.822i	1.882	0.248
15	-2.056±12.353i	1.966	0.164

Table 4.3. Dominant oscillation modes (with PSS)

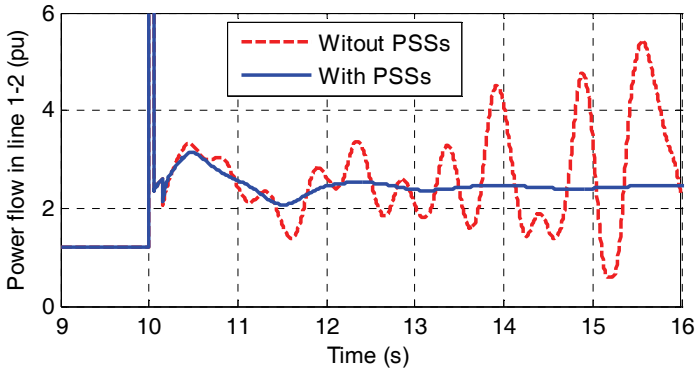


(a)

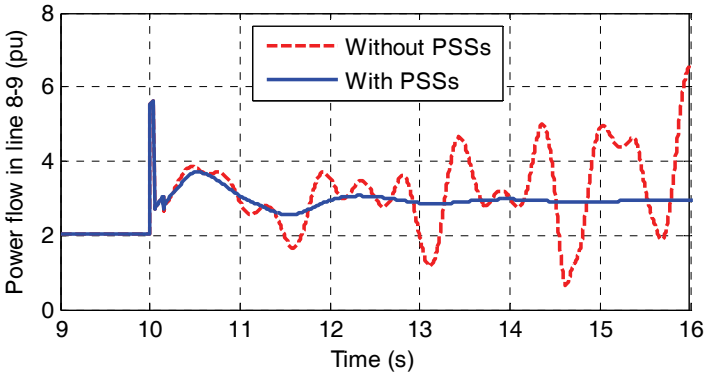


(b)

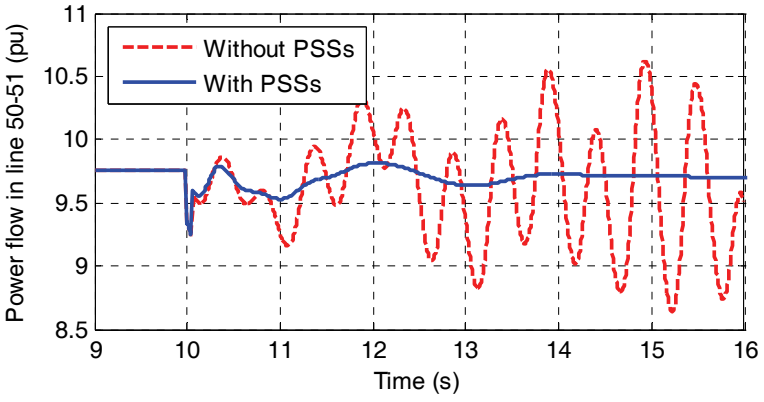
Fig. 4.7. Dynamic response of speed of G1-G16 to a line-to-ground fault. (a) without PSSs, (b) with PSSs.


Witout PSSs
With PSSs
Power flow in line 1-2 (pu)
Time (s)
0
2
4
6
9
10
11
12
13
14
15
16

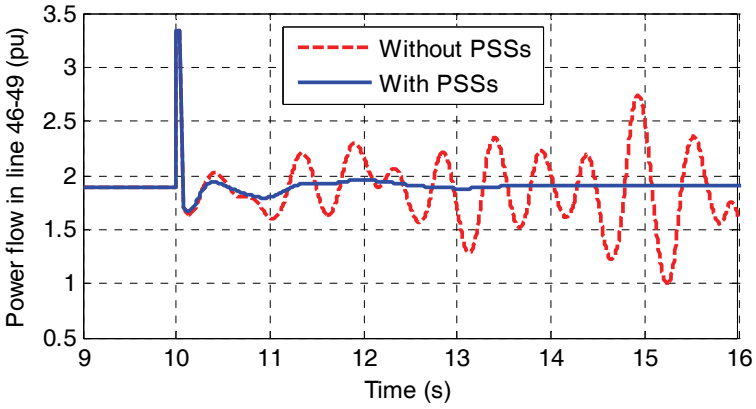
(a)


Without PSSs
With PSSs
Power flow in line 8-9 (pu)
Time (s)
0
2
4
6
8
9
10
11
12
13
14
15
16

(b)


Without PSSs
With PSSs
Power flow in line 50-51 (pu)
Time (s)
8.5
9
9.5
10
10.5
11
9
10
11
12
13
14
15
16

(c)



(d)

Fig. 4.8. Dynamic response of power flow in the interconnected backbone lines. (a) line 1-2, (b) line 8-9, (c) line 50-51, (d) line 46-49.

In order to evaluate the stabilization performance of PSSs on the LFO modes, the nonlinear simulation on the multi-machine test system has been performed by setting the line-to-ground fault nearby Bus-1. The fault starts from 10.0-s and continuous 50-ms. Figure 4.7 and 4.8 show the dynamic responses of the test system for such large disturbance. From Figure 4.7, it can be seen that the system without PSSs exists the serious power oscillations, which is directly reflected by the instability of machine speed shown in Figure 4.7(a). However, with the implement of PSSs, such oscillations are damped very well, which can be shown in Figure 4.7(b).

Furthermore, as for the multi-machine test system with five areas, the power flow in the backbone lines, which play an important role on the network interconnection, can be obtained as shown in Figure 4.8. From this, it can be seen that, the installation of PSSs can improve the system dynamic performance very well, and all the backbone lines can transmit the power stably.

5. Conclusion

This paper presents the power system stabilizer with the consideration of local, inter-area, and global mode to damp the potential power oscillation. Based on this, the eigenvalues analysis method has been introduced to analyze the damping effects of various arrangement schemes of such stabilizer. The case study on the typical 4-machines 2-area test system and 16-machines 5-areas shows that although the best arrangement scheme that install the stabilizer for every machine and area can obtain the best oscillation damping effect, it is not the economical solution scheme especial to the large power networks, and the scheme that arrange stabilizer for one area one machine is the optimal arrangement with the consideration of economical factor. This paper has a certain meaning to the optimal stabilizer arrangement for power networks, and the future researches on the arrangement rules with evolutionary algorithm and the coordinated FACTS device to obtain the better power oscillation damping effects could be concerned and performed.

6. References

Kunder, P. (1994). *Power System Stability and Control,* McGraw-Hill, 0-07-035958-X, New York, USA.

Anderson, P. M.; & Fouad, A. A. (2003). *Power System Control and Stability(Second Edition),* John Wiley & Sons, 978-0-471-23862-1, Manhattan, USA.

Bikash, P.; & Balarko, C. (2005). *Robust Control in Power Systems,* Springer Science & Business Media, 0-387-25949-X, New York, USA.

Schleif, F. R.; & White, J. H. (1966). Damping for the northwest-southwest tieline oscillations -an analogue study. *IEEE Trans. Power Appar. Syst.,* Vol. 85, No. 12, (Dec. 1966), 1239-1247, 0018-9510.

Zhang, X.P.; Rehtanz, C.; & Pal, B.(2006). *Flexible AC Transmission Systems: Modeling and* Control, Springer, 978-3540306061, Germany.

Rogers, G.(2000). *Power System Oscillation,* Kluwer Academic, 978-0-7923-7712-2, Norwell, USA.

Ray, S.; & Venayagamoorthy, G. K. (2008). Real-time implementation of a measurement-based adaptive wide-area control system considering communication delays. *IET Gener. Transm. Distrib.,* Vol. 2, No. 1, (Jan. 2008), 62-70, 1751-8687.

Kamwa, I.; & Grondin, R. (2002). PMU configuration for system dynamic performance measurement in large, multiarea power systems. *IEEE Trans. Power Syst.,* Vol.17, No. 2, (Mar. 2002), 59-69, 0272-1724.

Pourbeik, P.; & Gibbard, M. J. (1996). Damping and synchronizing torques induced on generators by FACTS stabilizers in multi-machine power systems. *IEEE Trans. Power Syst.,* Vol.11, No. 4, (Nov. 1996), 1920-1925, 0885-8950.

Pourbeik, P.; & Gibbard, M. J. (1998). Simultaneous coordination of power system stabilizers and FACTS device stabilizers in a multi-machine power system for enhancing dynamic performance. *IEEE Trans. Power Syst.,* Vol.13, No. 2 (May. 1998), 473-479, 0885-8950.

De Oliveira, R.; Ramos, R.; & Bretas, N.(2007). A mixed procedure based on classical and modern control to design robust damping controllers. *IEEE Trans. Power Syst.,* Vol. 22, No. 3 (Aug. 2007), 1231-1239, 0885-8950.

Pal, B. C.; Coonick, A.H.; & Cory, B.J.(1999). Robust damping of inter-area oscillations in power systems with superconducting magnetic energy storage devices. *IET Gener. Transm. Distrib.,* Vol. 146, No. 6, (Nov. 1999), 633-639, 1350-2360.

5

Network and System Simulation Tools for Next Generation Networks: A Case Study[1]

S. Mehta, Mst. Najnin Sulatan and K. S. Kwak
Inha University
Korea

1. Introduction

The modern information society will continue to emerge, and demand for wireless communication services will grow. Future generation wireless networks are considered necessary for the support of emerging services with their increasing requirements. Future generation wireless networks are characterized by a distributed, dynamic, self-organizing architecture (I. F. Akyildiz et al., 2006). These wireless networks are broadly categorized into different wireless networks according to their specific characteristics. Typical examples include Ad-Hoc/Mesh Networks, Sensor Networks, Cognitive Radio Networks, etc as shown in figure 1. These wireless networks could then constitute the infrastructure of numerous applications such as emergency and health-care systems, military, gaming, advertisements, customer-to-customer applications, etc. Not only their importance in military applications is growing, but also their impact on business is increasing. The emergence of these wireless networks created many open issues in network design too. More and more researchers are putting their efforts in designing the future generation wireless networks.

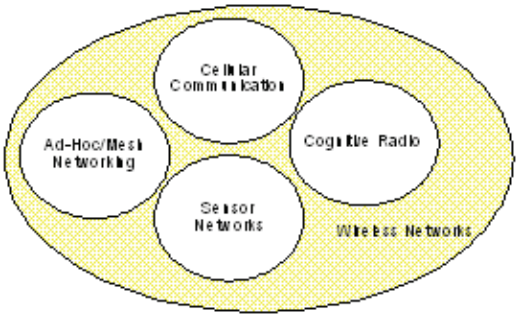


Fig. 1. Different kinds of wireless networks

[1] Some part of this chapter was published at AMS'09, Indonesia [24].

There are three main traditional techniques for analyzing the performance of wired and wireless networks; analytical methods, computer simulation, and physical measurement or a testbed measurement. Traditionally, formal modeling of systems has been via a mathematical model, which attempts to find analytical solutions to problems and thereby enable the prediction of the behavior of the system from a set of parameters and initial conditions. However, it is widely known that comprehensive models for wireless ad hoc networks are mathematically intractable. On its own, each individual layer of the protocol stack may be complex enough to discourage attempts at mathematical analysis. Interactions between layers in the protocol stack magnify this complexity. The construction of real testbeds for any predefined scenario is usually an expensive or even impossible task, if factors like mobility, testing area, etc. come into account. Additionally, most measurements are not repeatable and require a high effort.

Simulation is, therefore, the most common approach to developing and testing new protocol for a wireless network. Simulation has proven to be a valuable tool in many areas where analytical methods aren't applicable and experimentation isn't feasible. Researchers generally use simulation to analyze system performance prior to physical design or to compare multiple alternatives over a wide range of conditions. In context with networks, and especially wireless networks, simulators are used for the development and validation of new algorithms, such as routing algorithms in wireless networks, or protocols. Improvements of existing algorithms, as well as testing a networks capacity and efficiency under specific scenarios is also a simulators task. Many publications typically include performance simulations and commonly compare routing protocols. Simulators model the real world in a specific way. Their purpose is to ease the understanding of it, to surge its behavior and especially research its reactions on particular events. There are a number of advantages to this approach: lower cost, ease of implementation, and practicality of testing large-scale networks [2].

The goal of simulators is to achieve an "as real as possible" situation in order to make the simulation results realistic and therefore adaptable. Because it is impossible to collect and implement all the data and details playing a role within the real world, the simulators have to be trimmed. Now, the main difficulty is where to start cutting off details and where to end with it while dealing with simulation. The correct level of detail decides whether a simulation is useful or not, and therefore a difficult part in the development process. While less details in simulation could produce results which are deluding or in some cases even false, the effects of too many details can also make the simulation useless: Necessarily the implementation is more time- consuming and the simulation takes longer. When it comes to wireless network simulation, three main points are important: Firstly, the algorithms and protocols should be error free and have to be implemented in adequate detail, and secondly, the simulation environment, such as mobility schemes, must be realistic. Finally, a proper method is needed to analyze the collected data. Even though simulation is a powerful tool, it is still occupied with potential pitfalls [3]. To help overcome this, it is important to know the different tools available and their benefit and drawbacks there in associated. The goal of this paper is to give an over all short review to simulation system, especially discussion about commonly used simulation tools in system and network, and a cautionary guideline to avoid the pitfall associated with simulation for all who are using or will be using simulation tools for their research.

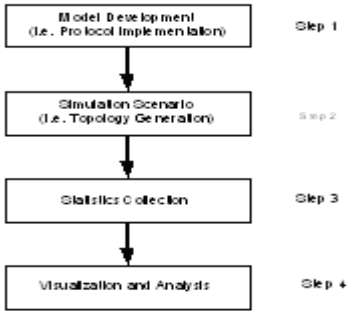


Fig. 2. Simulation Steps

There are four basic steps to run a simulation as shown in figure 2. First step is to develop a model (e.g. implementation of a protocol); second step is to create a simulation scenario (e.g. designing a network topology and traffic scenario); third step is to choose and collection of statistics, and finally fourth step is to visualize and analysis of simulation results which may be carried out after (or during, in some cases) the simulation execution. The problem with such approach is that one cannot guess in advance how many replications is needed for securing small errors of estimates, and if the errors are found to be too large, simulations need to be repeated. This is referred to as offline sequential analysis of simulation output data. Of course, this is not a very efficient way of data analysis. It is generally required that final results from any simulation are to analyze output data on-line, during simulation. Then, the simulation can be stopped when the statistical errors of the estimates become sufficiently small.

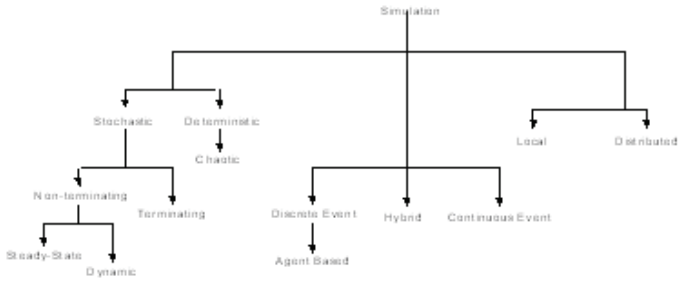


Fig. 2**.** Classification of simulation tools

Before going further, we present the classification of simulation tools which could be a good summary for those with little previous exposure to the topic [4]. As depicted in figure 3 simulation tools can be classified according to several criteria including:

- Stochastic or deterministic
- Steady State or dynamic
- Terminating or non terminating
- Discrete or continuous or hybrid
- Local or distributed

Stochastic simulation: Most of the realistic simulation tools The world is full of uncertainty, and most (if not all) realistic simulation models will incorporate some randomness as well as some element of time elapsing. Such tools can be used to examine a diverse set of applications. For example, the simulation may have been designed to model the operation of a customer service center, traffic patterns over a particular location grid, hospital facilities utilization, waiting times for customers arriving at a service center, the number of cars passing through an intersection during a 5 minute period, the efficacy of various strategies in combat warfare, the impact of changes in layout and equipment on production throughput, and more.

Deterministic simulation: Deterministic simulations use fixed, non-random values to specify the model and particular variant of the system under investigation. Because there is no randomness, the output is also fixed for any specific set of inputs. Chaotic model is the special case of deterministic model.

Non-terminating simulation: In a non-terminating system, the duration of the system is not finite. The Internet exemplifies a non-terminating system. Non-terminating simulations are used to simulate non-terminating systems. In a non-terminating simulation, there is no event to signal the end of a simulation, and such simulations are typically used to investigate the long-term behavior of a system. Non-terminating simulations must, of course, stop at some point, and it is a non-trivial problem to determine the proper duration of a non-terminating simulation. If the behavior of the system becomes fairly stable at some point, then there are techniques for analyzing the steady-state behavior of the system using non-terminating simulations.

Terminating simulation: Terminating systems are characterized by having a fixed starting condition and a naturally occurring event that marks the end of the system. An example of a terminating system is a work day that starts at 8 am and ends at 4 pm at a bank. For terminating systems the initial conditions of the system generally affect the desired measures of performance. The purpose of simulating terminating systems is to understand their behavior during a certain period of time, and this is also referred to as studying the transient behavior of the system.

Steady-State simulation: Steady-state models use equations defining the relationships between elements of the modeled system and attempt to find a state in which the system is in equilibrium. Such models are often used in simulating physical systems, as a simpler modeling case before dynamic simulation is attempted.

Dynamic simulation: Dynamic simulations model changes in a system in response to (usually changing) input signals.

Discrite event simulation: A discrete event simulation manages events in time. Most computer, logic-test and fault-tree simulations are of this type. In this type of simulation, the simulator maintains a queue of events sorted by the simulated time they should occur. The simulator reads the queue and triggers new events as each event is processed. It is not important to execute the simulation in real time. It's often more important to be able to access the data produced by the simulation, to discover logic defects in the design, or the sequence of events. Most of the network simulation tools fall under this category.

Agent-Based Simulators: This is a special class of discrete event simulator in which the mobile entities are known as agents. Whereas in a traditional discrete event model the entities only have attributes, agents have both attributes and methods (e.g., rules for

interacting with other agents). An agent-based model could, for example, simulate the behavior of a population of animals that are interacting with each other.

Continuous Simulators: This class of tools solves differential equations that describe the evolution of a system using continuous equations. These types of simulators are most appropriate if the material or information that is being simulated can be described as evolving or moving smoothly and continuously, rather than in infrequent discrete steps or packets. For example, simulation of the movement of water through a series of reservoirs and pipes can most appropriately be represented using a continuous simulator. Continuous simulators can also be used to simulate systems consisting of discrete entities if the number of entities is large so that the movement can be treated as a flow.

Hybrid Simulators: These tools combine the features of continuous simulators and discrete simulators. That is, they solve differential equations, but can superimpose discrete events on the continuously varying system.

Distributed simulator: Distributed models run on a network of interconnected computers, possibly through the Internet. Simulations dispersed across multiple host computers like this are often referred to as distributed simulations.

Local simulator: Local simulator models run on an individual machine or within an interconnected cluster.

2. Related Work

There are several surveys, comparisons, and also some case studies about wireless network and system simulators. They all differ with respect to the selection of evaluated simulators, the intention of the work, the focus of the potential comparison and the level of detail. Table 1 summarizes the previous related works.

Reference	Type of Study	Simulator Tools	Scope of Study
[5]	Comparison	Opnet, ns-2	Initialization, accuracy
[6]	Comparison	Opnet,ns-2, QualNet, OMNeT++, JSim, SSFNet	For critical infrastructure
[7]	Comparison	ns-2, TOSSIM	Models, visualization, architecture, components
[8]	Description	GloMoSim, ns-2, DIANEmu, GTNetS, J-Sim, Jane, NAB, PDNS, OMNeT++, Opnet, QualNet, SWANS	Overview
[9]	Comparison	SSF, SWANS, J-Sim, NCTUns, ns-2, OMNeT++, Ptolemy, ATEMU, Em-Star, SNAP, TOSSIM	Models, type of visualization
[10]	Description	ns-2, GloMoSim, Opnet, SensorSim, J-Sim, Sense, OMNeT++, Sidh, Sens, TOSSIM, ATEMU, Avrora, EmStar	Overview
[11]	Comparison	Opnet, ns-2, OMNeT++, SSFNet, QualNet, J-Sim, Totem	availability/credibility of models, usability
[12]	Case Study	Opnet, ns-2, testbed	Accuracy of results
[13]	Survey	In general simulation study	Credibility , accuracy

Table 1. Related works on Simulator comparison

All of the works listed in table 1 consider different simulators or differ in their aim from this paper. The works parented in [6, 8, 9, 10, 11, 13] are the close to our work as they include some common simulators J-Sim, OMNeT++, and ns-2, which we also consider for our study. However, [6] examines their suitability for simulating the failure of critical infrastructures like electricity or telecommunication networks. This is very unrelated to what we present here. A huge list of simulators is presented in [8, 10] however, they do not give a comparative study. Rather, their works consists of more or less description of each simulator tools independently. In [9] authors give an overview about the different issues in wireless networks on a general basis. Only at the end of their work they presented a table comparing the considered simulation tools according to different features such as their language, the available modules, and GUI support, etc. the most detailed comparison is presented in [11]. However, they consider all the simulators from an industrial research point of view, which are less relevant for academic researchers. They also miss several practical issues regarding the credibility and reliability of the tools. In [13] authors presented a survey study of more then 2200 research papers in the field of network simulation studies and point out several systematic flaws in that. We follow the similar kind of work line of [13] but with different aims. Our goal in this paper is to make a basic contribution to the wireless network community by a) Giving overall short overview of some widely used system and network simulation tools, b) comparing simulation tools on the basis of several features and a survey report of more then 800 research papers in the field of system and networks in recent years (2000~2008), c) listing our recommendations for the designers of protocols, models, and simulators.

The remainder of this paper is organized as follows. In Section 3, we provide a brief overview on "widely used" network and system simulation tools, and their comparisons and results from our survey. Finally, conclusions are presented in section 4.

3. System and Network Simulation Tools

For network protocol designers, it is often difficult to decide which simulator to choose for a particular task, especially for NGNs. Therefore, we conduct a survey to find a wireless system/network simulator that provides a good balance between availability of ready to use models, scripting and language support, extendibility, graphical support, easiness of use, etc. The survey is based on a collection of a number of criteria including published results, interesting characteristics and features. From our survey results, we broadly categories system and network simulators as: "Widely Used" simulators and "Other" simulators. We discuss more about these two categories in the later sections of this paper. The network simulators taken into consideration as "Widely Used"are Ns-2, GloMoSim, J-Sim, OMNet++, OPNet, and QualNet. While for physical layer[2], very few simulator tools are used. The theoretical, numerical, statistical analyses are vastly used for physical layer. As a simulator, NS2, TOSSIM, GloMoSim, Qualnet, OPNET, OMNET++ are used for the implementation of the error model at the physical layer [14]. However, none of them appears to implement sub-channels or to finely model an explicit packet detection and timing synchronization phase [15]. MATLAB and Monte Carlo based simulations are

[2] In rest of the paper we keep using terms "Physical Layer and "System" interchangeably, unless and otherwise specified.

"Widely Used" simulators in case of system. In this section, we will give a short overview and a comparative study about the eight "widely used" network and system simulators, respectively.

Ns-2: Ns2 is a discrete event simulator for networks. It began as ns (Network Simulator) in 1989 with the purpose of general network simulation. The core of core of the simulator and most of the network protocol models are written in C ++, and the rest is in OTcl. In general, C++ is used for implementing protocols and extending the ns-2 library. OTcl is used to create and control the simulation environment itself, including the selection of output data. Simulation is run at the packet level, allowing for detailed results. Ns2 provides OSI layers excluding presentation and session layers. It has a huge pool of available features, offering a large number of external protocols already implemented. Ns-2 does not scale well for sensor networks. This is in part due to its object-oriented design. While this is beneficial in terms of extensibility and organization, it is a hindrance on performance in environments with large numbers of nodes. Another drawback to ns-2 is the lack of customization available. Packet formats, energy models, MAC protocols, and the sensing hardware models all differ from those found in most wireless devices [16].

GloMoSim: GloMoSim was developed in 1986 fro mobile wireless networks at UCLA (California, USA). GloMoSim is written in Persec, which is an extension of C for parallel programming. New protocols and modules for GloMoSim must be written in Parsec too. GloMoSim respects the OSI standard. The ability to use GloMoSim in a parallel environment distinguishes it from most other wireless network simulators. Like ns-2, GloMoSim is designed to be extensible, with all protocols implemented as modules in the GloMoSim library. GloMoSim still contains a number of problems. While effective for simulating IP networks, it is not capable of simulating any other type of network. This effectively ensures that many wireless networks can not be simulated accurately. Additionally, GloMoSim does not support phenomena occurring outside of the simulation environment, all events must be generated from another node in the network. Finally, GloMoSim stopped releasing updates in 2000. Instead, it is now updated as a commercial product called QualNet [17].

J-SIM: J-Sim is a general purpose java based simulator developed by a team at the Distributed Realtime Computing Laboratory (DECL) of the Ohio State University. It is built according to the component-based software paradigm and written in Java. Everything in J-Sim is a component: a node, a link, a protocol. Each component can be atomic or composed of other components. Connection between components is done through ports. Actually, there are three possible ways to connect ports: one-to-one, one-to-many, and many-to-many. On a more abstract level, J-Sim distinguishes two layers. The lower layer Core Service Layer (CSL) comprises every OSI layer from network to physical, the higher layer comprises the remaining OSI layers. Initially designed for wired network simulation, its Wireless extension proposes an implementation of the IEEE 802.11 MAC—which is the only MAC supported so far. This extension turns J-Sim to a viable MANETs simulator. J-Sim also features a set of components which facilitates basic studies of wireless/mobile networks, including three distinct radio propagation models and two stochastic mobility models. J-Sim works on any operating system that turns Sun's Java SDK 1.5 or later and is open source [18].

OMNet++ : OMNet++ (Object Modular network Testbed in C++) is well designed discrete event simulation environment written in C++. OMNET++ is actually a general-purpose simulator capable of simulating any system composed of devices interacting with each

others. The mobility extension for OMNeT++ is intended to support wireless and mobile simulations within OMNeT++. This support is said to be fairly incomplete. OMNet++ is for academic and educational use. Modules are connected in a hierarchical nested fashion, where each module can contain several other modules. Modules can be defined as being either simple or compound. Simple modules are used to define algorithms, and make up the bottom of the hierarchy. Compound modules are a collection of simple modules that interact with one another, using messages. OMNeT++ provides a component-based, hierarchical, modular and extensible architecture. Components, or modules, are programmed in C++ and new ones are developed using the C++ class library which consists of the simulation kernel and utility classes for random number generation, statistics collection, topology discovery etc. OMNeT++ has a number of advantages over the other simulators. OMNeT++ accurately models most hardware and includes the modeling of physical phenomena. All layers of the protocol stack can be modified. Despite its apparent advantages, OMNeT++ has remained relatively obscure. The original implementation does not offer a great variety of protocols, and very few have been implemented, leaving users with significant background work if they want to test their own protocol in different environments. OMNet++ works on Linux, Unix-like systems and windows XP/2K [19].

OPNet: OPNet (Optimized Network Engineering Tools) Modeler is a discrete-event network simulator first proposed by MIT in 1986 and is written in C++. It is a well established and professional commercial suite for network simulation. It is actually the most widely used commercial simulation environment. However, it can be used free of charge by researchers applying to University Program of the product. Unlike ns-2 and GloMoSim, OPNET supports the use of modeling different network-specific hardware, such as physical-link transceivers and antennas. OPNET Modeler features an interactive development environment allowing the design and study of networks, devices, protocols, and applications. For this, an extensive list of protocols is supported. Particularly, MAC protocols include IEEE 802.11a/b/g and Bluetooth ones.OPNET can also be used to define custom packet formats. The simulator aids users in developing the various models through a graphical interface. The interface can also be used to model, graph, and animate the resulting output. One of the most interesting features of OPNet is its ability to execute and monitor several scenarios in a concurrent manner. However, OPNET also suffers from the same object-oriented scalability problems as ns-2. OPNet modeler runs on Windows XP/2K, Linux and Solaris platforms [20].

QualNet: QualNet network simulation software has been developed and marketed by Scalable Network technologies.. It is a commercial ad hoc network simulator based on the GloMoSim. It provides a comprehensive set of tools with many components for custom network modeling and simulation. Models in source code form provide developers with a solid foundation from which to build new functionality or to modify exiting functionalities. QualNet does have a range of wired as well as wireless models but its main strength is in the wireless area. QualNet also largely extends the set of models and protocols supported by the initial GloMoSim distribution. As it is built on top of GloMoSim, QualNet is written in Parsec [21].

MATLAB: MATLAB is an interactive software environment and programming language from The MathWorks which has been founded in 1984. MATLAB was written in C. It supports cross-platform operating system. It is used to make measurements, analyze and visualize data, generate arbitrary waveforms, control instruments, and build test systems. It

provides tools and command-line functions for data analysis tasks such as signal processing, signal modulation, digital filtering, and curve fitting. MATLAB and companion toolboxes provide engineers, scientists, mathematicians, and educators with an environment for technical computing applications. With MATLAB and Simulink, one can (a) develop digital signal processing (DSP) algorithms, (b) model and simulate systems, (c) automatically generate code for embedded DSPs, MCUs, GPPs, FPGAs, and ASICs, and (d) verify and validate the hardware and software implementations. But the drawback of this package is to deploy MATLAB functions as library files, which can be used with .NET or Java application building environment it is needed that the computer where the application has to be deployed needs MCR (MATLAB Component Runtime) for the MATLAB files to function normally. Another drawback is that M-code written for a specific release of MATLAB often does not run with earlier releases as it may use some of the newer features [22].

Monte Carlo based Simulatiors[3]: This method solves a problem by generating suitable random numbers and observing that fraction of the numbers obeying some property or properties. The name and the systematic development of Monte Carlo methods date from about 1944. Monte Carlo simulation methods are especially useful for modeling phenomena with significant uncertainty in inputs and in studying systems with a large number of coupled degrees of freedom. The method is useful for obtaining numerical solutions to problems which are too complicated to solve analytically. In the science and engineering communities, Monte Carlo simulation is often used for uncertainty analysis, optimization, and reliability-based design. But it avoids higher order statistics of the output sequences. So, approximate method of it is the main disadvantage of Monte Carlo method. By increasing the number of iterations by the costs of simulation time, any degree of precision can be easily achieved. Another limitation is the number of random numbers that can be produced by random number generating algorithm. To use it one can develop codes in MATLAB or C/C++ or Visual Basic or Java. The popularity of Monte Carlo methods has led to a number of superb commercial tools [23].

3.1 Comparison

In this sub section we summarize the most interesting capabilities, advantages, and drawbacks of existing tools for wireless networks in table 2. Table 2 has all simulators considered in the previous section listed in the consecutive columns and special features/capabilities in the context of all simulators in the consecutive rows, respectively.

Sr.N.	Tools / Features	NS2	GloMo-Sim	J-Sim	OMNet++	OPNet	QualNet	MATLAB
1	Applicability	Net./Sys.	Net./Sys.	Network	Net./Sys.	Net./Sys.	Net./Sys.	System
2	Interface	C++/OTcl	Parsec (C-Based)	Java/Jacl	C++/NED	C or C++	Parsec (C-Based)	C++

[3] "Monet Carlo based simulation" is representing a category of simulators rather then an individual simulation tool.

3	Available Modules	T/W/Ad/ WSNA	T/W/ Ad	T/W/Ad / WSNA	T/W/Ad	T/W/Ad/ WSN	T/W/Ad / WSNA	Data Acquisition Toolbox, Instrument Control Toolbox , Image Acquisition Toolbox
4	Mobility	Support	Support	Support	No	Support	Support	Support
5	Graphical Support	No or very limited visual aid	Limited Visual aid	Good visualization and debug facility	Good visualization and excellent facility for debug	Excellent graphical support, Excellent facility for debug.	Good graphical support, Excellent for debug.	Excellent graphical support, Excellent facility for debug.
6	Parallelism	No	SMP /Beowulf	RMI-based	MPI/PVM	Yes	SMP /Beowulf	Yes
7	License	Open Source	Open Source	Open Source	Free for academic and educational use	Free academic License for limited use	Commercial	Commercial
8	Scalability*	Small	Large	Small	Large	Medium	Very Large	Very Large
9	Documentation and user support	Excellent	Poor	Poor	Good	Excellent	Good	Excellent
10	Extendibility*	Excellent	Excellent	Excellent	Excellent	Excellent	Excellent	Excellent
11	Emulation	Limited	Not Direct	Yes	Limited	Not Direct	Yes	Yes

T: Traditional Models (eg. TCP/IP, Ethernet)
W: Wireless Support (eg. Propagation model, IEEE 802.11)
Ad: Ad-Hoc Support (eg. AODV, DSR)
WSN: Wireless Sensor Networks Support (eg. S-MAC, Direct Diffusion)
WSNA: Advance Wireless Sensor Networks Support (eg. Zigbee, Energy Model)
*Concerning table 2, as no exact metrics are available for scalability and extendibility, we define Very Large > Large> Medium > Small, and Excellent > Good > Poor, respectively.

Table 2. Comparison of Different Network and System Simulation Tools

Wireless Networks simulators exhibit different features and models. Each has advantages and disadvantages, and each is appropriate in different situations. In choosing a simulator from the available tools, the choice of a simulator should be driven by the requirements. Developers must consider the pros and cons of different programming languages, the means in which simulation is driven (event vs. time based), component-based or objectoriented architecture, the level of complexity of the simulator, features to include and not include, use of parallel execution, ability to interact with real nodes, and other design choices. While design language choices are outside of the scope of this paper, there are some guidelines that appear upon looking at a number of already existing simulators. Most simulators use a discrete event engine for efficiency. Component-based architectures scale

significantly better than object-oriented architectures, but may be more difficult to implement in a modularized way.
Defining each wired/wireless node as its own object ensures independence amongst the nodes. The ease of swapping in new algorithms for different protocols also appears to be easier in object-oriented designs. However, with careful programming, component based architectures perform better and are more effective. Generally, the level of complexity built into the simulator has a lot to do with the goals of the developers and the time constraints imposed. Using a simple MAC protocol may suffice in most instances, and only providing one saves significant amounts of time. If high-precision PHY layers are needed, then ns-2 (coupled with the highly-accurate PHY) is clearly the wisest choice. The number of nodes targeted also determines the choice of the simulation tool. Sequential simulators should not be expected to run more that 1,000 nodes. If larger scales are needed, then parallel simulators are a wise choice. Finally, most non-commercial simulators suffer from a lack of good documentation (NS2 is an exceptional case here) and support. Using a commercial one might help in case of troubles. Moreover, commercial simulators usually feature extensive lists of supported protocols, while open source solutions give full empowerment.

3.2 Analysis

In the previous section we provide the background on a number of different network simulators and present the comparison of some important features of each. In continuation of our research work, we present our survey results to take up on the credibility issues of simulation studies in wireless networks, and to alert the researchers on some common simulation issues and pitfalls. We conducted a survey on wireless networks, especially on Ad-hoc/Mesh/Sensor/Cognitive Radio networks studies published in some of the premiere conferences of the wireless networks from years 2000 to 2008. Table 3 lists the name of all conferences that we considered in our survey. We only included the full papers on PHY, MAC and Routing layers in our survey, not the poster and demonstration papers. We reviewed each paper individually avoiding word searches or other means of automatically gathering results. For consistency, the same person reviewed all of the papers; to validate the results and to correct the few inconsistencies we had a second person review all of the papers again.

Sr.No.	Confernce Name	Applicable Area	Average Acceptance Ratio*	Specialized Area	Years
1	ACM MobiCom	Network	≈13%	Ad-hoc/Mesh/Sensor Networks Tracks	2000~2008
2	ACM/IEEE MobiHoc	Network/System	≈15%		2000~2008
3	ACM Sigcomm	Network/System	≈15%		2000~2008
4	IEEE Infocom	Network	≈20%		2000~2008
5	IEEE Percom	Network	≈13%		2003~2008
6	IEEE GlobeCom	System	≈40%		2000~2008
7	IEEE WCNC	System	≈42%		2000~2008
8	IEEE ICC	System	≈34%		2007~2008
9	ACM SenSys	Networks/System	≈17%	Sensor Networks	2003~2008
10	EWSN	Networks/System	≈16%		2004~2008
11	IEEE CrownCom	Networks/System	≈35%	Cognitive Radio Track	2006~2008
12	IEEE DySpan	Networks/System	≈25%		2005~2008
13	IEEE CogART	Network	N/A		2008

14	IEEE MilCom	Networks/System	N/A		2005~2008
15	Mobile Networks and Applications	System	≈18%	Networks and Applications	2000~2008
16	MASCOTS	System	≈33%	Measurement, modelling and performance analysis of computer systems and communication networks	2000~2008
17	ACM SigMobile	System	≈27%	Mobility of systems, users, data, and computing	2000~2008

*Average Acceptance ratio is calculated over mentioned years, unless otherwise specified.
N/A: we couldn't provide the average acceptance ratio, as exact figure about the acceptance ratio was not mention on the respective conference sites.

Table 3. Name of conferences

Table 4 shows the detailed database of survey data; here, we categorized our data into mainly three categories: MAC layer, Routing layer and PHY layer (especially, for system simulators). Our database includes all related fields papers from above listed conferences. From our survey, we come across many simulator tools, and we broadly classified them into two main categories: "Widely Used" network simulators and "Other" network simulators, they are summarized in table 5.

Sr. No.	Conference Name	Years	Ad-Hoc/ Mesh Networks			Sensor Networks			Cognitive Radio Networks			Total
			Routing	MAC	PHY	Routing	MAC	PHY	Routing	MAC	PHY	
1	ACM MobiCom	2000~2008	25	19	-	10	5	-	-	-	-	59
2	ACM/IEEE MobiHoc	2000~2008	38	32	10	11	13	-	-	-	-	104
3	ACM Sigcomm	2000~2008	27	3	5	4	17	-	-	-	-	56
4	IEEE Infocom	2000~2008	45	26	-	9	6	-	-	-	-	86
5	IEEE Percom	2003~2008	5	10	-	5	10	-	-	-	-	30
6	IEEE GlobeCom	2000~2008	-	-	45	-	-	49	-	-	23	117
7	IEEE WCNC	2000~2008	-	-	-	-	-	21	-	-	7	28
8	IEEE ICC	2007~2008	-	-	-	-	-	9	-	-	18	27
9	ACM SenSys	2003~2008	-	-	-	24	-	7	-	-	-	31
10	EWSN	2004~2008	3	-	-	12	24	4	-	-	-	43
11	IEEE CrownCom	2006~2008	-	-	-	-	-	-	4	41	50	95
12	IEEE DySpan	2005~2008	-	-	-	-	-	-	1	34	26	61
13	IEEE CogART	2008	-	-	-	-	-	-	1	4	-	5
14	IEEE MilCom	2005~2008	-	-	-	-	-	14	-	17	8	39

15	Mobile Networks and Applications	2000~2008	-	-	12	-	-	-	-	-	-	12
16	MASCOTS	2000~2008	-	-	7	-	-	-	-	-	-	7
17	ACM SigMobile	2000~2008	-	-	8	-	-	-	-	-	-	8
		Total	143	90	87	75	75	104	6	96	132	808

Table 4. Survey Data

Sr.No.	"Widely Used" Network Simulators	"Other" Network Simulators	"Widely Used" System Simulators	"Other" System Simulators
1	NS2	Matlab	NS2Matlab	TOSSIM
2	GloMoSim	TOSSIM*	Monte Carlo	Own Simulators
3	J-Sim	Monte Carlo	-	MICA2
4	OMNet++	Own Simulators	-	Spectrum Analyzer
5	OPNet	Simulation Package not mentioned	-	Simulation Package not mentioned
6	QualNet	Rarely used simulators+	-	Rarely used simulators&

*TOSSIM falls under the category of Emulators.
+Rarely used simulators: Includes ROSS, JiST/SWAN, Prowler, Emstar, and EmSim, just to name a few. These simulators are not cited for more then 2/3 research papers in our survey so we put them under the tag named "Rarely used simulators".
& Rarely used simulators (System) : Includes Bayesian Estimator, MFC Coding, MDL/AIC, DUALFOIL, SWEET, BLUE-BCH Estimator, Microwave Studio, BELLHOP, QT Based Simulator, etc. , these simulators are not cited for more then 2/3 research papers in our survey so we put them under the tag named " Rarely used simulators".

Table 5. Network and System Simulator Tools

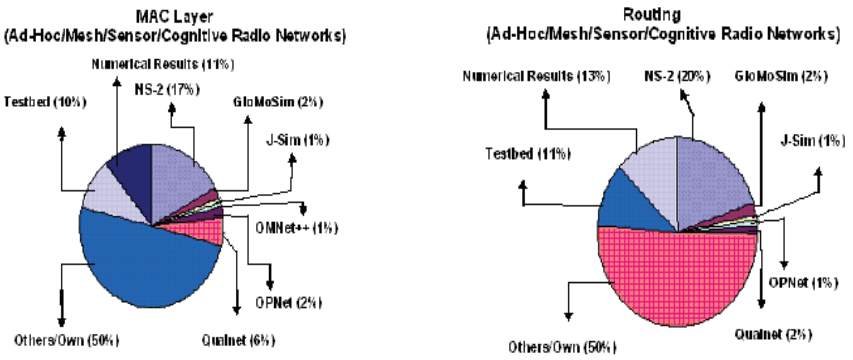


Fig. 3. Simulator Usage in MAC and Routing layers

3.2.1 Analysis: Network

Figure 4 shows the simulator usage results of our survey at different levels of protocol stack, especially at Routing and MAC layers. From figure 4, it is also interesting to know that testbed or experimental studies are also gaining popularity in recent years, and their usage ratio is almost same in Routing and MAC layers. It also shows a good start from the wireless

networks community to present more realistic, practical, and sound research results. But still there are many issues such as scalability, cost, area, etc., need to be addressed to make testbed or experimental setup widely accepted among the community. As of current research practice, simulation is currently the most feasible approach to the quantitative analysis of wireless networks. As we can see from figure 4 NS-2 is the most popular/used simulator among the "Widely Used" network simulator tools. To our surprise we find numerical/mathematical results are more dominating than "Widely Used" tools (NS2 is an exceptional case), as general trend is to present rigorous simulation results than mathematically sound results in wireless networks community. It is worth to note that these numerical results also include the theoretical aspect of the field. From figure 4 we can find a very interesting observation that in both the layers other/own category is at the top. To know the reason we further expand survey results on other/ own category as shown in figure 5.

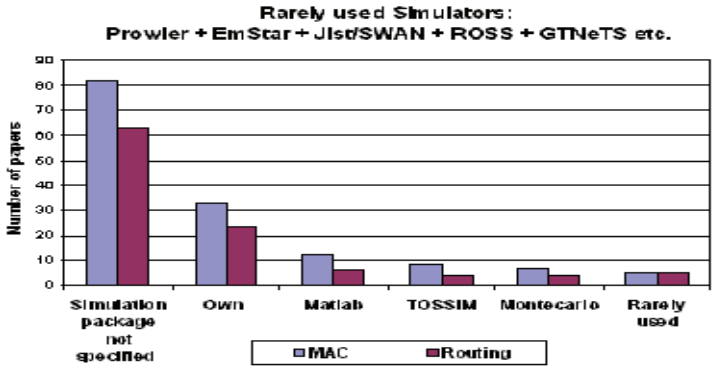


Fig. 4. Survey results on "Other/Own" Category

As we can see from figure 5 major part of other/own category is occupied by the simulator tools which are not specified in the papers. When the simulator used is not specified within a published paper, the repeatability and credibility of the simulation study are questionable. Second topmost category is "own" where researchers have used their self-developed or custom made simulation tools. It is also difficult, if not possible, to repeat a simulation study when the simulation is self developed and code is not available. Rest of the simulator tools, especially Matlab, TOSSIM, Montecarlo, and "rarely used" simulator tools, have a small portion of participation in wireless networks research. One very important fact come out in our survey is that a very few papers (hardly 3/4 papers) cited about the code availability, for whatever reasons but this issue really need an attention from the community. Further more, we obtained some interesting observations from our survey as shown in figure 6.

The execution and analysis of any experiment/simulation study must be based on mathematical principles and need to be statistically sound. For any experimental/ simulation study to be statistically sound must present the number of times simulation runs, confidence levels that exist in the results, and a list of any statistical assumption made. To our surprise, the large numbers of papers don't even bother to present this basic information regarding their research results. As we see from figure 6 nearly 150 papers aren't independently repeatable because of the lack of simulation's information. Additionally, the

papers often omitted simulation input parameters such as traffic model or type. As shown in figure 6 nearly 250 papers didn't specify any traffic model or type they have used. So, this lack of basic information raises many questions on the reliability and repeatability of wireless networks research.

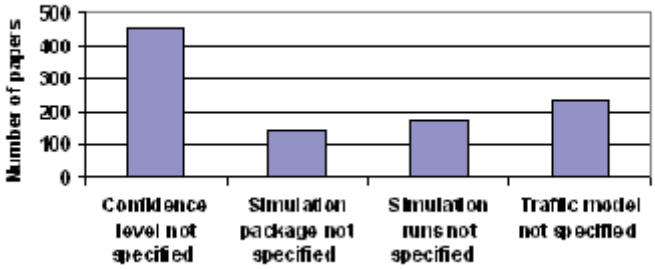


Fig. 5. Simulation issues

3.2.2 Analysis: System

Figure 7 shows the percentage of usage of various simulators in Physical layer. In this layer, sometimes it may not be possible to set up simulation environment. Theoretical analysis would be the best option for this type of situation. A large number of researchers adopt theoretical and numerical analysis in Physical Layer. In this survey, it is noticeable that testbed has approximately same popularity as in the case of Routing and MAC layers. By experiments it can be get more pragmatic and acceptable results. But experimental set up is not always feasible in many cases due to monetary and other limitations. In this regard, MATLAB is more prominent and widely used simulator in Physical layer as compared with Routing and MAC layers. It has many intelligent tools for various simulation purposes. Most of the communication systems can be simulated by using MATLAB. It has little practice in Network simulation. NS2 also plays a significant role in Physical Layer simulation. In addition, Monte Carlo is also an extensively used simulator in Physical Layer as opposed to Routing and MAC Layer. Some researchers also have interest with GloMoSim, QualNet and OPNet but in a little portion same as other two layers. Again, "other/ own" category is in the surprising top position. We would like to find out the reasons behind it.

By expanding this other /own category in figure 8, we can see that major parts of it didn't mention the simulator name. So, it makes same difficulties as in Routing and MAC layer. Second major category is "own" simulators those are developed by researchers themselves or custom made. With these kinds of self developed simulators and if codes are also not available, it is not possible to study these simulators again. Some simulators like M-AFC Coding, Bayesian State Estimator, MDL/ AIC, DUALFOIL, SWEET, BLUE_BCH Estimator, BELLHOP etc., are also used in Physical Layer simulations infrequently, we named them under 'rarely used" category. MICA2, Spectrum Analyzer and TOSSIM have some usage also. Except TOSSIM any other simulators among these are not used in Wireless Network Simulations. These are specially designed for Physical Layer simulations. As other two layers, we also give our attention on system parameters which play important role to get reliable and sound results from a simulation as shown in figure 9.

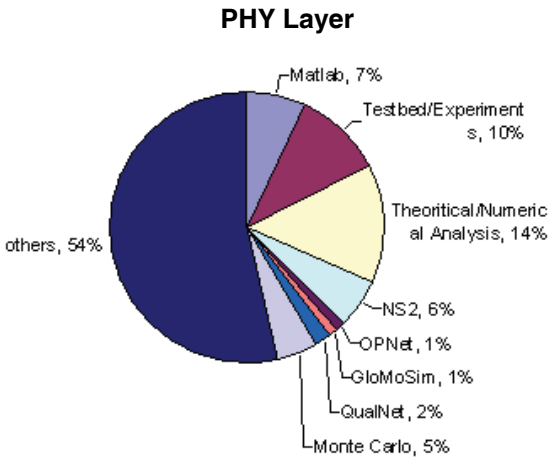


Fig. 6. Simulator usage in PHY Layer

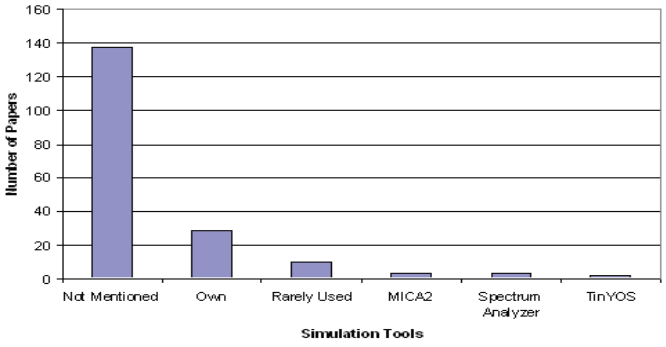


Fig. 7. Survey results on "Other/Own" Category (PHY)

It is already mentioned that results and analysis getting from a simulation study should be reliable and acceptable in a range. In general, a model of a physical system has error associated with its predictions due to the dependence of the physical system's output on uncontrollable or unobservable quantities. Confidence level is an important parameter for the simulation reliability. But from Figure 3 we can see that as like as Routing and MAC layers, most of the papers didn't mentioned the traffic models, acceptance levels, and other statistical parameters explicitly in Physical Layer. It is seen that around 225 papers didn't mention any confidence level in their simulation. Again, in more than 236 papers didn't mention what traffic model have been used. But in next generation networks traffic modeling will have to deal with two main issues: the radio resource management scheme and the effect of the user mobility in the traffic volume per cell. So, information about traffic model is necessary to further repeatability of a simulation and for reliable output. Number of independent run of any simulation has also an impact on accurate result. But a large portion of researchers didn't state any information about that. So, questions may be raised about the credibility of the simulation based analysis.

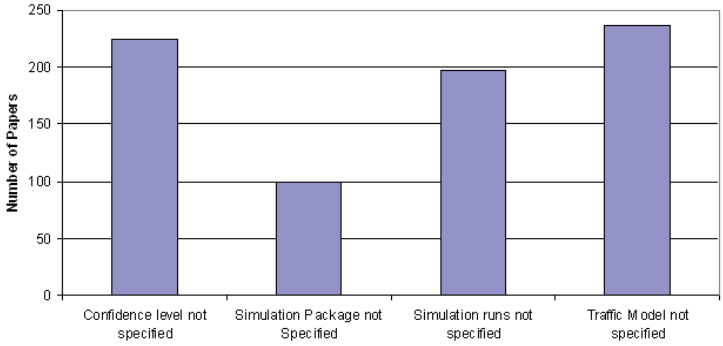


Fig. 8. Simulation issues (PHY)

To raise the awareness on the lack of reliability, repeatability, and credibility of simulation based studies we have developed a list of common issues and pitfalls as the starting point for improvement. We have written the list from our own experiences with simulations as well as the experience of others in the field. Some common issues and pitfalls are identified from our survey. We summarize these issues and pitfalls into the following categories: simulation setup and initial assumptions, simulation execution, and output analysis. They are summarized with our recommendations in table 6.

Category	Issues/Pitfalls	Recommendations
Simulation setup and initial assumptions	Network area, number of nodes, mobility Models, node distribution, traffic model, transmission range, bidirectional communication, capturing effect, simulation type: terminating vs. steady state, protocol stack model, RF propagation model, and proper variable definitions.	• Most of these issues can be easily solved by proper documentation. • Due to space limitation, sometimes publications can include only major settings. In this case authors can provide the external links or references, which include all the needed information. • Try to tune setting some parameters against an actual implementation if possible or improve the abstraction level of used models.
Simulation execution	Protocol model validation, PRNG validation, scenario initialization: empty caches, queues, and table; and proper statistics collection.	• Validating protocol models against analytical models or protocol specifications • Determining the number of independent runs required. • Proper setting and address of random number generators • Collecting data only after deleting transient values or eliminating it by

		proper preloading routing cache, queues, and tables.
Output analysis	Single set of data, Statistical analysis: autocorrelation, averages, aggregation, mean , and variance; confidence level	• Experiment should be run for some minimum number of times • Analysis should be based on sound mathematical principles • Provide proper confident interval for a given experiment.

Table 6. Important issues and recommendations

This paper summarizes the current state of practice, and identified some of the difficult issues that must be resolved to increase the reliability and credibility of simulation based studies. Further more, wireless community should take some concrete steps such as standardization of simulation tools and creating some universal virtual testbeds to resolve the points of consensus as mentioned above. Universal virtual tesbed could be a very useful for all the research groups around the globe and can also be used as standard measuring tool for wireless networks community.

4. Conclusions

In this paper, eight most "widely used" network and system simulators and their strengths and weaknesses were discussed based on a couple of papers and a survey. Then, the results of a survey of recent research publications on performance evaluation of networks were used to show that the majority of results of simulation studies of wireless networks published in technical literature have many pitfalls/issues. With this paper we documented these pitfalls and some important issues with some recommendations to increase the reliability and repeatedly of simulation studies. Finally, we hope, the results presented in this paper will motivate the researches to put their efforts in thorough descriptions of the simulation scenarios and taking care of pitfalls in simulation studies of wireless networks.

5. References

I. F. Akyildiz, W. Y. Lee, M.C.Vuran, and S. Mohanty. NeXt generation/dynamic spectrum access/cognitive radio wireless networks: A survey. Computer Networks 50 (2006) 2127–2159, May 2006.

J. Lessmann, P. Janacik, L. Lachev, and D. Orfanus. Comparative study of wireless network simulators. in proceeding of ICN, 2008, pp. 517-523.

J. Heidemann, K. Mills, and S. Kumar. Expanding Confidence in Network Simulations. IEEE Network, vol. 15, no. 5, 2001, pp. 58-63.

S. M. Sanchez. ABC's of output analysis. in proceeding of the 1999 winter simulator conference, 1999, pp. 24-32.

B. Schilling. Qualitative comparison of network simulation tools. Technical report, Institute of Parallel and Distributed Systems (IPVS), University of Stuttgart, January 2005.

S. Duflos, G. L. Grand, A. A. Diallo, C. Chaudet, A. Hecker, C. Balducelli, F. Flentge, C. Schwaegerl, and O. Seifert. Deliverable d 1.3.2: List of available and suitable simulation components. Technical report, Ecole Nationale Supieure des Tommunications (ENST), September 2006.

M. Karl. A comparison of the architecture of network simulators ns-2 and tossim. In Proceedings of Performance Simulation of Algorithms and Protocols Seminar. Institut fr Parallele und Verteilte Systeme, Abteilung Verteilte Systeme, Universit Stuttgart, 2005.

L. Hogie, P. Bouvry, and F. Guinand. An overview of manets simulation. In Electronic Notes in Theoretical Computer Science, Proc. of 1st International Workshop on Methods and Tools for Coordinating Concurrent, Distributed and Mobile Systems (MTCoord 2005), LNCS, pages 81–101, Namur, Belgium, April 2005. Elsevier.

E. Egea-Lopez, J. Vales-Alonso, A. Martinez-Sala, P. Pavon-Mari, and J. Garcia-Haro. Simulation scalability issues in wireless sensor networks. IEEE Communications Magazine, 44(7):64–73, July 2006.

D. Curren. A survey of simulation in sensor networks. Student project, www.cs.binghamton.edu/˜kang/ teaching/cs580s/david.pdf, 2007.

L. Begg,W. Liu, K. Pawlikowski, S. Perera, and H. Sirisena. Survey of simulators of next generation networks for studying service availability and resilience. Technical Report TRCOSC 05/06, Department of Computer Science & Software Engineering, University of Canterbury, Christchurch, New Zealand, February 2006.

G. F. Lucio, M. Paredes-Farrera, E. Jammeh, M. Fleury, and M. J. Reed. Opnet modeler and ns-2 - comparing the accuracy of network simulators for packet-level analysis using a network testbed. WSEAS Transactions on Computers, 2(3):700–707, July 2003.

K. Pawlikowski, H.-D. Jeong, and J.-S. Lee. On credibility of simulation studies of telecommunication networks. IEEE Communications, 40(1):132–139, January 2002.

M. Lacage and T. R. Henderson. Yet another network simulator. In WNS2, The workshop on ns-2: the IP network simulator, 2006.

G. F. Riley. The georgia tech network simulator. In MoMeTools ’03, pages 5–12, 2003.

http://www.isi.edu/nsnam/ns/

http://pcl.cs.ucla.edu/projects/glomosim/

http://www.j-sim.org/

http://www.omnetpp.org/

http://www.opnet.com/

http://www.scalable-networks.com/

http://www.mathworks.com/

http://www.solver.com/simulation/monte-carlo-simulation.

S.Mehta,Md. H. Kabir, Mst. N. Sultana, N. Ullah, and K.S. Kwak, “ A Case Study of Networks Simulation Tools for Wireless Networrks,” in procceding of AMS’09, May,pp. 661-666.

6

Super-Resolution Procedures in Image and Video Sequences based on Wavelet Atomic Functions

Volodymyr Ponomaryov and Francisco Gomeztagle
National Polytechnic Institute of Mexico
Mexico, Mexico-city

1. Introduction

The images and video sequences obtained from optical, radar, medical sensors, in digital photographs, high definition television, electron microscopy, etc. are formed in the electronic devices, which use different sensors, like x-ray systems, remote sensing cameras, radars, radiometers, US sensors, CCD, etc. (Bovik et al.; 2000, Chaudhuri, 2001; Chaudhuri & Manjunath 2005). So, the images and frames in the video sequences depend on spatial resolution that is defined as a number of pixels per square area in the camera (sensor). The temporal resolution is determined by the frame rate and the exposure time, which limits the maximum speed that can be observed correctly in video. Because of the physical limitations and high cost needed to improve the precision and stability of the imaging system by manufacturing techniques, many applications of image and video sequence data (Farsiu et al., 2004), such as those mentioned above, demand to develop additional methods and algorithms that should restore the resolution degraded in a sensor permitting better observations of the fine details, edges, and restoration of the colour properties. Super-resolution (SR) is defined as a reconstruction of a high-resolution (HR) image or frame in the video sequence from one or multiple low-resolution (LR) images/videos, which is relatively inexpensive to implement. Such methods are effective in the enhancement of the resolution by transcending the limitations of the sensors through digital image processing algorithms. Thus, SR restoration technology is a hot research topic in computer vision applications (Park et al., 2003, Zhang et al., 2010).

This chapter is devoted to analysis of the various ways and methods to get SR in the images or video sequences (Protter et al., 2009; Park et al., 2003). So, it is assumed that the images or frames are treated as LR ones, where the promising methods of super resolution to the entire image or area of interest should be employed recovering the data lost during acquisition stage. Finally, reconstructed data present more information for better visual understanding of selected areas, permitting a deeper analysis for various purposes (Baboulaz et al., 2009)

There are exist a lot of the algorithms in the SR (Franzen et al., 2001; Chaudhuri et al., 2001), among them, the nearest neighbour methods that employ the interpolation procedure with the closest pixels to approximating point; bi-linear interpolation (Hou et al., 1978) that

applies the mean averaging filter for neighbouring pixels in each a central pixel, and in this way obtaining the lost pixels; the bi-cubic algorithm that uses the cubic polynomial function for additional pixels; the methods based on spline technique (Lehmann et al., 1999; Phu et al., 2004) that deform the edges and wave them. Simple interpolation-based methods, such as bilinear or bicubic interpolation, etc. tend to produce HR images with jagged edges, these are a common artifacts for many SR algorithms. All these methods only apply the spatial pixels information. Other algorithms are known in literature, among them, warping, which is based on re-sampled operation on base on rectangular point spread function, and methods based on fuzzy logic theory (Tolpekin et al., 2008). Another group of methods is based on the Fourier transform with band limited function interpolation. Here, the restoration is realized by extension of the zeros, applying Discrete Fourier transform (DFT) (Crouse et al., 1998; Maeland et al., 1998; Landi et al., 2006) of size N for original sequence, filling up it with the zeros from N + 1 to 2N, and finally, calculating 2N points in the inverse DFT, that permits improving the detail and edge preservation in SR image. In similar way, this idea can be used employing the Discrete Cosine Transform (DCT) to find the lost pixels in an image or video frame, reconstructing SR image via inverse transform as in DFT method. The Wavelet based techniques have been introduced but mainly in specific applications (Crouse et al., 1998; Maeland et al., 1998; Landi et al., 2006, Reichenbach et al., 2003.; Chan et al., 2003; Lertrattanapanich et al., 2002; Ng et al., 2004).

The proposed here techniques take into account the spatial and spectral Wavelet pixel information permitting to reconstruct different video (Katsaggelos et al., 2007; Chaudhuri&Manjunath, 2005; Qin Feng-qing et al.; 2009, C. Wang et al, 2006) composition and texture nature, and, as it is observed from realized simulations, present good performance in terms of objective and subjective criteria (Chan et al., 2003; Lertrattanapanich et al., 2002; Ng et al., 2004).

Here, we describe in details the novel SR method applying the Wavelets based on atomic functions (WAF) (Gulyaev et al., 2007). Novel Wavelet families (*Fup, Up, Gk, Ξn, π*) that are employed in SR restoration present the better performance in the compression of different types of the images and video sequences due to its special approximation properties explaining in this chapter. Recently, WAFs have already demonstrated their successful performance in the diverse fields, such as windowing in radar processing, compression and recognition of medical images, speech reconstruction, image processing, etc. (Juarez et al.; 2008, Kravchenko et al.; 2008, Kravchenko et al., 2009). So, it is also expected better estimation of lost information and possible improvement during reconstruction in the SR procedure. Additionally, the most common Wavelet families, such as Daubechies, Symlets, Biorthogonal and Coiflets are tested also.

The idea applied in Wavelet based techniques is justified by such a proposition: If Wavelet transform is efficiently used in the compression and decompression of the images without significant lost of information, then it is supposed that the reconstruction of HR image or frame in a video sequence can be realized sufficiently well using the inverse Wavelet transform, so treating the initial LR image as a before compressed one. In such a way, the reconstruction of SR data is realized by extension of an image (frame) size up to 4 times in comparison with the original LR image.

Due to movement of an object or a scene during the video acquisition process, the frames are different from each other, so, utilizing the spatial sub pixel movement information between the frames, a spatial HR video sequence can be reconstructed from a LR video

(Shen Huanfeng et al.; 2007, Callico et al., 2008). Principally, this permits to restore the high frequencies behind the diffraction limit of a sensor. For neighbouring frames in the video sequences, which can be significantly different because of motion, the similar pixels are tested with the purpose to find the movement estimate (Jain et al., 1981). Such motion estimation is used to obtain the better estimates of the missing values. The apparent motion vectors are calculated between two neighbouring frames obtaining additional pixels. The precision of the registration stage is an important for the reconstructed image quality, because sometimes it is better to interpolate a LR image using classical algorithms than to reconstruct a HR image/frame from a set of images applying incorrect motion parameters.
In the chapter, the proposed methods are also investigated under criterion of real time implementation, where additionally to restoration quality, the time values needed to reconstruct is considered, so, only fast method in motion estimation and SR are employed here, like "block matching" that commonly is used (Gomeztagle et al., 2009).
To compare the robustness of the analyzed methods different test images and video sequences are studied, These image data present various physical characteristics, such as fine details, edges, texture, contrasts, smooth and rough background, etc. Test video sequences: "Toy", "Plant", "Walter", "Stephan" and "Flowers" have been investigated in greyscale and colour formats.
To get objective performance of reconstruction, the criteria: Pick Signal to Noise Ratio (PSNR), Mean Absolute Error (MAE), and Normalized Colour Difference (NCD) are employed (Bovik, 2000; Kravchenko et al.; 2009; Farsiu et al., 2006; Akgun et al., 2005; Wood et al., 2008).
Finally, the possibility of the real time processing is discussed implementing several promising frameworks on the Texas instruments Digital Signal Processing (TMS320C642, 2004).

2. Performance Criteria

There exist different objective measures that are used in evaluation of image restoration qualities. Here, to characterize different known and proposed SR algorithms, and compare their performances, several criteria are employed: the Peak Signal-to-Noise Ratio (*PSNR*) for the characterization of noise suppression and artifacts limitations, Mean Absolute Error (*MAE*) for quantization of edges and fine detail preservation, and the Normalized Color Difference (*NCD*) for the estimation of the color perceptual error (Bovik, 2000; Kravchenko et al., 2009) The PSNR is defined as:

$$PSNR = 10\log\left(\frac{(255)^2}{MSE}\right), \text{ dB}, \tag{1}$$

where the Mean Square Error (*MSE*) is the error measure for gray scale image of dimension MN (Bovik, 2000).
The *MAE* criterion is written as follows:

$$MAE = \frac{1}{M \times N}\sum_{x=0}^{M-1}\sum_{y=0}^{N-1}\left|\left[\hat{f}(x,y) - f(x,y)\right]\right| \tag{2}$$

Every measure, *PSNR* and *MAE* to get the objective criteria value employ the reference image HR $f(x,y)$ and other one $\hat{f}(x,y)$ obtained from SR algorithm.

NCD criterion should be calculated in the $L^*u^*v^*$ space (Katsaggelos et al., 2007; Kravchenko et al., 2009)) and is a measure of color errors:

$$\mathrm{NCD} = \frac{\sum_{i=1}^{M_1}\sum_{j=1}^{M_2}\left\|\Delta f_{Luv}(i,j)\right\|_{L_2}}{\sum_{i=1}^{M_1}\sum_{j=1}^{M_2}\left\|f^*_{Luv}(i,j)\right\|_{L_2}} \tag{3}$$

Here, $\left\|\Delta f_{Luv}(i,j)\right\|_{L_2} = \left[\left(\Delta L^*(i,j)\right)^2 + \left(\Delta u^*\right)^2 + \left(\Delta v^*\right)^2\right]^{1/2}$ is the norm of color error; ΔL^*, Δu^*, and Δv^* are the differences in the L^*, u^*, and v^* components, respectively, between the two color vectors that present the SR reconstructed image and original HR image for each a pixel *(i, j)* of an image; and $\left\|f^*_{Luv}(i,j)\right\|_{L_2} = \left[\left(L^*\right)^2 + \left(u^*\right)^2 + \left(v^*\right)^2\right]^{1/2}$ is the L_2 norm or magnitude of the original HR image pixel vector in the $L^*u^*v^*$ space. It has been proved that the *NCD* objective measure expresses well the color distortion (Kravchenko et al., 2009).

Since it is difficult to define the error criteria for an accurate quantization of SR image reconstruction, a subjective measure of the image distortion in form of subjective visual perception is used in this paper. It is presented by error image - the absolute difference between the original HR image and reconstructed SR one. So, subjective visual comparison of the images provides information about the spatial distortion and artifacts introduced by an algorithm employed, and present the performance of the analyzed technique when the SR image or SR frame of the video sequence are observed by the human visual system.

The motion estimation is one of the fundamental problems in the treatment of the digital video sequences (Wüst Zibetti et al. 2007; Callico et al., 2008; Kravchenko et al., 2009). The objective of motion estimation consists of calculating the field of motion vectors to describe the apparent movement between two images of the sequence. It is important to deal with apparent movement, because the dynamic changes (motions) of the images are the projection on 2D plane at discrete moments of time from 3D spatial-temporal scenes. This supposes a loss of information that does necessary to distinguish between the real movement that projects on the plane and the movement pretends that, well, to keep redundant information with the goal to improve the estimate. In this specific application, the estimations of the movements between the frames should be found, and the technique "block matching" is usually used (Callico et al., 2008). Because this technique is too expensive in computation charge, we apply the motion estimation only in areas where two images have differences. A set of pixels in a window of sizes 9 x 9 pixels in a first frame, that should be slipped into the next frame is used in order to find the minimum of the difference

according to criterion: $E(d)=\sum_{x\in R}\left|f_t(x)-\hat{f}_{t,\tau}(x+d(x))\right|$. In such a way, it is possible to obtain the redundant information from two blocks. So, we can use the information in the same zone of a scene that is found in two frames, in order to increase the sample size permitting the correct estimation of the lost pixels. The mentioned above algorithm in motion estimation permits to form the lost pixels in the SR reconstruction, and it is simple and sufficiently fast. There exist a lot of other algorithms with better performance but their computational charges are sufficiently bigger, this does not provide their real time implementation.

3. Wavelet Atomic Functions and their Properties

3.1 Atomic Functions

Let present novel family of the Wavelets, the WAF, firstly introducing basic atomic functions (*up*, fup_n, g_k, up_n, Ξ_n, π_n) used as mother functions in their Wavelets construction. The idea of AF was consisted of finding a function where the maximum and minimum of their derivatives should be similar to maximum of initial function. The result of such a mathematical problem is in infinitely differentiable solution of the differential equations with a shifted argument (Kravchenko et al., 2008, 2009; Gulyaev et al., 2007):

$Lf=\lambda\sum_{k=1}^{n}c(k)f\left[ax-b(k)\right]$, $|a|>1$, where L is a linear differential operator with constant coefficients. It has been shown that AFs take intermediate "place" between splines and classical polynomials. Similarly to *B*-splines AFs are compactly supported and similarly to polynomials they are universal from the point of view of their approximation properties. AFs are useful in numerical analysis, in the cases when an approximated function is smooth enough and the use of polynomials is inconvenient due to the fact that they are not compactly supported.

The simplest and most important AF is generated by infinite-to-one convolutions of rectangular impulses. To investigate such a convolution we use the Fourier transform. Applying standard Fourier transform, the rectangular impulse is represented as: $\varphi(x)=\frac{1}{2\pi}\int_{-\infty}^{\infty}e^{jux}\frac{\sin(u/2)}{u/2}du$; analogously, the *N-to-one* convolution of *(N+1)* identical rectangle impulses $\varphi(x)$ defines the compactly supported spline $\theta_N(x)$:

$$\theta_N(x)=\frac{1}{2\pi}\int_{-\infty}^{\infty}e^{jux}\left(\frac{\sin(u/2)}{u/2}\right)^{N+1}du. \tag{4}$$

The function *up(x)* is represented by Fourier transform for the infinite convolution of rectangular impulses with variable length of duration 2^{n-1} similar to eq. (1)

$$up(x)=\frac{1}{2\pi}\int_{-\infty}^{\infty}e^{jux}\prod_{k=1}^{\infty}\frac{\sin(u\cdot 2^{-k})}{u\cdot 2^{-k}}du \tag{5}$$

The Atomic Functions $fup_N(x)$ is defined by the convolution of compactly supported spline $\theta_N(x)$ and AF $up(x)$ in the interval $\left[-(N+2)/2,(N+2)/2\right]$:

$$fup_N(x)=\theta_N(x)*up(2x)=\theta_{N-1}(x)*up(x),\quad fup_0(x)\equiv up(x). \tag{6}$$

The Fourier transform of $fup_N(x)$ is written as follows:

$$fup_N(x)=\int_{-\infty}^{\infty}e^{jux}\left(\frac{\sin(u/2)}{u/2}\right)^N\prod_{k=1}^{\infty}\frac{\sin\left(u\cdot 2^{-k}\right)}{u\cdot 2^{-k}}du\,, \tag{7}$$

Next AF $\Xi_n(x)$, which is used here, is defined as a compactly supported solution of the equation:

$$y^n(x)=a\sum_{k=0}^{n}C_n^k(-1)^k\,y\left[(n+1)x+n-2k\right],\quad x\in\left[-1,\ \ 1\right]. \tag{8}$$

Using transforms analogous to those made in (Kravchenko et al., 2009), we obtain the following integral representation for AF $\Xi_n(x)$:

$$\Xi_n(x)=\frac{1}{2\pi}\int_{-\infty}^{\infty}\exp\{ixt\}\prod_{k=1}^{\infty}\left(\frac{\sin t(n+1)^{-k}}{t(n+1)^{-k}}\right)^n dt\quad,\ \Xi_1(x)=up(x). \tag{9}$$

Another AF $g_k(x)$ employed in this work is defined in (Kravchenko et al., 2009), as the compactly supported solution of differential equation in form of the Fourier transforms:

$$g_k(x)=\frac{1}{2\pi}\int_{-\infty}^{\infty}F_k(t)\exp\{-itx\}dt, \tag{10}$$

Where $F_k(t)=\prod_{j=1}^{\infty}\dfrac{k^2}{1-2\cos(2k/3)}\dfrac{\left[\cos\left(2t3^{-j}\right)-\cos(2t/3)\right]}{k^2-t^2 9^{1-j}}$; $a=\dfrac{2}{3}\dfrac{k^2}{1-\cos(2k/3)}, b=2a\cos(2k/3)$.

Next AF $\pi_m(x)$ considers the differential equation:

$\pi'_m(x)=a\left[\pi_m(x_1(m))+\sum_{k=2}^{2m-1}(-1)\pi_m(x_k(m))-\pi_m(x_{2m}(m))\right]$, where $x_k(m)=2mx+2m-2k+1, x\in R^1$, $k=\overline{1,2m};\ m=3,4,...,$, and can be presented using the Fourier transform:

$$F_m(t)=\prod_{k=1}^{m}\frac{\left[\frac{\sin(2m-1)t}{(2m)^k}+\sum_{v=2}^{m}(-1)^v\frac{\sin(2m-2v+1)t}{(2m)^k}\right]}{(3m-2)t/(2m)^k} \tag{11}$$

Finally, the AF $up_m(x)$ used below is the generalization of presented above AF $up(x)$, and can be characterized by their Fourier transform (Kravchenko et al., 2009):

$$up_m(x)=\frac{1}{2\pi}\int_{-\infty}^{\infty}e^{jxu}\prod_{k=1}^{\infty}\frac{\sin^2\left(\frac{mu}{(2m)^k}\right)}{\frac{mu}{(2m)^k}m\sin\left(\frac{u}{(2m)^k}\right)}du \quad , m=1, 2, 3, \quad up_1(x)=up(x) \tag{12}$$

Let explain the construction procedure for WAFs and their properties (Gulyev et al., 2007). Each a WAF with unit norm has such a structure:

$$\psi_\theta^p(x)=\frac{\exp(j\pi x)h_\theta^p(x)}{\left\|h_\theta^p(x)\right\|}. \tag{13}$$

The function $h_\theta^p(x)$ in eq. (13) is determined as:

$$h_\theta^p(x)=\frac{1}{2^p}\sum_{k=0}^{(p-1)/2}C_p^k\left(\theta(x+\frac{p-2k}{2})+\theta(x-\frac{p-2k}{2})\right) \quad \text{odd } p, \tag{14a}$$

$$h_\theta^p(x)=\frac{1}{2^p}\left[\sum_{k=0}^{(p-2)/2}C_p^k\left(\theta(x+\frac{p-2k}{2})+\theta(x-\frac{p-2k}{2})\right)+C_p^{p/2}\theta(x)\right] \text{ even p,} \tag{14b}$$

where $\widetilde{\theta}(\omega)$ is the spectrum of chosen atomic function $\theta(x)$ from presented above in eqs. (5), (7), (9) - (12).

3.2 Wavelet Key Properties

Inverse Discrete Wavelet Transform (IDWT) is applied in reconstruction of SR image. The DWT and IDWT are usually implemented employing the filter bank techniques in the scheme with only two filters for rows. The Wavelet decomposition algorithm applies two analysis filters $\widetilde{H}(z)$ (lowpass) and $\widetilde{G}(z)$ (highpass), and the reconstruction algorithm uses the complementary synthesis filters *H(z)* (lowpass) and *G(z)* (highpass). The highpass operators are obtained by simple shift and modulation presented as $\widetilde{G}(z)$ = *z H(-z)* and

$G(z) = z^{-1}\widetilde{H}(-z)$. The WT involves the scaling functions $\varphi(x)=\frac{2}{H(1)}\sum_{k\in Z}h_k\varphi(2x-k)$, where the wavelets themselves are linear combination of the scaling functions: $\psi(x)=\frac{2}{H(1)}\sum_{k}g_k\varphi(2x-k)$. The corresponding analysis and synthesis Wavelet basis functions are $\widetilde{\psi}_{i,k}=2^{-i/2}\widetilde{\psi}(x/2^i-k)$ and $\psi_{i,k}=2^{-i/2}\psi(x/2^i-k)$, respectively, where *i* and *k* are the translation and scale indices.

Frequency response allows appreciating the behavior of the synthesis filters (in SR problem) in a graphic way to appreciate the differences of the different Wavelet families used.

Approximation order implies that the scaling function $\varphi(x)$ reproduces all polynomials of degree lesser or equal to $n = L - 1$. The stability of the wavelet representation and its underlying multi-resolution bases are depended on translations the scaling functions and how wavelets form Riesz bases (Meyer, 1990). To analyze this the *Cross-correlation* function is defined as a 2π periodic function $a_{\varphi1,\varphi2}(\omega)$: $a_{\varphi1,\varphi2}(\omega)=\sum_{k\in Z}e^{-kj\omega}\varphi_{12}(k)$, where $\varphi_{12}(x)=\int\varphi_2(\xi)\varphi_1(\xi+x)d\xi$.

Riesz bounds. The tightest upper and lower bounds, $B < \infty$ and $A > 0$ of the autocorrelation filter are the Riesz bounds of $\varphi(x)$ and given by: $A^2=\inf_{\omega\in[0,2\pi]}a_\varphi(\omega)$, $B^2=\sup_{\omega\in[0,2\pi]}a_\varphi(\omega)$, satisfying to the next equations: $A=\inf_{c\in\ell^2}\frac{\left\|\sum_{k\in Z}c_k\varphi(x-k)\right\|_{L^2}}{\|c\|_{\ell^2}}$, and $B=\sup_{c\in\ell^2}\frac{\left\|\sum_{k\in Z}c_k\varphi(x-k)\right\|_{L^2}}{\|c\|_{\ell^2}}$. The existence of the Riesz bounds ensures that the underlying basis functions are in L^2 and that they are linearly independent (in the ℓ^2 space). The Riesz basis property expresses equivalence between the L^2-norm of the expanded functions and the ℓ^2-norm of their coefficients in the wavelet or scaling function basis. There is a perfect norm equivalence (Parseval's relation), if and only if $A = B = 1$, so, in this case the basis is orthonormal.

Projection angle θ between the synthesis and analysis subspaces V_a and $\widetilde{V}_a$ is defined as: $\cos\theta=\inf_{f\in\widetilde{V}_a}\frac{\|\mathrm{P}_a f\|_{L^2}}{\|f\|_{L^2}}=\frac{1}{\sup_{\omega\in[0,2\pi]}\sqrt{a_\varphi(\omega)\cdot a_{\widetilde{\varphi}}(\omega)}}$; this fundamental quantity is scale-independent, and it allows comparing the performance of the biorthogonal projection $\widetilde{\mathrm{P}}_a$ with that of the optimal least squares solution P_a for a given approximation space V_a (Kravchenko et al., 2008): $\forall f\in L^2,\quad \|f-\mathrm{P}_a f\|_{L^2}\le\|f-\widetilde{\mathrm{P}}_a f\|_{L^2}\le\frac{1}{\cos\theta}\|f-\mathrm{P}_a f\|_{L^2}$.

The biorthogonal projector will be essentially as good as the optimal one (orthogonal projector onto the same space) provides that *cos θ* is close to one.

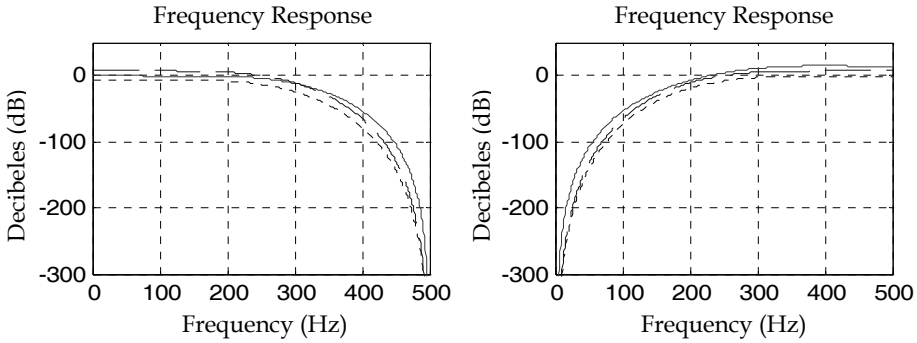


Fig. 1. Frequency responses: Wavelet 9/7 (solid line), Daubechies 8 (dotted line), and Symlet 8 (dashed line).

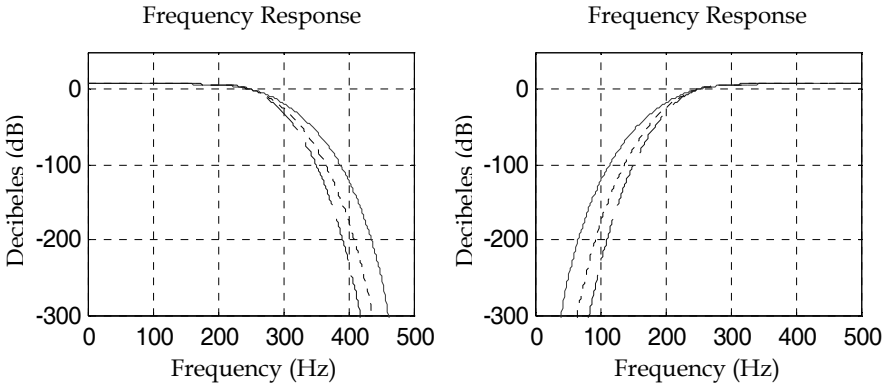


Fig. 2. Frequency responses: WAFs up(*t*) (solid line), $\text{fup}_2(t)$ (dotted line), and eup(*t*) (dashed line).

Figure 1 and 2 expose the frequency responses for some classical Wavelets and WAFs, accordingly, showing that Daubechies and Symlet filters are more selective that the Wavelet 9/7 filters.

It has been observed that the WAF filters have an respond function answer in more selective frequency that the better classical, potentially permitting the best approximation quality in the SR problem.

Finally, Table 1 presents the key properties of the different Wavelets used in SR of the images (Kravchenko et al., 2008). So, the approximation property of estimation is characterized by relative error $\delta = 2(1-r)$, where *r* is correlation coefficient that is equal to projection cosine in this case, the calculations have shown that WAF based on *eup(x)* can potentially produce relative variance error of 0,00464 (6.8% in RMS value), Wavelet *Db8*

gives the value of 0,02242 (of about 15% in RMS value), and *Wavelet 9/7* presents the value of 0,03234 (more than 18% in RMS value). The existence of the limits Riesz bounds demonstrates that the coefficients of the analysis and synthesis filters are lineally independent. The found projection cosine shows that the WAFs are near to the ideal value, this implies that they are better semi-orthogonal and the "most independent". These properties permit to expect that the families of WAFs have sufficiently better acting in approximation problems such as SR one than the traditional families.

Key properties for different Wavelet filters												
Type	**Wavelet 9/7**		**Daubechies 8**		**Symlet 8**		**WAF *up(t)***		**WAF *fup* $_2$*(t)***		**WAF *eup(t)***	
	Dec.	Rec.	Dec.	Rec.	Dec.	Rec.	Dec.	Rec.	Dec.	Rec.	Dec.	Rec.
Approximation Order	4		4		4		4		4		4	
Projection cosine	0.98387		0.98879		0.98781		0.99176		0.99472		0.99769	
Riesz Bounds	0.926	0.943	0.833	0.849	0.880	0.896	0.792	0.806	0.713	0.726	0.641	0.653
	1.065	1.084	1.267	1.290	1.273	1.295	1.514	1.542	1.802	1.834	2.145	2.183

Table 1. Summary of key properties of different Wavelet families.

4. Some Promising Frameworks in Image and Video Super-Resolution

Resolution enhancement via probabilistic deconvolution of multiple degraded images (Sroubek et al. 2006). This approach consists of employing a stochastic fusion method that performs multichannel blind deconvolution (MBD) and SR simultaneously. LR image z_k is modeled by unknown blurring the ideal image $\boldsymbol{u}$, and shifting the result by an unknown vector contaminated by Gaussian noise. This model is a very realistic description of remote sensing observation process where many LR satellite sensors (channels) are employed, and can be rewritten as:

$$z = Gu + n = Ug + n\,, \tag{15}$$

Where $z \equiv \left[z_1^T,\ldots,z_K^T\right]^T, G \equiv \left[zG_1^T,\ldots,G_K^T\right]^T,^T \quad n \equiv \left[n_1^T,\ldots,n_K^T\right]^T, g \equiv \left[g_1^T,\ldots,g_K^T\right]^T$ and $\boldsymbol{U}$ is a block-diagonal matrix with $\boldsymbol{K}$ blocks each performing convolution with an image $\boldsymbol{u}$. According to the Bayes solution, the relation between a priori probabilities $\boldsymbol{p(u)}$, $\boldsymbol{p(g)}$ and the a posteriori probability is $\boldsymbol{p(u,g|z)=const\ p(z|u,g)p(u)p(g)}$, where the conditional pdf $\boldsymbol{p(z|u,g)}$ is follows from Gaussian noise approximation, prior Gibbs pdf for $\boldsymbol{u}$, where the last is defined as

$$p(u)=\begin{cases}\dfrac{1}{z}\exp\left\{-\dfrac{1}{2\sigma_u^2}u^T L(v)u\right\} & if\ u\in C_u\\ 0 & otherwise\end{cases},\quad C_u\equiv\{u|0\le u\le 255\}, \tag{16}$$

and the following prior distribution for g is used:

$$p(g)=\begin{cases}\dfrac{1}{Z}\exp\left\{-\dfrac{1}{2}g^T\Im^T D^{-1}\Im g\right\} & if\ g\in C_g\\ 0 & otherwise\end{cases} \tag{17}$$

The optimal MBD solution is defined as a maximum a posteriori (MAP) estimate. It does not require any knowledge of the blurring functions and the input channels might be mutually shifted by an unknown vector. Allowing only translational between-channel the MAP estimation is given as:

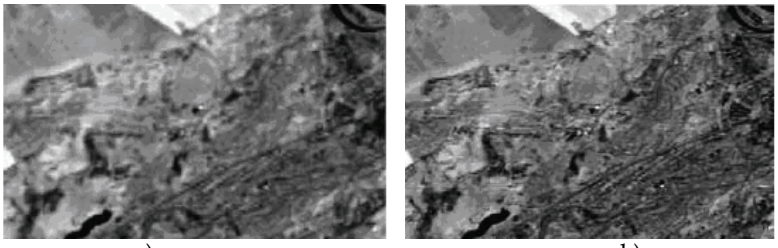
$$\{\hat{u},\hat{g}\}=\arg_{u\in C,g\in Cg}\min\left\{(z-Gu)^T\sum{}^{-1}(z-Gu)+\frac{1}{\sigma_u^2}u^T L(v)u+g^T\Im^T D^{-1}\Im g\right\} \tag{18}$$

The genetic algorithms are applied adopting an approach of alternating minimizations over u and g. The proposed AM-MAP algorithm alternates between two steps:

$$1.\ \hat{u}=\left(G^T\Sigma^{-1}G+\frac{1}{\sigma_u^2}L(v)\right)^{-1}G^T\sum{}^{-1}Z\wedge u\in C_u, \tag{19a}$$

$$2.\ \hat{g}=\left(U^T\Sigma^{-1}U+\Im^T D^T\Im\right)^{-1}U^T\sum{}^{-1}z\wedge g\in C_u, \tag{19b}$$

Here, a decimation operator matrix D is introduced to model a LR acquisition of digital sensors by performing convolution with a 2X 2 uniform mask returns every second pixel and down-sampling of images. In the discrete case, the acquisition model becomes $Z=\mathrm{DGu}+\mathrm{n}$. The steps in the AM-MAP algorithm are the same, except the G, U and $\Im$ are replaced with DG, DU and $\Im D$, respectively. An iterative fusion algorithm was developed recovering a HR image from misaligned and blurred input channels. The fusion problem is formulated as the MAP estimation with the prior probabilities derived from the variational integral and from the mutual relation of co-prime channels. The simulation results of an approach expose that framework can form high-quality fused images. Recovering SR and blind deconvolution, the method can restore the images as it shown in Fig.3 The data source for simulation (playing the role of the *ideal* image) was the 300×300 SPOT HRV image covering the north-western part of Prague (Czech capital). LR acquisitions are formed blurring image by randomly generated 6×6 motion masks, corrupted by AWGN of SNR = 50 dB and resolution decimated by factor of two to obtain images of size 150×150. Six such images were generated and used as input channels´ data.



a) b)

Fig. 3. a) Standard MBD fusion, followed by linear interpolation; (b) fusion using AM-MAP with the SR.

The result of fusion using the blind deconvolution approach in (Flusser, 2003) and applying linear interpolation afterward is depicted in Fig. 3(a). The current algorithm with the SR extension performs better and gives a more accurate representation of the original image, as illustrated in Fig. 3(b).

SR Reconstruction Algorithm for Surveillance Images (Zhang et al., 2010) is presented as an edge-preserving maximum a posteriori (MAP) estimation based SR algorithm using a Weighted Directional Markov image prior model for a region of interest (ROI) from several LR surveillance images. In many surveillance video applications, it is of interest to recognize an object that is selected as a ROI where edges of the object are often very important. Standard Gaussian Markov Random Field regularization (GMRF) in the MAP-based SR cannot effectively preserve sharp edges in the estimated images. Different techniques have been proposed such as Huber-Markov regularization and bilateral-TV regularization, Weighted Directional Markov image prior model, which utilizes the weights for different directional smoothness measures of the edge pixels (Chan et al., 2003, Lertrattanapanich et al., 2002). Typically, the imaging process involves warping, followed by blurring and down-sampling to generate LR images from the HR image. So, the LR image can be represented as $y_k = DB\ M_k z + n$, where $\boldsymbol{M_k}$ is warp matrix, $\boldsymbol{B}$ is camera blur matrix, $\boldsymbol{D}$ is down-sampling matrix, and $\boldsymbol{n}$ is noise, k= *1,2,....P*, with P being the number of LR images. It is assumed that the motion of the ROI during the sequence is a globally translational motion and the motions of all points can often be modelled by a parametric model.

In most situations, the problem of SR is an ill-posed inverse one because the information contained in the observed LR images is not sufficient to solve the HR image, so the ill-posed problem should be stabilized to be well-posed. The MAP method, which can easily include image prior or regularization, is an efficient framework to describe the SR problem.

The maximization of this posterior probability distribution is equivalent to such a problem:

$$\hat{z} = \arg \ \min\left[\sum_k \left\| y_k - A_k z \right\|^2 + \lambda \Gamma(z)\right], \tag{20}$$

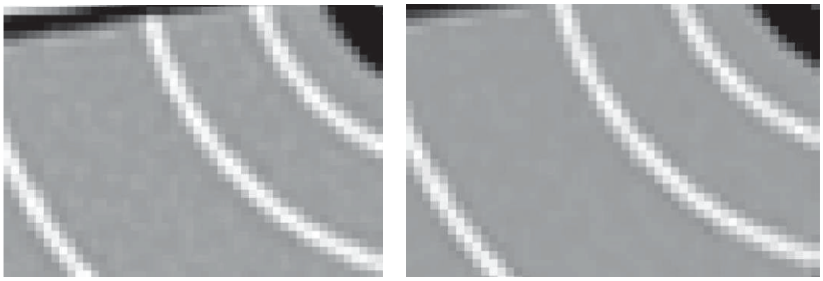
where the first term is the data fidelity term, and $\Gamma(z)$ is the regularization term.

Here, the CG optimization utilizes conjugate direction instead of local gradient to search for the minima permitting faster convergence when compared to the steepest descent method. The gradient of presented above function is denoted as

$$r\left(\hat{z}^{n}\right)=\sum_{k} A_{K}^{T}\left(A_{k}\hat{z}^{n}-y_{k}\right)+\lambda\nabla\Gamma\left(\hat{z}^{n}\right), \tag{21}$$

where the right term of the gradient $\nabla\Gamma\left(\hat{z}^{n}\right)$ is the derivative of the regularization term with respect to z and can be approximated from the estimated HR image. While the left term $A_k^T\left(A_k\hat{z}^n-y_k\right)=M_k^T B^T D^T\left(DBM_k\hat{z}^n-y_k\right)$ can be computed using basic image operations such as warp, blur and sampling instead of sparse matrices multiplications. The matrices $\boldsymbol{M_k, B, D}$ model the principal image formation process: the image warping, blurring and down-sampling, respectively. The implementation of their transpose matrices is also very simple D^{T} is implemented by up-sampling the image without interpolation, i.e., by zero padding. For a convolution blur, B^{T} is the convolution with the flipped kernel of the imaging blur kernel $b(\mathrm{i},\mathrm{j})$; If M_{k} is implemented by backward warping, then M_K^T should be the forward warping of the inverse motion. Thus, the gradient of the cost function with the Weighted Markov Random Field (WMRF) regularization is computed in an efficient manner and the CG optimization technique is used without explicit construction of these large matrices.

The simulation results presented in following figures show the effectiveness of the current proposal (Fig. 4). The image was reconstructed using MRF and WMRF regularizations, and this scenario can be thought of as a SR reconstruction problem with a resolution enhancement factor of one. For the quadratic function chosen as the function of the smooth measure, the reconstruction result of GMRF regularization is shown in Figure 4(a) and Figure 4(b), and GMRF regularization achieves the smallest MSE error of 18.79 ion comparison with other investigated techniques.



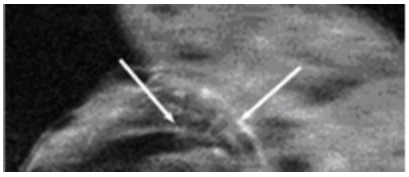
a) b)

Fig. 4. Simulation results of deblurring using different regularizations. (a) HMRF regularization. (b) HWMRF regularization.

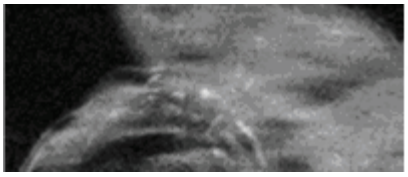
Filters	GMRF	GWMRF	HMRF	HWMRF
First experiment	18.79	14.61	12.91	9.73
Second experiment	77.54	72.02	71.85	62.51

Table 2. MSE error of super- resolution reconstruction using different regularizations.

Representation of HR Images from Low Sampled Fourier Data (LandiI et al., 2006). B-spline functions are proposed to use for the parametric representation of HR images from low sampled data in the Fourier domain. To solve the ill-conditioned linear system arising from the method an efficient regularization method is proposed demonstrating this in applications to dynamic Magnetic Resonance images (MRI) that acquires a time series of images of the same slice of a body where the data are low sampled in the Fourier space to fasten the acquisition. The image representation problem from limited Fourier data is classically addressed as an interpolation problem through zero padding in the Fourier domain but this born in substantially degraded HR images truncation artifacts, including ringing and blurring artifacts. In current method, the desired HR image is represented through a B-spline parametric model where the coefficients are determined. In practical applications, the band-limited interpolator based on *sinc* functions is not suitable for the interpolation of space-limited images. Therefore, several *sinc*-approximating functions: the ideal, windowed and truncated *sinc* interpolation, linear, quadratic, cubic and cubic B-spline interpolation as well as Lagrange and Gaussian methods are discussed in literature and compared the qualitative and quantitative error determinations, computational complexity evaluations, and run time measurements. The cubic B-spline interpolation has very good Fourier properties, small interpolation error and moderate computational cost. It is supposed with a regularization method it can be achieved a further advantage of B-splines over the other sinc-approximating basis functions. In image processing, when using a low-pass filter to perform image denoising, the original noisy image is firstly Fourier transformed and then the high frequency content of its spectrum is ignored. Since noise is mainly distributed over the high frequencies, all frequencies outside a circle of a prescribed radius are set to zero, and then the image is reconstructed by a 2DIFT.



a) Keyhole method, zoom.



b)B-spline Keyhole method with regularization, zoom.

Fig.5. Zoomed parts of the reconstruted images.

Current method can also be used for HR of a single image, i.e in the process of obtaining a HR image given a LR one. In this simple version of the SR problem, the missing high frequency details have to be estimated in order to obtain an image with more pixels. The current B-spline model-based method called B-spline Zero-Padding (BZP) method uses B-spline basis functions to represent HR images from LR data collected in the Fourier space:

$$s(x) = \sum_{l=0}^{N_{low}-1} \alpha_l \beta_l(x), \tag{22}$$

The Keyhole-like methods were tested on real data in the *Mouse* test problem (see data sets in the site: http://mri.ifp.uiuc.edu/V/). Data consist of six data sets from a *Mouse* breast

with a big tumour: a baseline reference data set $D_B(k_x, k_y)$ and an active reference data set $D_A(k_x, k_y)$ of 256 × 256 samples, and four low-sampled data sets $D_t(k_x, k_y)$ of 32 × 256 samples, one for each dynamic section, acquired by a MR spin-echo technique after injecting a contrast agent. Figs. 5(a) and 5(b) show a zoom images by increasing the resolution of a factor 4 where it is evident that the BZP method preserves the quality of the image while the ZP method degrades the image, by introducing the artefacts indicated by the arrows.

Noniterative Interpolation-Based SR Minimizing Aliasing in the Reconstructed Image (Sanchez-Beato et al., 2008). A sampling theory framework is proposed with a pre-filtering step to allow deal with more general data models and also a specific method for SR that uses Delaunay triangulation and B-splines to build the SR image. It has been confirmed the interpolation problem solving in the case of the de-blurring with the translational motion, and with the rotations and shifts where the PSF is rotationally symmetric. The algorithm raises the following: first build a continuous function using Delaunay triangulation and then it should be projected it on the space of polynomial B splines of degree. A cubic B-spline was used, which has a good tradeoff between computational complexity and close behavior to the sampling system of *sinc* functions.

This spline $\beta^3(x)=\begin{cases} 2/3-|x|^2+|x|^3/2 & 0\le|x|<1 \\ (2-|x|)^3/6 & 1\le|x|<2 \\ 0 & 2\le|x| \end{cases}$ is employed to find the expressions needed to calculate the $a(i,j)$ coefficients.

To find the $c(i,j)$ coefficients, the impulse response of the B-spline digital filter of order seven is needed. This B-spline has Z-transform

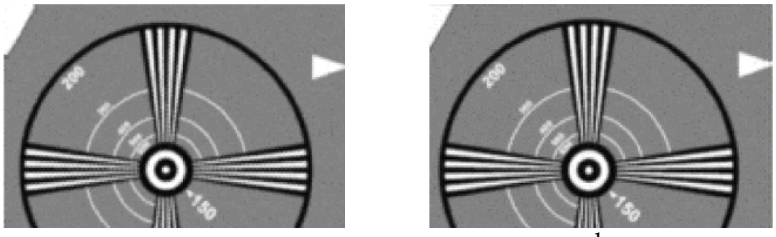
$$S^7(z)=\frac{5040z^3}{1+120z+1191z^2+2416z^3+11191z^4+120z^5+z^6}, \tag{23}$$

which can be implemented as two recursive filters, one causal and another one anti-causal. The denominator of $S^7(z)$ has six real roots, three of them inside the unit circle and the other three outside. Separating the denominator of (23) in its causal and anti-causal parts gives: $S^7(z)=\frac{5040z^3}{1+\alpha_1z^{-1}+\alpha_2z^{-2}+\alpha_3z^{-3}}\cdot\frac{1}{\gamma_1+\gamma_2z+\gamma_3+z^3}=\frac{D(z)Y(z)}{X(z)D(z)}$, permitting to implement the *IIR* filter in such a form:

$$\begin{aligned} d(n)&=5040x(n)-\alpha_1d(n-1)-\alpha_2d(n-2)-\alpha_3d(n-3) \\ y(n)&=\frac{1}{\gamma_1}\left(d(n)-\gamma_2y(n+1)-\gamma_3(n+2)-y(n+3)\right) \end{aligned} \tag{24}$$

Initially, the first equation is run to find the $d(n)$ coefficients and then final output of the filter is obtained in the second equation. For images, this filter is applied once in the

direction x and once in the y direction. The proposed method is non-iterative, scalable and can prevent the presence of aliasing artifacts when the HR image is under-sampled suppressing the high frequency noise. Also, the Delaunay triangulation provides a very strong protection against a possible ill-conditioning of the problem. There is no parameter involved in the reconstruction that is an advantage because, in MAP methods, the gradient descent step and different parameters for the regularization prior are needed. The method is highly parallelizable, once the triangulation is done, each defined triangle independently of the others can be processed.



a b

Fig. 6. Super-resolved images for experiment one with SR factor 2: (a) Image with Tikhonov regularization, PSNR = 18:80dB, (b) Current method, PSNR = 19:04 dB.

The Fig.6 presents the visual results, where one can see that described method achieves the best PSNR among all tested algorithms. The borders of the lines that converge to the centre of the image present aliasing artefacts in all other methods but in current.

Estimation of the Parameters in Regularized Simultaneous SR (Zibetti et al., 2010). A method for automatic determination of the regularization parameters for the class of simultaneous SR algorithms is based on the joint maximum a posteriori (JMAP) estimation technique. This classical technique JMAP has the drawback: it can be unstable and may generate multiple local minima.

It assumes that the frame in the temporal instant k can be represented by the frame in the temporal instant j, with the motion compensated, plus a new information $e_{k,j}$, which cannot be obtained from the frame in the instant j. The motion model is defined as $\boldsymbol{f}_k = \boldsymbol{M}_{k,j}\,\boldsymbol{f}_j + \boldsymbol{e}_{k,j}$ where $\boldsymbol{f}_k$ and $\boldsymbol{f}_j$ are vectors that represent the frames in the temporal instants k and j, respectively. The matrix $\boldsymbol{M}_{k,j}$, of size $M \times M$, represents the motion transformation, or warping.

The simultaneous algorithms estimate the entire sequence of HR frames in a single process. This approach allows the inclusion of the motion matrix in the prior term, improving the quality of the estimated image sequence presenting solution as minimum of cost function:

$$\hat{f} = \arg\min_{f} \left\{ \left\| g - Df \right\|^2 + \lambda_R \left\| Rf \right\|^2 + \lambda_M \left\| Mf \right\|^2 \right\}, \tag{25}$$

where $\boldsymbol{g}$ is LR sequence, $\boldsymbol{f}$ is the HR sequence, $\boldsymbol{D}$, $\boldsymbol{R}$ are block diagonal matrixes, and $M = \begin{bmatrix} I - M_{1,2} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & I - M_{L-1,L} \end{bmatrix}$ is the first order motion difference. Because the regularization

parameters λ_M, λ_R should be known or estimated, this leads to different HR resolution algorithms.

Joint maximum a posterior (JMAP) estimation is given as a point of posterior density function:

$$\hat{f},\hat{\theta},\hat{\beta}_R,\hat{\beta}_M = \arg\max_{f,\theta,\beta_R,\beta_M} p(f,\theta,\beta_R,\beta_M / g), \tag{26}$$

where θ is data hyperparameter, and β_R, β_M are the hyperparameter of the image prior density. Using Gaussian approximations of acquisition noise, and densities $p(g/f,\theta)$, $p(f/\beta_R,\beta_M)$, it easy to find the classical JMAP estimate, according to algorithm

$$\hat{f} = \arg\min_f \ln\left(\left\|g-Df\right\|_2^2\right) + \frac{r}{LN} + \ln\left(\left\|Rf\right\|_2^2\right) + \frac{m}{LN}\ln\left(\left\|Mf\right\|_2^2\right) \tag{27}$$

Unfortunately, the cost function in is non-convex and estimation is unstable. Here, to stabilize the JMAP estimation an improved solution by modelling the JMAP hyper parameters with a Gamma prior distribution is proposed. In the JMAP method, the density of the data or the prior density of the images is connected with the density of its respective hyperparameter. For example, consider only the use of the smoothness prior, $p\left(f|\beta_R\right)$, which enforces the HR images to be smooth. The associated hyperparameter β_R, defines "how smooth" is the resulting image. However, when an uniform density is assigned it is implicitly assumed that an over smooth image, like a constant intensity value image, when $\beta_R \to 0,$ is as likely to occur as a noisy image, like the one produced by a completely un-regularized estimation, when $\beta_R \to \infty,$. The other extreme choice for $p\left(\beta_R\right)$ is a Dirac delta function, i.e., an impulse in a fixed value for β_R. Among several candidates, the Gamma density has practical and theoretical advantages over the alternatives. The Gamma densities for the hyperparameters are given by $\rho\left(\theta = \frac{\theta^{(a-1)}b^{-a}}{\Gamma(a)}e^{-\frac{\theta}{b}}\right)$, and $p\left(\beta_{R,}\beta_M\right) = \frac{\beta_R^{(c-1)}\beta_m^{(h-1)}d^{-c_i-h}}{\Gamma(c)\Gamma(h)}e^{-\left[\frac{\beta_R}{d}+\frac{\beta_M}{i}\right]}$, where *a, c* and *h* are the scale factors, *b, d* and *i* are the shape factors, and $\Gamma(\mathrm{x})$ is the gamma function.

Using Gamma density functions in JMAP, the estimate it can be found as the solution:

$$\hat{f} = \arg\min_f \left(\left\|g-Df\right\|_2^2\right) + \mu_R\left(\left\|Rf\right\|_2^2\right) + \mu_M\left(\left\|Mf\right\|_2^2\right), \tag{28}$$

where parameters μ_R, μ_M depend on Gamma densities presented above.

Initial Conditions	n := 0; f_0 := initial HR image guess $\lambda_0^R =$ initial image smoothness parameter λ^R; $\lambda_0^M =$ initial motion similarity parameter λ^M; $b = D^T g$ calculate data	
initiate iterations	$A_n = D^T D + \lambda_n^R R^T R + \lambda_n^M M^T M$	MTM calculate matrix
	f_n solve via $CG\left(A_n f = b\right)$	new HR image
	$r_n^D = g - Df_n$, $r_n^R = Rf_n$	calculate data error and image smoothness
	$r_n^M = Mf_n$	calculate motion difference
	$\lambda_{n+1}^R = \mu R \left\| r_n^D \right\| 2 / \left\| r_n^D \right\| 2$, $\lambda_{n+1}^M = \mu M \left\| r_n^D \right\| 2 / \left\| r_n^M \right\| 2$	new λ^R and λ^M

Table 3. First implementation CG + parameter updating.

The first solution of the problem is presented in Table 3. The second approach (Table 4) has shown to be much faster than the first one.

Initial Conditions	$n = 0$ n := 0; $f_0 =$ HR image guess; $\lambda_0^R =$ image smoothness parameter λ^R; $\lambda_0^M =$ motion similarity parameter λ^M; $r_0 = D^T \left(Df_0 - g\right) + \lambda_0^R R^T Rf_0 + \lambda_0^M M^T Mf_0$ gradient $p_0 = -r_0$ initial search direction; $e_{0=} \left\| r_0 \right\|^2$	
NL-CG iterations	$h_n = D^T Dp_n + \lambda_0^R R^T Rp_0 + \lambda_0^M M^T Mp_0$	step search A
	$\tau_n = p_n^T r_n / p_n^T h_n$	step search B
	$f_{n+1} = f_n + \tau_n p_n$	HR image update
	$r_{n+1}^D = f_n + \tau_n p_n$; $r_n^D = Df_{n+1} - g$; $r_{n+1}^M = Mf_{n+1}$	gradient part update A; gradient part update B; gradient part update C.
	$\lambda_{n+1}^R = \mu R \left\| r_{n+1}^D \right\|_2 / \left\| r_{n+1}^R \right\|_2$, $\lambda_{n+1}^M = \mu M \left\| r_{n+1}^D \right\|_2 / \left\| r_{n+1}^M \right\|_2$	new λ^R and λ^M
	$r_{n+1} = D^T r_{n+1}^D + \lambda_{n+1}^R R^T r_{n+1}^R + \lambda_{n+1}^M M^T r_{n+1}^M$	final gradient update
	$e_{n+1} = \left\| r_{n'1} \right\|^2$, $\beta_n = e_{n+1} / e_n$, $p_{n+1} = -r_{n'1} + \beta_n p_n$	search direction update
	$n = n + 1$	

Table 4. Second Implementation: NL-CG (image and parameter updated together) Present the algorithm in adequate form.

The well known *Flowers* sequence was studied in simulation experiments where the motion was estimated using the optical flow method, and in this case, linear interpolated versions of the LR images were employed. The estimated motion vectors are not completely reliable therefore, occlusions and motion errors occur in several places in the sequence.

a) b)

Fig. 7. Visual results comparing classical JMAP and the Gamma_JMAP method in two different repetitions. (a) JMAP(SNR=18.6dB; b) Gamma_JMAP (SNR=19.3dB).

In simulations, different methods in SR were probed: JMAP - The classical JMAP approach. The L-MD - The method in based in the L-Curve, designed for multiple parameters in general inverse problems. The G_JMAP-1 - Described method with the procedure, as shown in Table 3. The G_JMAP-2 - Described method with direct minimization, as shown in Table 4. Fig. 7 exposes simulation results with the tested methods. It has been noticed the instability of the JMAP and of the L-MD methods.

5. DSP Implementation

Different promising algorithms of SR realized from LR images or frames from videos have been implemented in real time mode. The heart of the EVM DM642 is a Digital Media Processor, which is based in line of C64xx Digital Signal Processors (DSPs) manufactured by Texas Instruments. DM642 is characterized by a big set of integrated peripherals inside a chip, it includes three video ports interfaces, a I2C bus controller, a multichannel serial audio port, a 64-bit EMIF, a 10/100 Ethernet Controller MAC, and a PCI interface. Characteristics card includes: A TMS320DM642 DSP at 720 MHz, 32 Mb in SDRAM, 4 Mb in Linear Memory Flash, 2 video decoders, 1 video coder, FPGA implementation to screen display, double UART with RS-232 drivers, an stereo codec, Ethernet card, 32 Kb EEPROM I2C, 8 programmable LEDs, several input-output video formats, etc. (TMS320DM642, 2004, Enry Shen, 2005). Communication of Code Composer Studio with EVM is achieved using an external emulator via JTAG connectors. Figure 8 exposes the EVM DM642 block diagram architecture. To get the processing time values, note that the TMS320DM642 DSP has a clock of 720 MHz, realizing 1.39 instructions per cycle, obtaining 570 millions instructions/sec. It is important to clarify that it can get up maximum 2,147,483,648 instructions per second.
Using Simulink module of Matlab a project is created, where the DSP model (in this case DM342EVM) and its respective Task Bios are selected; later, this bios configuration can be changed on Code Composer Studio. Inside of the function, there are three modules: video capture, subsystem realizing SR reconstruction on base of Wavelet frameworks, and video display. Next, a CCS project is formed in Simulink. The Matlab sends call to CCS, and send

the project on C. To realize the video sequence processing on DSP, first, it should be changed the MatLab program into "C" code for Code Composer Studio via Simulink. Once the project Composer Code Studio has been created, the necessary changes are arranged with purpose to obtain the processing time values. The results of time execution of the designed and reference frameworks are exposed in the next section.

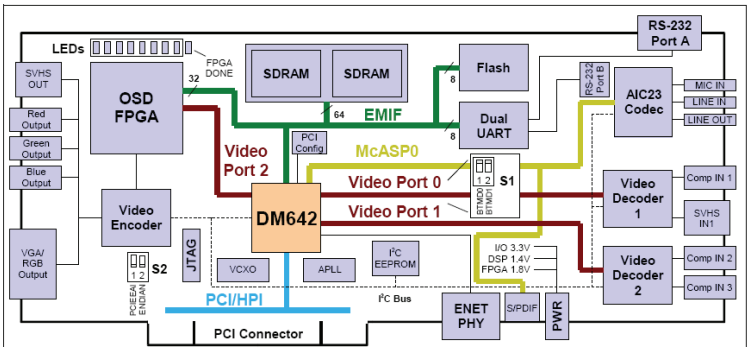


Fig. 8. Block diagram of the EVM DM642.

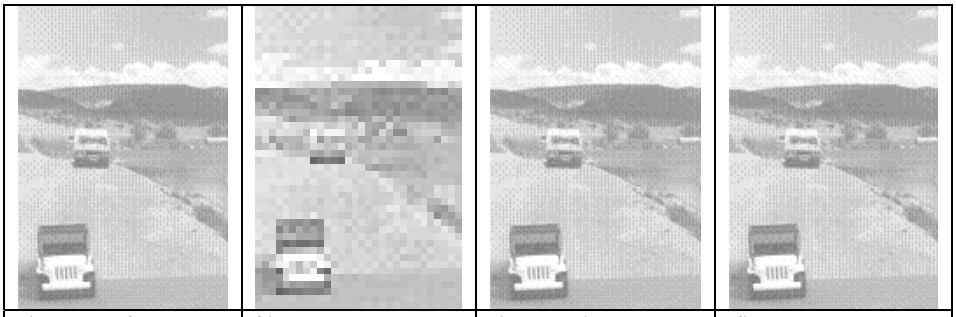
6. Simulation Results

Numerous statistical simulations, which have been realized, are consisted of the tests of the SR procedures in several video sequences that are widely used in the experiments: *Toy* (256x256 pixels, 8 bits), *Stephan* and *Flowers* (352 x 240 pixels, 8 bits). Additionally, the *Flowers* sequence has been reconstructed in gray scale and color formats. The first video sequence, contain an image with toys moving with defined borders, and plain background. The second one shows a tennis player, there are a lot fine details at objects like the racket, faces of the audience, etc., they are moving in several directions, moreover, the edges of the letters in the stands are exposing a visual reference. The third sequence exposes a tree with edges and field of numerous flowers that contain a lot of fine details. This sequence is used in gray and color formats. In all cases, the initial LR images are obtained reducing the original image HR size in four times, applying the summing and averaging the values of every four pixels. The reconstruction SR process should restore the initial sizes, applying different SR techniques. Finally, the original HR and reconstructed SR images are compared. The objective qualities of SR for the proposed and reference algorithms are applied according to criteria: *PSNR, MAE,* and color *NCD* measure. We also use the subjective visual comparison in form of error image to compare the capabilities of noise suppression and the artifacts´ limitations, also, the detail´s preservation of the different algorithms. There were realized numerous simulation experiments using different methods of SR reconstruction, but for each video sequence we only present below the better selected results that put in following tables (from 5 to 8) and figures (from 9 to 12). So, the reconstructed results in the SR problem for different test images using following techniques: Bi-cubic, Nearest neighbor, Warp, DCT, FFT, Sinc, Fuzzy-ELA, Recursive logic, and Wavelets based on classical families Biorthogonal, Daubechies, Symlets, Coiflets, and finally, the framework proposed show the best values in criteria PSNR and MAE for different WAF: Ξ_n, fup_n, π_n, g_n, and up_n. Finally, the

values of needed processing time for SR procedures implemented in hardware are exposed for better algorithms.
In the SR frame of the video sequence *Toy*, one can observe the simulation results in Table 5 and Fig.9 comparing the original HR and LR images that the better performance in terms of objective criteria PSNR and MAE, as well as in subjective perception are presented employing SR reconstruction on the base of WAF fup_1 and DCT algorithm. It is clearly observed in the images (see Fig.9) better subjective perception for WAF fup_1 in comparison with DCT technique.

		Algorithm	MAE	PSNR	Algorithm	MAE	PSNR
No. of frame	1	fup_3	*12.36*	*34.76*	**DCT**	*12.09*	*34.90*
	2		*12.67*	*34.65*		*12.45*	*34.59*
	3		*12.78*	*34.56*		*12.43*	*34.61*
	1	fup_1	**11.38**	**34.91**	**Daubechies**	13.20	34.50
	2		**11.35**	**34.90**		13.44	34.37
	3		**11.51**	**34.90**		13.55	34.34

Table. 5. Objective criteria values for video sequence *Toy*.



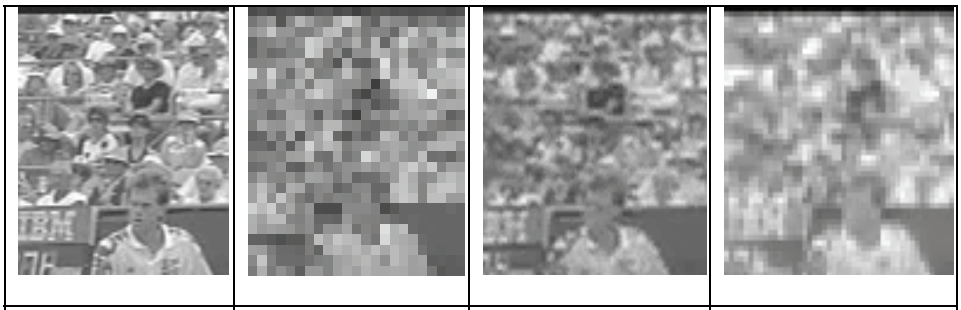
a) Original HR. b) Input LR. c) WAF fup_1. d) DCT.

Fig. 9. Visual perception results for the video sequence *Toy*.

Video sequence *Stephan* has several features, one of which is that there are several movements in it as above mentioned. The simulation results expose the best reconstruction in SR process in the case of WFA usage. This is confirmed as in objective criteria for fine details (see Table 6, MAE), as in subjective perception analyzing SR reconstructed images and corresponding error images (Fig.10).
In the video sequence *Flowers* (see Table 7 and Fig.11), there is clear the difference between LR image and HR one, because of numerous fine features presented in the flowers. Also, there are the well-defined borders of the house. Better results among all analyzed algorithms in terms of the objective criteria are presented by DCT algorithm and when the WAF π_6 is employed, but observing the SR images and their error images it can viewing that mentioned WAF exposes in some areas slightly better visual subjective performance.

No. of Frame		Algorithm	MAE	PSNR	Algorithm	MAE	PSNR
	100		**5.10**	**80.10**		*6.59*	*80.43*
	200		**3.01**	**82.29**		*3.98*	*83.44*
	300	π_5	**2.61**	**83.05**	biorthogonal	*4.81*	*82.23*
	100		***5.90***	***79.44***		51.73	72.38
	200		***3.18***	***82.01***		23.68	73.47
	300	Ξ_3	***2.88***	***82.48***	Coiflets	29.90	72.77

Table. 6. Objective criteria values for video sequence *Stephan*.



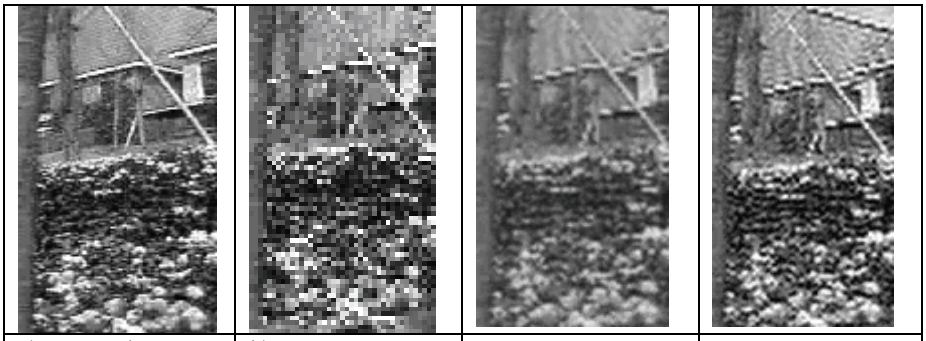
a) Original HR. b) Input LR. c) WAF π_5. d) Biorthogonal.

Fig. 10. Visual perception results for the video sequence *Stephan*.

The main difference of the color video sequence *Flowers* (Fig.12) in comparison with their gray variant (Fig.11) is additional color information that permits to see more precisely the fine details for flowers in different colors. For example, analyzing SR image restored employing the WAF π_5 or Ξ_3, one can see better visual performance in comparison with any another SR restoration algorithm. For example, SR procedure based on Biorthogonal Wavelet presents blurry frames with the pixels having a sideways movement type, moreover, the values in the error images are greater for the SR algorithm in case Biorthogonal than when the SR based on the WAF π_5 and Ξ_3 are employed.

No. Of frames		Algorithm	MAE	PSNR	Algorithm	MAE	PSNR
	10		11.40	76.88		**9.57**	**78.27**
	15		9.11	78.18		**8.65**	**78.55**
	20	π_5	9.97	77.31	DCT	**8.05**	**78.66**
	10		*9.97*	*77.31*		63.08	72.53
	15		*9.11*	*78.18*		54.22	72.60
	20	π_6	*11.40*	*76.88*	Coiflets	59.97	72.51

Table. 7. Objective criteria values for video sequence *Flowers* (gray).



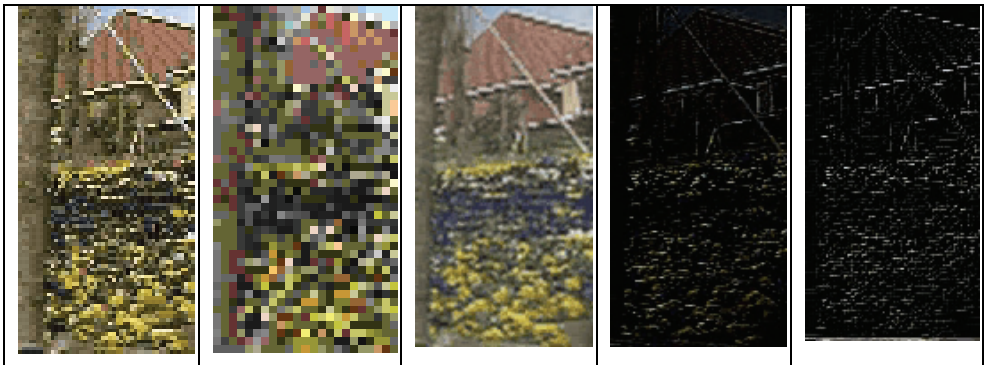
a) Original HR. b) Input LR. c) WAF π_5. d) DCT.

Fig. 11. Visual perception results for the video sequence *Flowers*.

No. of Frame		Algorithm	MAE	PSNR	NCD	Algorithm	MAE	PSNR	NCD
	10		**11.00**	**79.17**	**0.09**		***11.08***	***76.74***	***0.197***
	15		**10.24**	**79.43**	**0.09**		***12.12***	***76.24***	***0.205***
	20	**Biorthogonal**	**9.88**	**79.49**	**0.08**	Ξ_3	***10.52***	***76.73***	***0.182***

Table. 8. Objective criteria values for video sequence *Flowers* (colour).

Finally, the results of real time implementation are presented for the SR algorithms on DSP. Table 9 exposes the values of processing time for different better SR algorithms tested here. The first and second columns mark the class and type of algorithm, in the third and fourth columns, the results obtained in Matlab implementation are presented; the fifth and sixth columns show the DSP processing time values, and, finally the seventh and eighth columns view the processing results on DSP serial processing.



a) Original HR. b) Input LR. c) WAF Ξ_3. d) WAF Ξ_3. Error image. Biorth. Error image.

Fig. 12. Visual results for the *Flowers* (colour) sequence.

One can observe that dividing the WAF configuration process into two parts, the serial processing on DSP reduces the processing time values, presenting the better processing

performance, resulting in 25.43 frames per second for WAF technique. This can be considered as practically on line real-time processing.

Processing Time Values, sec				Processing Time Values, sec			
Time in seconds		*Matlab*		*DSP*		*Serial Processing on DSP*	
Type	*Algorithm*	*Time/ frame*	*Frame/ sec*	*Time/ frame*	*Frames/s*	*Time/frame*	*Frames/ s*
Others	***Bicubic***	*0.03*	*33.33*	*0.07074*	*14.1363*	*0.03338*	*29.958*
	Nearest neighbour	*0.03*	*33.33*	*0.07074*	*14.1363*	*0.03358*	*29.780*
	Bilineal	*0.03*	*33.33*	*0.07074*	*14.1363*	*0.03338*	*29-958*
	DCT	*0.084*	*11.90*	*0.20028*	*4.9930*	*0.03343*	*29.913*
Wavelets	***Biorth.***	*0.075*	*13.33*	*0.17685*	*5.6545*	*0.02989*	*33.456*
	Coiflets	*0.075*	*13.33*	*0.17685*	*5.6545*	*0.02989*	*33.456*
	Daubechies	*0.075*	*13.33*	*0.17685*	*5.6545*	*0.02989*	*33.456*
	Symlets	*0.075*	*13.33*	*0.17685*	*5.6545*	*0.02989*	*33.456*
WAF	$\Xi_i(x)_{,}\ i=\overline{2,4}$	*0.04*	*25.00*	*0.09432*	*10.6022*	*0.03932*	*25.432*
	$fup_i, i=1,4;8$	*0.04*	*25.00*	*0.09432*	*10.6022*	*0.03932*	*25.432*
	$g_i,\quad i=\overline{2,6}$	*0.04*	*25.00*	*0.09432*	*10.6022*	*0.03932*	*25.432*
	$\pi_i,\quad i=\overline{2,6}$	*0.04*	*25.00*	*0.09432*	*10.6022*	*0.03932*	*25.432*
	$up_i,\quad i=\overline{2,4,7,8}$	*0.04*	*25.00*	*0.09432*	*10.6022*	*0.03932*	*25.432*

Table 9. Values of processing time in hardware implementation.

7. Conclusion

It has realized a review of several promising SR methods. Some of them are usually focused on solving the issue of super resolution only for some specific type of image, or images that are used in specific applications, others, need to have additional prior information about the image, perform some sort of convergence of information, or realize some training, this, before processing the information and get the SR image. The proposed method can be applied to any kind of image or video sequence frame without any a priory information, permitting to realize the SR process over the region of interest of an image, a sequence of images and video, it is not depended on the type of application where it was obtained, and it much less interfere with results from the final purpose image. Additionally, numerous simulation and real time implementation results have shown that the proposed framework based on the WAFs is effective in performing the image registration and super-resolution for different real-life video sequences, demonstrating better robust performance in the frames with different nature and texture characteristics, such as edges, fine details, and different types of movements. Real time implementation of the proposed framework on

MatLab and DSP platforms has confirmed the processing velocity of about 25 frames/sec for all investigated video sequences.

Acknowledgement

The authors thank the National Polytechnic Institute of Mexico and CONACYT (grant 81599) for their support to realize this work.

8. References

Akgun T., Altunbasak Y., Mersereau R. M. (2005). Super-resolution reconstruction of hyperspectral images, *IEEE Trans. on Image Proc.* Vol.14, No.11, 2005, pp.1860–1875, ISSN: 1057-7149.

Baboulaz L. and Dragotti P. L. (2009). Exact feature extraction using finite rate of innovation principles with an application to image super-resolution. *IEEE Trans. on Image Proc.*, vol. 18, No. 2, 2009, pp. 281–298, ISSN: 1057-7149.

Bovik A.,Ed.(2000).*Handbook of Image and Video Process*.Academic.ISBN:0-12-119790-5, MA.

Callico G.M., Lopez S., Sosa O., Lopez J.F., and Sarmiento R. (2008). Analysis of fast block matching motion estimation algorithms for video super-resolution systems. *IEEE Trans. on Consumer Elec.*, Vol.54, No. 3, 2008, pp.*1430-1438,* ISSN*:* 0098-3063*.*

Chan R., Chan T., Shen L., Shen Z. (2003). Wavelet algorithms for high-resolution image reconstruction, *SIAM Journal on Scientific Computing*, Vol. 24, No.4, 2003, 1408–1432, ISSN: 1064-8275.

Chaudhuri S. (2001). *Super-Resolution Imaging*, Kluwer Academic Publ., ISBN: 0-7923-7471-1. MA, USA.

Chaudhuri S. and Manjunath J. (2005). *Motion-Free Super-Resolution*. Springer, ISBN-13: 978-0387-25587-3.New York.

Crouse M. S., Nowak R. D., and Baranuik R. G. (1998). Wavelet based signal processing using hidden Markov models, *IEEE Trans. on Signal Proc.* Vol.46, No. 4, 1998, pp. 886-902. ISSN: 1053-587X.

Enry Shen. Code Compouser Studio (optimization), Texas Instruments (2005), *Proc. Of 8th Texas Instruments Developer Conference,* India, Bangalore, 2005.

Farsiu S., Robinson D., Elad M., and Milanfar P. (2004).Advances and chanllenges in super-resolution,"*Int.J.Imaging Syst. Techn.* Vol.14, No.2, 2004, pp.47–57, ISSN: 0899-9457.

Farsiu S., Elad M., and Milanfar P. (2006). Video-to-video dynamic superresolution for grayscale and color sequences. *EURASIP J. Appl. Signal Process*. Vol.10, 61859_1-10, 2006, ISSN: 1110-8657.

Flusser F., Sroubek, J. (2003). Multichannel blind iterative imagerestoration. *IEEE Trans. Image Process*. Vol.12, No.9, 2003, pp.1094–1106, ISSN:1110-8657.

Franzen O., Tuschen C., Schroeder H.(2001).Intermediate image interpolation using polyphase weighted median filters, *Proc. SPIE*, vol. 4304, pp.306–317, ISSN 0277-786X. 2001.

Gomeztagle F., Kravchenko V., and Ponomaryov V. (2009). Super-Resolution Procedures in Images and Video sequences Applying Wavelet Atomic Functions" in *Proc. of IASTED 20th Symposium on Modelling and Simulations*, pp.187-189, ISBN: 978-0-88986-798-7. Banff, Alberta, Canada, 2009.

Gulyaev Yu.V., Kravchenko V.F., Pustovoit V.I. (2007). A New Class of WA-Systems of Kravchenko–Rvachev Functions, *Doklady Mathematics*, Vol.75, No.2, 2007, pp.325–332, ISSN:1064-5624.

Hou H. S. and Andrews H. C. (1978).Cubic splines for image interpolation and digital filtering, *IEEE Trans. on Accoust., Speech and Sig. Proc.*, Vol.26, No.6, 1978, 508– 517, ISSN:0096-3518.

Jain J.R. and Jain A.K. (1981). Displacement measurement and its application in interframe image coding. *IEEE Trans. On Commun.*, Vol. 29, No. 12, 1981, pp. 1799-1808, ISSN: 0090-6778.

Juarez C., Ponomaryov V., and Kravchenko V.F. (2008). Wavelets Based on Atomic Function used in Detection and Classification of Masses in Mammography, *Lect. Not. in Art. Intel.*, LNAI 5317, 2008, pp.295-304, ISSN:0302-9743.

Katsaggelos A., Molina R., and Mateos J. (2007). *Super Resolution of Images and Video, Synthesis Lectures on Image, Video, and Multimedia Processing*. Morgan & Claypool, ISBN:1-5982-90843. New York.

Kravchenko V.F., Ponomaryov V.I., Sanchez-Ramirez J.L. (2008). Properties of Different Wavelet Filters used for Ultrasound and Mammography Compression, *Telecom. and Radio Engineering.*, Vol.87, No.10, 2008, pp.853-865, ISSN:0040-2508.

Kravchenko Victor, Perez-Meana Hector, and V .Ponomaryov. (2009). *Adaptive Digital Processing of Multidimensional Signals with Applications*, FizMatLit Edit., ISBN:978-5-9221-110-6. Moscow (available in http://www.posgrados.esimecu.ipn.mx).

Landi G. and Loli Poicolomimi E. (2006). Representation of High Resolution Images from Low Sampled Fourier Data: Applications to Dynamic MRI. *J Math Imaging Vision*, 2006, pp. 27–40, ISSN:0924-9907.

Lertrattanapanich S., Bose N. K. (2002). High resolution image formation from low resolution frames using Delaunay triangulation, *IEEE Trans. on Image Proc.,* Vol. 11, No.12, 2002, pp. 1427–1441, ISSN: 1057-7149.

Luisier F., Blu T., and Unser M. (2007). A new sure approach to image denoising: Inter-scale orthonormal wavelet thresholding. *IEEE Trans. on Image Proc.,* V. 16, No. 3, 2007, pp. 593–606, ISSN: 1057-7149.

Ng M. K., Sze C. K., Yung S. P. (2004). Wavelet algorithms for deblurring models. *Intern. Jour. of Imaging Systems and Techn*. V. 14, No.3, 2004, pp.113–121, ISSN 0010-4620 .

Maeland E. (1998). On the comparison of interpolation methods, *IEEE Trans. on Med. Imag.* Vol.7, No.3, 1998, pp.213–217, ISSN: 0278-0062.

Meyer Y.(1990). *Ondelettes,* Hermann, Paris, France (in French), ISBN: 0-8218-13870.

Park S., Park M., and Kang M. (2003). Super-resolution image reconstruction: A technical overview. *IEEE Sign. Proc. Mag.*, vol. 20, No. 3, 2003, pp. 21–36, ISSN: 1053-5888.

Phu M.Q., Tischer P.E., Wu H.R. (2004). A median based interpolation algorithm for deinterlacing, *Proc. of Int. Symp.on Intellig. Sign. Proc. and Commun. Systems,* pp.390–397, ISBN: 0-7803-8639-6. Seoul. 2004.

Protter M., Elad M., Takeda H., and Milanfar P. (2009). Generalizing the non-local-means to super-resolution reconstruction. *IEEE Trans on Image Proc.*, vol. 16, No. 2, 2009, pp. 36–51, ISSN: 1057-7149.

Reichenbach S.E.. Geng F. (2003). Two-dimensional cubic convolution. *IEEE Trans on Image Proc*. Vol.12, No.8, 2003, pp.857–865, ISSN :1057-7149.

Qin Feng-qing, He Xiao-hai, Chen Wei-long, Yang Xiao-min, Wu Wei. (2009). Video superresolution reconstruction based on subpixel registration and iterative back projection. *J. of Elect. Imaging,* Vol.18, No.1, 2009, 013007_1-11, ISSN: 1017-9909.

Sanchez-Beato A. and Pajares G. (2008). Noniterative Interpolation-Based Super-Resolution Minimizing Aliasing in the Reconstructed Image. *IEEE Trans. On Image Proc.*, Vol.17, No. 10, 2008, p.1817-1826, ISSN : 1057-7149.

Shen Huanfeng, Zhang Liangpei, Huang Bo, and Li Pingxiang. (2007). A MAP Approach for Joint Motion Estimation, Segmentation, and Super Resolution. *IEEE Trans. On Image Proc.*, V.16, No. 2, 2007, pp. 479-490, ISSN : 1057-7149.

Sroubek F., Flusser J. (2006). Resolution enhancement via probabilistic deconvolution of multiple degraded images,Institute of Information Theory and Automation, *Pattern Recognition Letters,* Vol. 27, No.4 , 2006, pp.297-283, ISSN: 0167-8655.

TMS320DM642 Evaluation Module with TVP Video Decoders (2004). *Technical Reference* 507345-0001 Rev. B, December 2004.

Tolpekin V.A and Hamm N.A.S. (2008). Fuzzy super-resolution mapping based on Markov random fields. *Proc. Of IEEE Int. Geosc. & Remote Sensing Symp.* pp.96-99, ISBN: 978-1-4244-2807-6. 2008.

Wang C. and Xue P. (2006). Improved super-resolution reconstruction from video. *IEEE Trans.Circuits Syst. Video Technol.* Vol.16, No.11, 2006, pp.1411–1422,ISSN:1051-8215.

Wood S.L., Lee S.-T., Yang G., Christensen M.P., Rajan D. (2008). Impact of measurement precision and noise on superresolution image reconstruction, *Appl. Opt*., Vol. 47, No.10, 2008, pp.1638–1648, ISSN: 0003-6935.

Wood Sally L. (2009). Super-resolution image reconstruction for a steerable array of sub-imagers. *Digital Signal Processing.* Vol. 19, , No.6, 2009, pp. 923–933, ISSN:1051-2004.

Wüst Zibetti M. V., and Mayer Joceli A. (2007). Robust and Computationally Efficient Simultaneous Super-Resolution Scheme for Image Sequences. *IEEE Trans. On Circ. And System for Video Tech.*, Vol. 17, No. 10, 2007, pp.1288-1300, ISSN: 1051-8215.

Zhang Liangpei, Zhang Hongyan, Shen Huanfeng, Li Pingxiang. (2010). A super-resolution reconstruction algorithm for surveillance images. *Signal Processing*. 2010. doi: doi:10.1016/j.sigpro.2009.09.002, Ref.: SIGPRO3898, ISSN:0165-1684.

Zibetti W., M.V., Bazan V. F.S., Mayer, J. (2010). Estimation of the Parameters in Regularized Simultaneous Super-resolution, *Pattern Recognition Letters*. 2009, doi: 10.1016/j.patrec.2009.12.009. ISSN: 0167-8655.

7

Order Statistics - Fuzzy Approach in Processing of Multichannel Images and Video Sequences

Francisco Gallegos, Volodymyr Ponomaryov and Alberto Rosales
National Polytechnic Institute of Mexico
Mexico

1. Introduction

There exist different applications of the image processing, such as medical imaging, high definition television, virtual reality, remote sensing, ultrasound and radar imaging, etc. In these applications, it is necessary to restore an image (or frames of video sequence) and decrease a noise influence exploiting the filtering algorithms that form a part of a general image processing system. The images are corrupted by noise, in sensors employed or maybe, during signal transmission. Also, several kinds of noises are produced by natural phenomena (atmospheric, scattering, interference, etc.). Usually, real noises are described by different models, there exist impulsive, additive and multiplicative (speckle) ones. So, the image pre-processing efficient scheme should be one of important part in any vision application permitting to suppress a noise, saving the image performances, such as, edge and fine features preservation, and also the chromaticity properties for the multichannel (multispectral) images. This demands to have several efficient filtering schemes, which depend on noise type and priory information, in a pre-processing stage of image or video sequence processing system. The main objective of present chapter is to exhibit several justified approaches in restoration of the images and video sequences, which are usually affected by noise of different nature, which can be efficiently used in different applications of the multichannel (multispectral) images and sequences. Here, multispectral image is defined in such a way, where each a pixel is represented by a number of channels that carry out information about its spectral content. Multispectral images span the domain of images from conventional three-channel colour images to hyperspectral imagery with hundred of bands/channels used in remote sensing applications, medicine, spectrometry, etc.
In literature, there exist a lot of algorithms that process two dimensional (2D) images (Franke et al., 2000); (Russo & Ramponi, 1996); (Schulte et al., 2006, 2007a, 2007b, 2007c); (Shaomin & Lucke, 1994); (Nie & Barner, 2006); (Morillas et al., 2005, 2006, 2008a, 2008b, 2009); (Camarena et al., 2008, 2010); (Ma et al., 2007); (Amer & Schröder, 1996); (Xu, 2009). We compare the proposed 2D fuzzy framework with recently presented *2D-INR* filter based on fuzzy logic (Schulte et al., 2007b), where a noise is detected preserving the fine features and edges in an image. Also, other promising classes of 2D processing algorithms are employed as comparative ones: *2D-AMNF*, *2D-AMNF2* (Adaptive Multichannel Filters)(Plataniotis & Androutsos et al., 1997); (Plataniotis & Venetsanopoulos, 2000); *2D-*

AMNIF (Adaptive Multichannel Filter using Influence Functions) (Ponomaryov et al., 2005); (Plataniotis & Venetsanopoulos, 2000); *2D-GVDF* (Generalized Vector Directional Filter) (Trahanias & Venetsanopoulos, 1996); *2D-CWVDF* (Centered Weighted Vector Directional Filters) (Lukac et al., 2004); and finally, *2D-VMF_FAS* (Vector Median Filter Fast Adaptive Similarity) (Smolka et al., 2003). These techniques have demonstrated the better results among a lot of others known in literature. The principal drawback of all 2D processing algorithms is that they use only one frame of a video sequence and principally cannot use temporal correlation that exists between neighbouring frames to distinguish and decrease noise or motion in an image. This does not permit to suppress a noise efficiently, as well as to preserve the fine image features and restore the image chromaticity properties. Temporal information should be taken into consideration in the processing of the neighbouring frames together but straight averaging in temporal area the corresponding pixels to smooth a noise may introduce "ghosting" artifacts in the presence of camera and object motion. Such artifacts can be removed by motion compensation where a number of algorithms have been proposed with different computational complexity (Balster et al., 2006); (Jovanov et al., 2009); (Kravchenko et al., 2009); (Mélange et al., 2008); (Zlokolica et al., 2005). Thus, a desirable video noise filter should distinguish noisy and motional pixels as well as collect enough similar pixels adaptively from temporal to spatial directions.

In this chapter, the fuzzy set theory and fuzzy logic that offer us a powerful tool for representing and processing human knowledge and intuition, incorporating them into the design are employed, which cannot be done using classical mathematical modelling. The fuzzy metric is considered more effective in comparison with classical measures, moreover, due to the non-stationarity of images and serious problems in distinguishing between noise, motions, and fine features and edges, fuzzy modelling is considered quite appropriate in video sequence filtering. Here, classical binary decisions are replaced by a gradual transition, which is more appropriate when dealing with complex systems.

Unfortunately, a methodology, which gathers the advantages of each one of powerful techniques (order vector statistics and fuzzy set theory) usually employed in processing of images or video sequences, providing the better suppression noise capability, fine features preservation, as well as chromaticity characteristics, is not developed sufficiently. In present chapter, promising scheme is designed unifying the directional gradients and pixel angular divergence measure together with the robust vector order statistics processing techniques described previously (Ponomaryov, 2007); (Ponomaryov et al., 2010). The employing the designed fuzzy rules with fuzzy measure of a motion in a form of the membership degree in a 3D sliding-window gives the opportunity to preserve well the fine image features and restore the chromaticity properties. General operations of novel approach consist of the selection made in fuzzy means for any spectral band of an image: if there exist the edges and fine features, or noise, or may be some movement in the central pixel into sliding processing window. So, the framework does it possible to distinguish these characteristics inherent in multispectral images (or frames) using fuzzy rules designed in this chapter. They are applied to fuzzy-directional values to resolve the hypothesis: if a central pixel component is a corrupted one or not. In case of a corrupted pixel happened, some procedures in substitution of a central component with one of its neighbours are realized according to justified in fuzzy matter selection.

We also realize the adaptation of several 2D algorithms in filtering of 3D video data: *3D-MF*, *3D-VGVDF* (Trahanias & Venetsanopoulos, 1996), *3D-VVMF* and *3D-VVDKNNVMF* (Ponomaryov, 2007). Additionally, we have implemented the 3D-VKNNF, 3D-VATM (Zlokolica et al., 2006), and *3D-VAVDATM* filters (Ponomaryov, 2007). Other fuzzy Logic techniques 3D algorithms (Saeidi et al., 2006), and (Zlokolica et al., 2006) are analyzed during modelling and in the simulation experiments. The first framework that is used to smooth Gaussian noise is the designed ***FDARTF_G*** (Fuzzy Directional Adaptive Recursive Temporal Filter for Gaussian Noise) that preserves the fine features, edges and chromaticity properties, and the second one, ***3D-FCF*** (Fuzzy Temporal Spatial Colour Filter) operates in similar way as ***FDARTF_G*** only with some modifications for impulsive noise decreasing. To justify the effectiveness of introduced 3D techniques, the comparison with the better filtering frameworks that exist in video sequence processing were used (Zlokolica et al., 2006); (Ponomaryov et al., 2009); (Schulte et al., 2006b); (Schulte et al., 2006a); (Mélange et al., 2008). Reference filters: "Fuzzy Motion Recursive Spatio-Temporal Filter" (*FMRSTF*) (Zlokolica et al., 2006); an adaptation of *FMRSTF* employing only angles instead of gradients, named as "Fuzzy Vectorial Motion Recursive Spatio-Temporal Filter" (*FVMRSTF*); "Video Generalized Vectorial Directional Processing" (*VGVDF*) (Trahanias et al., 1996), "Video Median M-type K-Nearest Neighbour" (*VVDKNNVMF*) described in (Ponomaryov, 2007) were used as comparative in suppression of Gaussian noise, and algorithms *3D-MF, 3D-VGVDF, 3D-VVMF, 3D-VVDKNNVMF, 3D-VKNNF, 3D-VATM, 3D-VAVDATM* filters were used as comparative ones to evaluate ***3D-FCF*** rendering during the simulations and modelling experiments. Numerical simulations have shown the better performance of original framework that outperforms existed methods in suppression of a noise of different nature increasing performances of a colour image and/or video data. The objective criteria used in modelling and simulation experiments of the different filtering algorithms are the Peak Signal-to-Noise Ratio (*PSNR*), Mean Absolute Error (*MAE*) and Normalized Colour Difference (*NCD*), (Plataniotis & Venetsanopoulos, 2000); (Ponomaryov, 2007). Additionally, the subjective visual criterion in form of error of reconstructed multichannel image is used.

Several designed promising algorithms as well as better existed ones were implemented on the DSP platform realizing analysis of the sequences or images in a real time environment (Mullanix et al., 2003); (Gallegos-Funes et al., 2009); (Kravchenko et al., 2009).

The current chapter is organized as follows: Sec. 2 presents the model of noise usually employed in image processing applications and defines the objective criteria: *PSNR*, *MAE* and *NCD*. Sec. 3 exposes some promising recent schemes for simultaneous processing of different kinds of 2D-3D images and video sequences corrupted by noises (Gaussian and impulsive). Sec. 4 explains the original 2D-3D procedures to suppress additive and impulsive noises using two neighbouring frames for the motion, fine detail and edges, and noise detection in multichannel images and video sequences. Here, the numerous experimental results of modelling and simulations in form of the objective and subjective measures are presented, justifying the effectiveness of several proposed and existing approaches, and also the implementation of the better promising algorithms on the DSP platform realizing analysis of the sequences or images in a real time environment is discussed. A brief conclusion is drawn in Sec. 5.

2. Noise and Performance Criteria

Real-world still images and video sequences are affected by random fluctuations in intensity, colour, texture, object boundary, or shape, and also by blurring blocking, and colour distortions. There are a lot of complex reasons for these fluctuations and distortions, often due to factors, such as non-uniform lighting, random fluctuations in object surface orientation and texture, sensor limitations, etc. The processing of such images or frames in video sequences can be treated as a problem of statistical inference, which requires the definition of a statistical model corresponding to the image and noise pixels employing the random field models. Combined with various frameworks for statistical inference, such as maximum likelihood (ML) and Bayesian estimation, random field models are used in image restoration, enhancement, classification, segmentation, compression and synthesis. The general model of image-noise representation consists of the random field definition that represents the multidimensional signal and the random process, together with the joint density that models the corruption mechanism (Bovik, 2000).
Images are relatively broadband signals where the visual information may be at mid-to-high spatial frequencies, and significant image details: edges, lines, and textures typically contain higher frequencies. The classical but no efficient approach in noise suppression influence is the linear filtering algorithms where for a given filter type, different quality of smoothing can be received by adjusting the bandwidth of a linear filter.

2.1 Additive Noise

Optimal methods of linear filtering theory is useful when the corruption could be represented as a Gaussian process and the criterion of accuracy is the mean square error (*MSE*), but this assumption is not correct in most applications, for example in digital systems. Gaussian noise is a part of almost any signal where an additive Gaussian noise generally assumes zero-mean Gaussian distribution and is usually introduced during video acquisition. The additive model is most appropriate when the noise in this model is independent of an image. There are many applications of the additive model: thermal noise, photographic noise, and quantization noise, etc.

2.2 Impulsive Noise in Image

It is assumed that the noise process is impulsive noise if as a result many of the signal values do not change at all or change slightly and some signal values change dramatically, in other words, the change is clearly visible (Astola & Kuosmanen, 1997). In practice, the same number of bits is used to represent the noisy and the noise-free signal, usually 8 bits or 256 levels 0, 1,..., 255. The realistic impulsive noise is modelled as bit errors in the signal values during transmitting the images or video sequences over noisy digital links. It is easy to calculate for a binary symmetric channel with a given crossover probability that the contribution to the *MSE* from the most significant bit is approximately 3 times that of all the other bits. Impulses are also referred to as outliers.
Several types of impulsive models usually can be used. Some of them need the detail a priori information about the degradation process in each a channel for multichannel (or colour) multidimensional image. In our opinion, the complex models that need several parameters, which should be determined a priori or during the processing stage, have low tolerance, and so, such a model can produce confusion during the interpretation of filtering

results (Ponomaryov at el, 2005); (Ponomaryov, 2007); (Kravchenko et al., 2009). Below, we use the simple and in the same time the most severe model of impulsive noise from point of view of image degradation. This model needs only prior information about the probability p of random spikes appearance, which are independent in each a channel. Additionally, the amplitude of impulsive noise is modelled as uniformly distributed random value within the interval of given values (0-255) for each a channel in the case of colour images.

2.3 Mathematical Solutions Applied in Image-Noise Models

We use the simplest model for additive Gaussian noise degradation

$$x_{\Sigma}(i,j) = x_0(i,j) + n(i,j), \tag{1}$$

where $x_0(i,j)$ is original image (or sequence frame), $x_{\Sigma}(i,j)$ is degraded image, and $n(i,j)$ is Gaussian additive noise. Also, such a model for noise influence in the case of impulse noise degradation is employed (Ponomaryov, 2007); (Kravchenko et al., 2009):

$$x(i,j) = n_i\left(x_0(i,j)\right),\ n_i\left(x_0(i,j)\right) = \begin{cases} \text{random values with probability } P \\ x_0(i,j) \text{ another case} \end{cases}, \tag{2}$$

where $x_0(i,j)$ is original image (or sequence frame), $x(i,j)$ is degraded image, and $n_i\left(x_0(i,j)\right)$ the above presented function.

In the case of multiplicative noise degradation, the model (2) can be represented in the form (Kravchenko et al., 2009):

$$x_{speckle}(i,j) = n_i\left(\varepsilon_m(i,j)\cdot x_0(i,j)\right), \tag{3}$$

where $n_m(i,j)$ denote multiplicative (speckle) noise.

The eqs. (1-3) represent the basic models in degradations by noise. For multichannel images it is necessary to apply eq. (2) for each a channel.

In the case of multidimensional image representation, the model (2)-(3) is changed, and for 3D discrete image can be rewritten as follows:

$$x_{speckle}(i,j,k) = n_i\left(x_0(i,j,k)\cdot\varepsilon_m(i,j,k)\right), \tag{4}$$

where $n_i\left(x_0(i,j,k)\right)$ is the functional $n_i\left(x_0(i,j,k)\right) = \begin{cases} \text{noise } \mathrm{n_i} \text{ with probability } p \\ x_0(i,j,k), \text{ otherwise} \end{cases}$, and $x_{speckle}(i,j,k)$ is a noisy observation (i.e., the recorded image) of the 3-D function $x_0(i,j,k)$

(i.e., the noise-free image that has to be recovered), $\varepsilon_m(i,j,k)$ is the corrupting multiplicative (speckle) noise component.

2.4 Objective and Subjective Criteria

To model and evaluate different filters and compare their performances, several criteria are used, such as: the peak signal-to-noise ratio (*PSNR*) for the evaluation of noise suppression; the mean absolute error (*MAE*) for quantification of edges and fine feature preservation and the normalized colour difference (*NCD*) (Plataniotis & Venetsanopoulos, 2000):

$$PSNR = 10 \cdot \log\left[\frac{(255)^2}{\mathrm{MSE}}\right]\mathrm{dB}, \tag{5}$$

$$MAE = \frac{1}{M_1 M_2}\sum_{i=1}^{M}\sum_{j=1}^{N}\left\|x(i,j) - x_0(i,j)\right\|_{L_1}, \tag{6}$$

where $MSE = \frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}\left\|x(i,j) - x_0(i,j)\right\|_{L_2}^2$ is the mean square error, M, N are the image dimensions, $x(i,j)$ is the 3D vector value of the pixel (i,j) in the filtered colour image, $x_0(i,j)$ is the corresponding 3D vector value of the pixel in the original uncorrupted image, and $\|\cdot\|_{L_1}$, $\|\cdot\|_{L_2}$ are the L_1- and L_2-vector norms, respectively;

$$NCD = \frac{\sum_{i=1}^{M}\sum_{j=1}^{N}\left\|\Delta E_{Luv}(i,j)\right\|_{L_2}}{\sum_{i=1}^{M}\sum_{j=1}^{N}\left\|E^*_{Luv}(i,j)\right\|_{L_2}}. \tag{7}$$

Here, $\left\|\Delta E_{Luv}(i,j)\right\|_{L_2} = \left[\left(\Delta L^*(i,j)\right)^2 + \left(\Delta u^*\right)^2 + \left(\Delta v^*\right)^2\right]^{1/2}$ is the norm of colour (or multichannel) error; ΔL^*, Δu^*, and Δv^* are the difference in the L^*, u^*, and v^* components, respectively, between the two colour vectors that present the filtered image and uncorrupted original one for each a pixel *(i,j)* of an image; and $\left\|E^*_{Luv}(i,j)\right\|_{L_2} = \left[\left(L^*\right)^2 + \left(u^*\right)^2 + \left(v^*\right)^2\right]^{1/2}$ is the L_2 norm or magnitude of the uncorrupted original image pixel vector in the $L^*u^*v^*$ space. It has been proved that the NCD objective measure expresses well the colour distortion (Plataniotis & Venetsanopoulos, 2000).

3. Some Efficient Frameworks in Video Sequences Processing

Let present several promising approaches that are used in video sequence filtering.

3.1 Motion-Compensated 3-D LLMMSE Filter

In this approach (Yin et. al., 2007), an image-noise model is supposed to be a sum of an image and the signal-independent, additive, spatio–temporal invariant white noise. Uniform temporal filtering area is adaptively grown according to the motion estimation status of the adjacent candidate frames. So, the frames with higher temporal correlation are motion-compensated to the current one. The pixel aggregation algorithm is used to include the homogeneous adjacent pixels and exclude the outlier (noisy) pixels. An adaptive weighted local mean and variance improve the filtering performance. When a pixel within the filtering support deviates from the current pixel beyond a defined threshold in terms of intensity, its weight is decreased to deemphasize its contribution to the local mean and variance estimation.

The spatio–temporal LLMMSE estimate of the pixel at the spatial position *i, j* of the *k*-th frame is given by adaptive Wiener filtering algorithm

$$\hat{x}(i,j,k) = \overline{x}(i,j,k) + \frac{\hat{\sigma}^2_{x_0}}{\hat{\sigma}^2_{x_0} + \hat{\sigma}^2_n}\left[x(i,j,k) - \overline{x}(i,j,k)\right], \tag{8}$$

where $\overline{x}(i,j,k)$ is the mean estimate of the current pixel in the local spatio–temporal area, which is a cuboid window centered about the current pixel. In the same area, the variance estimate of $x(i,j,k)$ can be computed, and also local estimate of dispersion can be found $\hat{\sigma}^2_{x_0} = \max\left[\hat{\sigma}^2_x - \hat{\sigma}^2_n\right]$. The robust block-matching motion estimator is employed here, where all candidate motion vectors are checked to select the right motion vector within an adaptively uniform area with enough spatial gradients. With the motion field obtained, the adjacent frames are compensated with respect to the current frame selecting the data used in filtering stage. The 8 per 8 blocks are used for motion estimation, finally presenting the results in the form of the dark blocks that mark the temporal stationary data in the current data, which form the temporal filtering samples; on the other hand, the white blocks represent the regions containing temporal non stationarity on the data. In general, the more adjacent frames a filter are used, the higher denoising capability it can achieve. However, the more temporal blurring can be due to increasing imperfection of motion compensation. In general, the candidate frames having higher temporal correlations with respect to the current frame are selected to grow the temporal data to be filtered.

3.2 Inter-frame Model of Wavelet Coefficients

In this approach (Yin et. al., 2007), an image-noise model is supposed to be a sum of image (Mahbubur Rahman et. al., 2007). In order to take into account the correlation between the wavelet (WL) coefficients of any two neighbouring frames, a joint statistical model in form bivariate Gaussian distribution for the video wavelet coefficients can be used. The joint density function takes into account one of the essential variabilities of the video WL coefficients of the neighbouring frames (the motion). So, the video WL coefficients are zero-mean conditionally independent bivariate Gaussian random variables with slow-varying variance and covariance. This model is a base for developing a bivariate maximum a posteriori (MAP) estimator for spatial filtering of a noisy video.

Let define $f_j(k)$ as the WL coefficients for a given sub-band of the j-th frame, where, for simplicity k is used to represent two-dimensional spatial indexes. The WL coefficients of the previous neighbouring frames are denoted as $f_i(k)$. Because the correlation coefficient represents the linear relationship between the two random processes, so, for a given sub-band video WL coefficients, the amount of motion that exists between any two frames can be indirectly measured by correlation coefficient. So, it is preferable to use the sub-band dependent correlation parameter r as an index of the motion. The higher the value of r is, the lower the amount of motion between the sub-bands of the two neighbouring frames will be, and vice versa.

To define the joint density function WL coefficients for the current frame and any of the previous frames, the motion index is used. The bivariate Gaussian density function with a strong dependency between two random processes is elliptic, so, the coefficients of any two frames with very little motion can be modelled using this density function. For a relatively large motion, this coefficient can be assumed to be zero.

The joint PDF (Probability Density Function) of WL coefficients for the current frame and any of the previous frames is written as:

$$w(x_i, x_{i-1}) = \begin{cases} \dfrac{1}{2\pi\sigma_1\sigma_2\sqrt{1-r^2}} \exp\left[-\dfrac{1}{2\sqrt{1-r^2}} \left\{ \dfrac{x_i^2}{\sigma_1^2} + \dfrac{x_{i-1}^2}{\sigma_2^2} + 2r\dfrac{x_i}{\sigma_1}\dfrac{x_{i-1}}{\sigma_2} \right\} \right] & ,if \quad 0 < r < 1 \\ \dfrac{1}{2\pi\sigma_1\sigma_2} \exp\left[-\dfrac{1}{2} \left\{ \dfrac{x_i^2}{\sigma_i^2} + \dfrac{x_{i-1}^2}{\sigma_{i-1}^2} \right\} \right] & ,if \quad -1 < r < 0 \end{cases} . \tag{9}$$

If the pixels of the video frames are corrupted by additive white Gaussian noise and the variance is unknown, it may be estimated by applying the median absolute deviation method in the highest sub-band of noisy WL coefficient. Noisy WL is presented as a sum of coefficients of a frame and noise.

First, let develop a bivariate MAP estimator to estimate the image WL coefficients of the current frame denoted as $x_i(k)$, applying the correlation information of the j-th previous frame into account. The variances and covariances are estimated from the bivariate maximum likelihood (ML) estimator. In the second step, the estimated coefficients $x_i(k)$ are passed through a recursive temporal averaging filter for additional noise reduction. At the last step, the denoised coefficients, denoted as $\hat{x}_i(k)$, are inverse transformed to obtain the denoised video frame. The bivariate MAP estimator for WL coefficients is defined in the current frame from the noisy versions of the current frame and the i-th previous frame and can be written using eq. (8) and Gaussian PDF for noisy observation:

$$\hat{x}_j(k)\Big|x_{\Sigma,i}(k) = \begin{cases} \dfrac{\bar{\sigma}_j^2(k)\bar{\sigma}^2(k)}{\bar{\sigma}_j^2(k)\bar{\sigma}^2(k) + \sigma_n^2\left[\bar{\sigma}_i^2(k) + \sigma_n^2\right]} \left[x_{\Sigma,i}(k) + \bar{r}(k)\dfrac{\bar{\sigma}_i(k)}{\bar{\sigma}_j(k)}\dfrac{\sigma_n^2}{\bar{\sigma}^2(k)} x_{\Sigma,i}(k) \right], if \ o < r < 1 \\ \dfrac{\bar{\sigma}_j^2(k)}{\bar{\sigma}_j^2(k) + \sigma_n^2} x_{\Sigma,j}(k), if \ -1 < r < 0 \end{cases} , \tag{10}$$

where $\bar{\sigma}^2(k) = \left[1 - \bar{r}^2(k)\right]\bar{\sigma}_i^2(k) + \sigma_n^2, \quad \text{for} \quad i = j-1, j-2, .., j-l$.

Therefore, the statistical model for the near frame video WL coefficients is considered as locally independent and identically distributed (i.i.d.) bivariate Gaussian distribution with conditional mean, variance, and covariance that are calculated locally for each index *k*.

3.3 Wavelet-Domain and Motion-Compensated Video Denoising

This video denoising approach (Jovanov et al., 2009) exploits the idea of the motion estimation resources from the video coding module for video denoising. A novel motion field filtering step refines the accuracy of the motion estimates to a degree that is required for video denoising. Additionally, a novel robust temporal filtering against errors in the estimated motion field is proposed. Here, it is assumed that the video sequences are contaminated with the additive white Gaussian noise, with zero mean and known variance. The denoising approach is based on spatio-temporal filtering that combines WL spatial filtering, which is preceded by pixel-domain temporal filtering.

The basic idea in temporal algorithm is to compare the MAD between the corresponding blocks with the average MAD, and decide if motion is present or not. The proposed motion filtering method is particularly effective in suppressing spurious background motion vectors. The threshold *THR* for motion detection in the *k*-th frame in this filtering step is used to decide whether motion exists in each block. If the MAD<*THR*, both motion vector components are set to zero. Otherwise, the motion vector keeps its original value.

The idea of Motion Compensated Temporal Filter is to control switching between weaker and stronger temporal smoothing based on a motion detection variable. At positions where no motion was detected, a standard recursive temporal filter is applied. At moving positions the filtering is realized, but this time along the estimated motion trajectory, using different filter coefficients. This covers the situation when the estimated motion is not perfect permitting a different degree of temporal smoothing for moving and for non-moving areas. The proposed motion compensated filter is written as

$$\hat{d}_{i,j}^{k} = (1-m_{i.j}^{k})(\alpha(1+\varepsilon_{i,j}^{k})d_{i,j}^{k} + (1-\alpha)(1-\varepsilon_{i,j}^{k})d_{i-p,j-q}^{k-1}) + +m_{i,j}^{k}(\beta(1+\varepsilon_{i,j}^{k})d_{i,j}^{k} + \\ +(1-\beta)(1-\varepsilon_{i,j}^{k})d_{i-p,j-q}^{k-1}) \quad , \quad (11)$$

where α and β are the fixed parameters in recursive filter in static and moving areas. The values $1+\varepsilon_{i,j}^{k}$ and $1-\varepsilon_{i,j}^{k}$ are data driven factors for these parameters. The factor $1+\varepsilon_{i,j}^{k}$ increases the influence of current frame pixel value $d_{i,j}^{k}$ in the case when prediction error ε is large. The influence of $d_{i-p,j-q}^{k-1}$ on the filtering result is in this case simultaneously suppressed through $1-\varepsilon_{i,j}^{k}$ (which is close to zero). Otherwise, when the prediction error ε is small, the factor $1-\varepsilon_{i,j}^{k}$ is close to 1, enforcing smoothing along the estimated motion trajectory.

At the second stage of framework, the temporal filter is combined with a wavelet domain spatial filter using a fuzzy-logic version of the spatially adaptive Probability Shrink that is applied to each wavelet coefficient a shrinkage factor, which is a function of two measurements: the coefficient magnitude and a local spatial activity indicator that indicates the fine feature changes.

3.4 Video denoising by fuzzy motion and detail adaptive averaging-FMDAF

A fuzzy-rule-based algorithm for the denoising of video sequences (Melange et al., 2008) corrupted with additive Gaussian noise constitutes a fuzzy-logic-based improvement of a recent detail and motion adaptive multiple class averaging filter (MCA) (Zlokolica et al., 2003). Last framework to avoid the spatio-temporal blur, only takes into account neighbouring pixels from the current frame in case of detected motion. So, to preserve the details, the filtering should be less strong when large spatial activity is detected in a current window. The filtering window used in the framework is a $3\times3\times2$ sliding window, consisting of pixel windows in the current and previous frames. The output of the proposed filter for the central pixel in the window is determined as a weighted adaptive average of the pixel values in the $3\times3\times2$ window:

$$\hat{x}_t(r)=\frac{\sum_{r'}\sum_{t'=t-1}Q(r',t',r,t)x_{\Sigma t'}(r')}{\sum_{r'}\sum_{t'=t-1}Q(r',t',r,t)}\,. \tag{12}$$

The absolute greyscale difference (gradient) between the two spatial-temporal pixel positions is computed as $\Delta(r',t',r,t)=\left|x_{\Sigma,t'}(r')-x_{\Sigma,t}(r)\right|$; the function indicating the local detail amount is presented by $\delta(r,t)=\left\{\sum_{r'}\left[x_{\Sigma,t}(r')-\overline{x}_{\Sigma,t}(r)\right]^2\right\}^{1/2}$; and the motion indicator $m(r,t)=\left|\overline{x}_{\Sigma,t}-\overline{x}_{\Sigma\ ,t-1}\right|$ is measured as the absolute difference between the average grey values in the windows for the current and previous frames. In the MCA filter, the pixels are classified into four discrete index classes, depending on the $\Delta(r',t',r,t)$ value. When details are detected in a region, higher weights area assigned to pixels that are similar to the pixel being filtered (pixels from the lower index classes, which have smallest $\Delta(r',t',r,t)$ values), preserving these details. In homogeneous regions, the difference in weight compared to pixels from the higher index classes will be smaller, and strong filtering is to be performed. Exponential model for averaging function $Q(r',t',r,t)$, which depends on the amount of detail, motion, and class index inversely proportional, has been used.

Fuzzy motion and detail adaptive video filter, *FMDAF* employs the idea of MCA framework and the values $\Delta(r',t',r,t)$, $\delta(r,t)$ and $m(r,t)$. The model of exponential functions is changed by fuzzy logic framework with linguistic variables, introducing one fuzzy set *Large Difference* for the values $\Delta(r',t',r,t)$. If a difference $\Delta(r',t',r,t)$ has a membership degree one in the fuzzy set *Large Difference*, then this means that this difference is large for sure. A membership degree equal to zero exposes the certainty that the difference is not large.

A linguistic variable "*Large*" that has been proposed for the difference $\Delta(r',t',r,t)$, is also introduced for the motion value $m(r,t)$ and for the detail value $\delta(r,t)$ defining the fuzzy sets *Large Motion* and *Large Detail*. A linguistic variable "*Reliable*" to indicate whether a given neighbourhood pixel is reliable to be used in the filtering of the central window pixel, and is represented by the fuzzy set "*Reliable Neighbourhood Pixel*". Finally, the weight $Q(r',t',r,t)$ for the pixel at position (r',t') is now defined as the degree, to which it is reliable to be used

in the filtering of the central window pixel, i.e., its membership degree in the fuzzy set "*Reliable Neighbourhood Pixel*". The presented fuzzy rule 1 or 2 are depended on whether current $t' = t$ or previous $t' = t - 1$ frames positions.

Fuzzy rule 1. Assigning the membership degree in the fuzzy set "*Reliable Neighbourhood Pixel*" of the pixel at spatial position r' in the current frame $(t' = t)$ of the window with central pixel position *(r, t):*

IF [the detail value $\delta(r,t)$ is large AND the difference $\Delta(r',t',r,t)$ is not large] ***OR*** [the detail value $\delta(r,t)$ is not large)] ***THEN*** the pixel at position (r',t') is a *Reliable Neighbourhood Pixel* for the filtering of the central window pixel.

Fuzzy rule 2. Assigning the membership degree in the fuzzy set "*Reliable Neighbourhood Pixel*" of the pixel at spatial position r' in the previous frame $(t' = t - 1)$ of the window with central pixel position *(r, t):* ***IF*** { [the detail value $\delta(r,t)$ is large ***AND*** the difference $\Delta(r',t',r,t)$ is not large] ***OR*** [the detail value $\delta(r,t)$ is not large] } ***AND*** the motion value $m(r,t)$ is not large ***THEN*** the pixel at position (r',t') is a *Reliable Neighbourhood Pixel* for the filtering of the central window pixel.

Finally the described framework FMDAF adapts better to motion than the RMCA method as results reported in paper (Melange et al., 2008) indicates.

4. Fuzzy-Angular Deviation Frameworks in Denoising of Video Sequences

4.1. Additive Noise Suppression

4.1.1. 2D Spatial Noise Filtering

The filtering procedure includes the Histogram Calculation, Noise Estimation, and Spatial Algorithm Operations. A mean value $\overline{x}_\beta$ ($\beta = (Red, Green, Blue)$ in a colour image) is found in a sliding 3x3 processing window; later, the angle between two vectors deviation is computed agree to (Ponomaryov et al., 2007), mean value ($X = \{\overline{x}_R, \overline{x}_G, \overline{x}_B\}$), and central pixel $Y = \{x_{cR}, x_{cG}, x_{cB}\}$) is calculated. Finally, the probabilities: p_j, the mean value μ, the variance σ_β^2, and standard deviation (*SD*) $\sigma_\beta^{'} = \sqrt{\sigma_\beta^2}$ should be calculated. Two processing windows: large 5x5, and into it, small 3x3 one, are employed in this scheme.

Let denote as $\theta_i = A(x_i, x_c)$ the angle deviation x_i in respect to x_c, where $i = 0,1,...,8, i \neq c, c = central\ pixel$. The Spatial Algorithm is employed realizing the following ***IF-THEN*** rule for filtering the first frame only: $IF(\theta_1\ AND\ \theta_3\ AND\ \theta_4\ AND\ \theta_6 \geq \tau_1)\ OR\ (\theta_0\ AND\ \theta_2\ AND\ \theta_5\ AND\ \theta_7 \geq \tau_1)$ *THEN Mean Weighted Filtering ELSE Spatial Filtering Algorithm*. The "*AND*" operation is defined as "Logical AND", the "*OR*" operation is "Logical OR". The *Mean Weighted Filtering Algorithm* is realized using angle deviations as weight criteria (Ponomaryov et al., 2007).

If the spatial algorithm is selected, the processing is realized in each a colour plane using locally adapted *SD* σ_β around of mean value $\overline{x}_{\beta 5\times 5}$ found in sliding 5x5 processing window, adjusting it as follows: If $\sigma_\beta < \sigma'_\beta$ then $\sigma_\beta = \sigma'_\beta$ otherwise $\sigma'_\beta = \sigma_\beta$.

Let introduce for a central pixel $x_c = x(i, j)$ of a current sample the following neighbours in eight cardinal directions: *N*, *E*, *S*, *W*, *NW*, *NE*, *SE*, *SW* (Schulte et al., 2007b), and also similarity measures for each a given plane ($\beta = (Red, Green, Blue)$):

$$\nabla_{(k,l)\beta}(i,j) = \left|\nabla_{\beta}(i+k, j+l) - \nabla_{\beta}(i,j)\right|, \; k,l \in \{-1,0,1\}. \tag{13}$$

These gradients are called "*main gradient values*", and the point *(i, j)* is "*the centre of the gradient values*". Two "*derived gradient values*" are proposed, permitting to avoid blur in presence of the edges (Schulte et al., 2007b). Finally, these three gradient values are connected into one value called "*fuzzy vectorial-gradient value*" under IF-THEN rule: IF $\nabla_{\gamma\beta} < T_{\beta}, T_{\beta} = 2 \cdot \sigma_{\beta}$, THEN it is calculated the angle deviation in each γ direction from eight mentioned for main and derived vectorial values involved.
Let define the membership function to obtain "*Fuzzy Main and Derived Vectorial-Gradient Values*":

$$\mu_{BIG} = \begin{cases} \max\{x, y\}, \; if \; \nabla_{\gamma\beta} < T_{\beta} \\ 0, \; otherwise \end{cases}, \text{ where } x = \alpha_{\gamma(M,D1,D2)\beta}, \; y = 1 - \left[\nabla_{\gamma(M,D1,D2)\beta} \Big/ T_{\beta} \right],$$
$$\alpha_{\gamma\beta} = 2/[1 + \exp(\theta_{\gamma\beta})], \; M = Main\ value, \; D1 = Derived1, \; D2 = Derived2, \tag{14}$$

and $\theta_{\gamma\beta}$ is the angle deviation between vector pixels [255, 255, $x_{\gamma\beta}$] and [255, 255, $x'_{\gamma\beta}$] for each a colour channel. Finally, the process to obtain "*Fuzzy Vectorial-Gradient Values*" is defined as the **Fuzzy Rule 1_2D_G**:
Fuzzy Rule 1_2D_G: *Fuzzy Vectorial-Gradient value* is defined as $\nabla_{\gamma\beta}\alpha_{\gamma\beta}$, in such a way
IF ($\nabla_{\gamma\beta M}$ is BIG AND $\nabla_{\gamma\beta D1}$ is BIG) OR ($\nabla_{\gamma\beta M}$ is BIG AND $\nabla_{\gamma\beta D2}$ is BIG) THEN $\nabla_{\gamma\beta}\alpha_{\gamma\beta}$ is true.
Final step in filtering a noise is realized employing a *Weighted Mean* procedure with found weights:

$$y_{\beta out} = \sum_{\gamma} \omega_{\gamma} x_{\gamma\beta} \Big/ \sum_{\gamma} \omega_{\gamma} \quad y_{\beta out} = \sum_{\gamma} \omega_{\gamma} x_{\gamma\beta} \Big/ \sum_{\gamma} \omega_{\gamma}, \; \omega_{\gamma} = \nabla_{\gamma\beta}\alpha_{\gamma\beta}. \tag{15}$$

4.1.2. 3D Spatio-Temporal Noise Filtering

The "*Temporal Algorithm*" is designed realizing the motion detection in past and present frames of a video sequence for better preservation of the image characteristics.
The angle deviations and gradient values related to a central pixel in the present frame respect to its neighbours from past frame are found according to the first expression in the following equation:

$$\left(\theta_i^1 = D(x_i^{t-1}, x_c^t), \nabla_i^1 = \left|x_i^{t-1} - x_c^t\right|\right), \ \left(\theta_i^2 = D(x_i^{t-1}, x_i^t), \nabla_i^2 = \left|x_i^{t-1} - x_i^t\right|\right),$$
$$\left(\theta_i^3 = D(x_i^t, x_c^t), \nabla_i^3 = \left|x_i^t - x_c^t\right|\right) \tag{16}$$

where $i = 1, 2, ..., 8$, $x_{\beta c}^t$ is a central pixel channel in the present frame, and *t-1* and *t* mark the past and present frames, respectively. The angle and gradient values in both frames are calculated according to second equation in (16). Finally, the same parameters for the present frame are only employed, eliminating operations in past frame as in the third expression in eq. (16).

The Gaussian membership functions in the fuzzy sets SMALL and BIG for gradients and angular deviations are defined as:

$$\mu_{SMALL}(\theta) = \begin{cases} 1, if\ \theta < \theta_1 \\ \exp[-(\theta - \theta_1)^2 / 2\sigma^2], otherwise \end{cases}, \ \mu_{SMALL}(\nabla) = \begin{cases} 1, if\ \nabla < \nabla_1 \\ \exp[-(\nabla - \nabla_1)^2 / 2\sigma^2], otherwise \end{cases}, \tag{17a}$$

$$\mu_{BIG}(\theta) = \begin{cases} 1, if\ \theta < \theta_2 \\ \exp[-(\theta - \theta_2)^2 / 2\sigma^2], otherwise \end{cases}, \ \mu_{BIG}(\nabla) = \begin{cases} 1, if\ \nabla < \nabla_2 \\ \exp[-(\nabla - \nabla_2)^2 / 2\sigma^2], otherwise \end{cases}, \tag{17b}$$

where $\theta_1 = 0.2, \theta_2 = 0.9, \nabla_1 = 60, \nabla_2 = 140$, and $\sigma^2 = 0.1$ for θ and $\sigma^2 = 1000$ for ∇, and the numerical values of parameters are chosen according to the optimum values of the *PSNR* and *MAE* criteria.

The designed fuzzy rules (see Fig.1) are used to detect the *movement presence* and/or *noise* analyzing pixel by pixel, and to form a sample of pixels with similar structures for the subsequent filtration. The fuzzy rules were designed to detect changes in magnitude and angle deviations between central and neighbouring pixels in *t* and *t-1* frames. Procedure for fuzzy rules is as follows:

Fuzzy Rule 2_3D_G: Definition of the Fuzzy Vectorial-Gradient value $SBB_{\beta i}$: IF θ^1 is SMALL AND θ^2 is BIG AND θ^3 is BIG AND ∇^1 is SMALL AND ∇^2 is BIG AND ∇^3 is BIG THEN *SBB* is true (Fig. 1 b)).

Fuzzy Rule 3_3D_G: Definition of the fuzzy Vectorial-Gradient value $SSS_{\beta i}$: IF θ^1 is SMALL AND θ^2 is SMALL AND θ^3 is SMALL AND ∇^1 is SMALL AND ∇^2 is SMALL AND ∇^3 is SMALL THEN *SSS* is true (Fig. 1 c)).

Fuzzy Rule 4_3D_G: Definition of the fuzzy Vectorial-Gradient value $BBB_{\beta i}$: IF θ^1 is BIG AND θ^2 is BIG AND θ^3 is BIG AND ∇^1 is BIG AND ∇^2 is BIG AND ∇^3 is BIG THEN *BBB* is true (Fig. 1 d)).

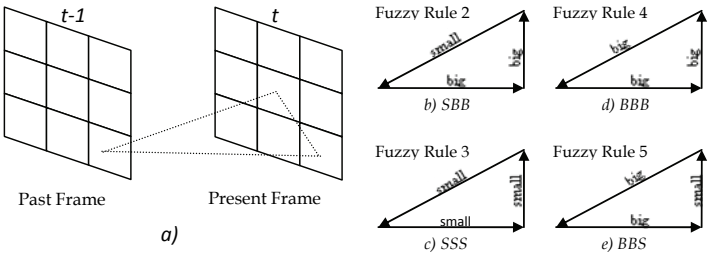


Fig. 1. Fuzzy Rules 2-5 in determination of the motion confidence in neighbouring frames. *a)* past and present Frames; *b)* Fuzzy Rule 2, *SBB*; *c)* Fuzzy Rule 3, *SSS*; *d)* Fuzzy Rule 4, *BBB*; *e)* Fuzzy Rule 5, *BBS*.

Fuzzy Rule 5_3D_G: Definition of the fuzzy Vectorial-Gradient value $BBS_{\beta i}$ **:** IF θ^1 is BIG AND θ^2 is BIG AND θ^3 is SMALL AND ∇^1 is BIG AND ∇^2 is BIG AND ∇^3 is SMALL THEN *BBS* is true (Fig. 1 e)).

For the reconstruction of edges and fine details in the image, we use the following processing procedure: a) calculate the *SD* (σ'') in the double 3×3×2 window of the neighbouring images, and b) compare the current *SD* with previous using the following rule: IF $\{(\sigma''_{RED} \geq 0.4 * \sigma'_{RED})\text{AND}(\sigma''_{GREEN} \geq 0.4 * \sigma'_{GREEN})\text{AND}(\sigma''_{BLUE} \geq 0.4 * \sigma'_{BLUE})\}$, THEN fuzzy rules 2, 3, 4, and 5; OTHERWISE, weighted mean filter. The latter filter is applied using 17 pixels from the 3×3×2 window. Using this procedure, it is possible to select the areas containing fine details and contours and subsequently filter the pixels from this area according to the fuzzy logic algorithms. The *SD* values are updated using the following sensitivity parameter α : $\sigma'_\beta = \alpha(\sigma_{total}/5) + (1-\alpha)\sigma'_\beta$, $\sigma_{total} = (\sigma''_{RED} + \sigma''_{GREEN} + \sigma''_{BLUE})/3$. This parameter is chosen as follows: α = 0.125 for the weighted mean filter and the fuzzy rule *SSS*, α = 0.875 for *SSB* and *BBS*, and α = 0.875 in the case of *BBB* if the motion–noise confidence value is (*motion_noise)=1*; α = 0.125 if *(motion_noise)= 0*, and α = 0.5 in other cases or if the fuzzy rule is not applied.

If number of pixels with fuzzy value *SBB*, or *SSS*, or *BBS*, or *BBB* is the biggest one against those that present other ***IF-THEN*** conditions, it should be employed the next filtering algorithm only for such the pixels that satisfy the established ***IF*** condition :

$$y_{\beta out} = \sum_{i=1}^{\#pixels} x_{\beta i}^{t-1} \cdot SBB_{\beta i} \Big/ \sum_{i=1}^{\#pixels} SBB_{\beta i} \text{ , or } y_{\beta out} = \sum_{i=1}^{\#pixels} 0.5(x_{\beta i}^{t-1} + x_{\beta i}^{t}) \cdot SSS_{\beta i} \Big/ \sum_{i=1}^{\#pixels} SBB_{\beta i} \text{ , or}$$

$$y_{\beta out} = \sum_{i=1}^{\#pixels} x_{\beta i}^{t} \cdot (1 - BBS_{\beta i}) \Big/ \sum_{i=1}^{\#pixels} (1 - BBS_{\beta i}) \text{ ,} \quad (18)$$

or, if the number of pixels with $BBB_{\beta i}$ value is the biggest one. Here the filtering results $y_{\beta out} = (1-\alpha)x_{\beta c}^{t} + \alpha x_{\beta c}^{t-1}$. In eq. (18), $\#pixels$ are the number of pixels that satisfy to mentioned *IF-THEN* condition; $x_{\beta i}^{t-1}$, $x_{\beta i}^{t}$ represent each a pixel in the past and present

frames that satisfy to mentioned *IF-THEN* condition; $y_{\beta out}$ is the output in spatial temporal filtering. If there is no majority in pixels for any Fuzzy Rule, only the mean of central pixels from present and past frames are used.

During numerous simulations, different video colour sequences *Miss America* (MA), *Flowers* (F) and *Chair (C)* in RGB colour space (24 bits) and QCIF format (176x144 pixels in a frame) are used to qualify effectiveness of the proposed approach in suppression of a noise and compare it with known techniques. Mentioned video sequences present different texture characteristics; permitting a better understanding of the robustness of the proposed and existed filtering schemes. Video sequences were contaminated by Gaussian noise of different intensity from 0.0 to 0.05 in their *SDs*. The filtered frames were evaluated according to *PSNR*, *MAE*, *NCD* objective criteria, and also in subjective matter.

The proposed Fuzzy Directional Adaptive Recursive Temporal Filtering for Gaussian noise named as ***FDARTF_G*** was compared with another similar one, the *FMRSTF*, and with the *FVMRSTF* (Fuzzy Vectorial Motion Recursive Spatial-Temporal Filtering Using Angles) that is the modification of *FMRSTF*, which combines the gradients and angles in processing. Other two reference filters were: *VGVDF_G*, adapted to process three frames, and the *VVDKNNVMF* filter presenting good efficiency in comparison with other filtering procedures. The data presented in Table 1 show that the proposed algorithm effectively suppresses the low-intensity additive noise and is the best according to the majority of filtration criteria for the video sequences *Flowers* and *Miss America*. Fig. 2 presents filtering results for sequence *Miss America* through 100 frames, where the better noise suppression in form of *PSNR* measure can be observed for novel filtering scheme.

Criteria	Flowers Frame 20, Gaussian noise SD = 0.005					Miss America Frame 20, Gaussian noise SD = 0.005				
	FMRSTF	*FVMRSTF*	***FDARTF_G***	*VVDKNNVMF*	*VGVDF*	*FMRSTF*	*FVMRSTF*	***FDARTF_G***	*VVDKNNVMF*	*VGVDF*
PSNR	26,192	26,01	**27,31**	25,36	25,46	29,93	29,91	**32,51**	29,80	30,66
MAE	9,638	9,83	**8,50**	8,78	8,96	5,82	5,83	**4,46**	6,18	5,55
NCD	0,016	0,017	**0,015**	0,015	0,017	0,02	0,02	**0,016**	0,021	0,02
	SD = 0.01					SD = 0.01				
PSNR	24,36	24,34	**25,72**	24,63	24,72	27,686	27,68	**30,06**	27.61	28.66
MAE	11,93	11,97	**10,44**	9,92	10,15	7,48	7,5	**6,07**	8.14	7.21
NCD	0,0206	0,0208	0,0187	**0,0169**	0,0193	0,026	0,026	**0,021**	0,028	0,026

Table 1. Simulation results for proposed framework and comparative filters.

4.2. Impulsive Noise Suppression

4.2.1. 2D Noise Filtering

Similar as in additive noise suppression idea is realized in framework used in impulsive noise suppression. It is based on the fuzzy-set theory and directional characteristics of an image, angular deviations of the image pixels in neighbouring multichannel video frames when the final filtered image frames are formed. At the first stage the spatial filtration of the initial frame of a sequence is performed. The following time stage realizes the combined processing of current neighbouring frames of the sequence. This processing uses the fuzzy set theory, which makes it possible to improve noise suppression. At the final stage, the

spatial filtration mechanism in each current frame is employed again. Let consider gradients and angular deviations of pixels in order to estimate the similarity between pixels within sliding processing window in verification if the central pixel is distorted by noise or free of noise. For each of directions $\gamma = \{N, E, S, W, NW, NE, SE, SW\}$ with respect to the central pixel x_c^β, we introduce the gradient $\nabla^\beta_{(k,l)} x(i,j) = \left| x_C^\beta(i,j) - x^\beta(i+k, j+l) \right|$ where $(i,j) = (0,0)$ within the processing window, with the index β determining the image components (red *(R)*, green *(G)*, and blue *(B)*), $(k,l) \in \{-1,0,1\}$. We also introduce the basic gradient and four related gradients calculated with respect to the former one, the index values being $(k,l) \in \{-2,-1,0,1,2\}$ for each direction γ (see Fig. 3). Fig. 3 shows pixels in processing procedure for *SE* direction for the basic and four related components.

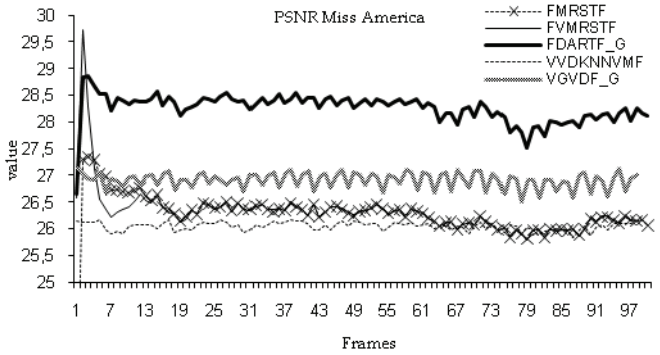


Fig. 2. *PSNR* criterion values for proposed and reference filters for 100 frames of colour video sequence *Miss America* contaminated by Gaussian noise with *SD*=0.015.

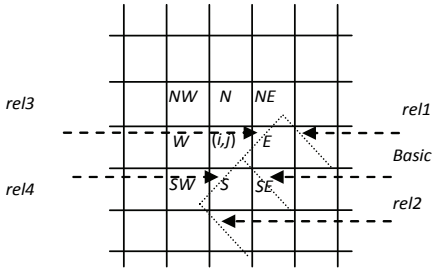


Fig. 3. Basic and related directions for gradients and angle variance values.

Let determine the angular deviation θ_γ^β between the multichannel vectors $x_1 = (x_1^R, x_1^G, x_1^B)$ and $x_2 = (x_2^R, x_2^G, x_2^B)$ for each a colour component along the direction γ (for the SE direction, in this case) agree to procedure given in (Ponomaryov et al., 2007). Also, let define the basic gradients $\nabla^\beta_{(1,1)} x(i,j) = \nabla^\beta_{SE(b)}$, $\theta^\beta_{\gamma=SE(b)}$, the related gradients, and the angular deviations:

$$F^{\beta}_{(0,2)}x(i-1,j+1)=F^{\beta}_{SE(r_1)},\ F^{\beta}_{(2,0)}x(i+1,j-1)=F^{\beta}_{SE(r_2)},\ F^{\beta}_{(0,0)}x(i-1,j+1)=F^{\beta}_{SE(r_3)},$$
$$F^{\beta}_{(0,0)}x(i+1,j-1)=F^{\beta}_{SE(r_4)}, \tag{19}$$

where the operator F determines gradient ∇ or angular deviation θ. Analogously, we find the gradients and the angular deviations for the basic value and four related values of other directions γ.

We now introduce two fuzzy sets SMALL (S) and BIG (B). Then, we use the Gaussian membership functions for both gradients and angular deviations in these sets:

$$\mu(F^{\beta}_{\gamma}\ \text{SMALL})=\begin{cases}1, & F^{\beta}_{\gamma}<med_F\\ \exp\left\{-\left[(F^{\beta}_{\gamma}-med_F)^2/2\sigma_F^2\right]\right\}, & \text{other case}\end{cases}, \tag{20a}$$

$$\mu(F^{\beta}_{\gamma}\text{BIG})=\begin{cases}1, & F^{\beta}_{\gamma}>med_F\\ \exp\left\{-\left[(F^{\beta}_{\gamma}-med_F)^2/2\sigma_F^2\right]\right\}, & \text{other case}\end{cases}. \tag{20b}$$

Here, σ_1^2=1000, med_1=60 and med_2=10 for the fuzzy gradients ($F^{\beta}_{\gamma}=\nabla^{\beta}_{\gamma}$) in the BIG and SMALL fuzzy sets, σ_2^2=0.8, med_3=0.615 and med_4=0.1 for the fuzzy angular deviations $\left(F^{\beta}_{\gamma}=\theta^{\beta}_{\gamma}\right)$ in the BIG and SMALL fuzzy sets. The novel fuzzy rules developed are based on both gradients and angular deviations. They are applied to determine whether the central pixel is *a noise*, or a *no-noise pixel*, or a *local movement*.

The 2D fuzzy rule 1_2D determines the value of the fuzzy gradient-angular measure $\nabla^{\beta F}_{\gamma}\theta^{\beta F}_{\gamma}$:

IF (∇^{β}_{γ} *is* B$\otimes\ \nabla^{\beta}_{\gamma(r_1)}$ *is* S$\otimes\ \nabla^{\beta}_{\gamma(r_2)}$ *is* S$\otimes\ \nabla^{\beta}_{\gamma(r_3)}$ *is* B$\otimes\ \nabla^{\beta}_{\gamma(r_4)}$ *is* B) $\otimes_1$ (θ^{β}_{γ} *is* B$\otimes\ \theta^{\beta}_{\gamma(r_1)}$ *is* S$\otimes\ \theta^{\beta}_{\gamma(r_2)}$ *is* S$\otimes\ \theta^{\beta}_{\gamma(r_3)}$ *is* B$\otimes\ \theta^{\beta}_{\gamma(r_4)}$ *is* B), THEN $\nabla^{\beta F}_{\gamma}\theta^{\beta F}_{\gamma}$ is BIG, where $A\otimes B=A$ AND B, $A\otimes_1 B=\min(A,B)$.

Combining eight fuzzy gradient-angular measures for each of the directions, we introduce the *noise factor* r^{β}.

The 2D fuzzy rule 2_2D: IF $\nabla^{\beta F}_{N}\theta^{\beta F}_{N}$ *is* B$\oplus\ \nabla^{\beta F}_{S}\theta^{\beta F}_{S}$ *is* B$\oplus\ \nabla^{\beta F}_{E}\theta^{\beta F}_{E}$ *is* B$\oplus\ \nabla^{\beta F}_{W}\theta^{\beta F}_{W}$ *is* B$\oplus\ \nabla^{\beta F}_{SW}\theta^{\beta F}_{SW}$ *is* B$\oplus\ \nabla^{\beta F}_{NE}\theta^{\beta F}_{NE}$ is B$\oplus\ \nabla^{\beta F}_{NW}\theta^{\beta F}_{NW}$ *is* B$\oplus\ \nabla^{\beta F}_{SE}\theta^{\beta F}_{SE}$ *is* B THEN r^{β} is BIG, where $A\oplus B=\max(A,B)$.

Depending on whether a pixel is the *noisy* or is *noise-free*, we use the following filtration algorithm:

$$IF\ r^{\beta} \geq 0.3,\ fuzzy\ logic\ algorithm, otherwise\ y_{output}^{\beta} = x_C^{\beta} \tag{21}$$

Fuzzy pixel weights for the algorithm are given in the form $\rho\left(\nabla_{\dot{\gamma}}^{\beta F}\theta_{\dot{\gamma}}^{\beta F}\right) = 1 - \nabla_{\gamma}^{\beta F}\theta_{\gamma}^{\beta F}$, which determines the value of the membership function for the fuzzy set **NO BIG** (*noise free*). At the same time, the weights for the central pixel are chosen as $\xi_c^{\beta F} = M\sqrt{1-r^{\beta}}$. The spatial filtration algorithm based on the fuzzy logic includes the following operations:

1) Calculation of fuzzy weights on the basis of the ordering of pixels in the 3×3 window: $x_{\dot{\gamma}}^{\beta} = \left\{x_{SW}^{\beta}, \ldots, x_{(i,j)}^{\beta}, \ldots, x_{NE}^{\beta}\right\}$, where the ordering statistics $x_{\dot{\gamma}}^{\beta(1)} \leq x_{\dot{\gamma}}^{\beta(2)} \leq \cdots \leq x_{\dot{\gamma}}^{\beta(9)}$ are determined from the inequality $\rho\left(\nabla_{\dot{\gamma}}^{\beta F}\theta_{\dot{\gamma}}^{\beta F}\right)^{(1)} \leq \rho\left(\nabla_{\dot{\gamma}}^{\beta F}\theta_{\dot{\gamma}}^{\beta F}\right)^{(2)} \leq \cdots \leq \rho\left(\nabla_{\dot{\gamma}}^{\beta F}\theta_{\dot{\gamma}}^{\beta F}\right)^{(9)}$.
2) Determination of the quantities $sum^{\beta} += \rho\left(\nabla_{\dot{\gamma}}^{\beta F}\theta_{\dot{\gamma}}^{\beta F}\right)$ for *j*= *9, 8,…, 1*, by decreasing j from 9 until $sum^{\beta} \geq \rho_0$, $\rho_0 = \left(\sum_{\dot{\gamma}} \rho\left(\nabla_{\dot{\gamma}}^{\beta F}\theta_{\dot{\gamma}}^{\beta F}\right) + M\sqrt{1-r^{\beta}}\right)\Big/2$, $M = 3$. In this case, the pixel ordering number j satisfying this condition determines the *j*-th pixel chosen as a result of the filtration $x_{\dot{\gamma}}^{\beta(j)} = y_{output}^{\beta}$.
3) If $j \leq 2$, then the fuzzy weights are calculated with the allowance for the threshold $\rho_1 = \left(\rho_0 - \rho\left(\nabla_{\dot{\gamma}}^{\beta F}\theta_{\dot{\gamma}}^{\beta F}\right)^{(1)} - \rho\left(\nabla_{\dot{\gamma}}^{\beta F}\theta_{\dot{\gamma}}^{\beta F}\right)^{(2)}\right)\Big/2$ for the parameter $sum^{t\beta} += \rho\left(\nabla_{\dot{\gamma}}^{\beta(j)}\theta_{\dot{\gamma}}^{\beta(j)}\right)$, with j = 9, 8, …, 1 decreasing until $sum^{t\beta} \geq \rho_1$. The ordering number j satisfying this condition determines the *j*-th pixel chosen as a result of the filtration $x_{\dot{\gamma}}^{\beta(j)} = y_{output}^{\beta}$.

4.2.2. 3D Impulsive Noise Filtering

The three-dimensional (3D) algorithm (*3D-FCF*) realized at the second time stage in processing video sequences is determined by filtration of neighbouring frames. This makes it possible to estimate the degree of movement and the noise level in the central pixel as a result of the application of the 5×5×2 sliding window that contains two neighbouring frames. We calculate difference values for both the gradients and the angular deviations between the *(t – 1)*-th and the *t*-th frames.
Using the algorithm developed in the Section 4.2.1., we can derive the 3D methodology; detailed development of this algorithm is described in (Ponomaryov et al., 2009).

Using the pixels in both frames of the sequence we can compute the motion estimation and noise level present in the central pixel. In this way it is possible to filtering the noisy pixel or not filtering it because of no noise and no movement present in the central sample.
Membership values are computed in same way; defining fuzzy sets SMALL (*S*) and BIG (*B*). This means that we deal again with the *no-movement* situation or *no-noise* situation in the

pixel sample that is subjected to processing. Gaussian functions are also used for the membership function. Again, they determine the fuzzy gradient-angular difference values. The fuzzy rules developed by this filter, are based on the difference values of both gradients and angular deviations. These rules are applied with the goal to determine whether the central pixel is a *noise* or it is *no-noise*. Otherwise, we deal with *local movement*.

There are four fuzzy rules designed to determine if the central pixel is in movement, is noisy or lacks both. The first 3D fuzzy rule designed determines the value of the first fuzzy gradient-angular difference; it characterizes the confidence level for a *movement-noise event* as applied to the central pixel when the values of fuzzy gradients and angular differences along direction γ are analyzed. The second 3D fuzzy rule characterizes the confidence level with respect to the *no movement-no noise event* as applied to the central pixel along direction γ. In this case, the regions are classified as homogeneous ones, edges, and fine feature regions. The third 3D fuzzy rule allows us to estimate the existence and the level of movement or noise in the central pixel on the basis of fuzzy gradient-angular values for all directions. Finally the fourth 3D fuzzy rule determines the time stage of the video sequence filtration, in which the *j*-th pixel should be chosen as the final result. This is true if the pixel satisfies the conditions that provide the reconstruction of fine details and contours when fuzzy ordering statistics are used. In this case, the pixel nearest to the central one among all neighbouring pixels in the *t*-th and (*t*–1)-th frames of the video sequence is chosen.

The temporal stage of the filtration consists in selecting two pixels agree to 3D fuzzy rules designed. These two pixels are averaged to provide the filtering temporal result.

The characteristics of the *3D-FCF* filter proposed and algorithms well known in the literature were studied with the use of standard criteria. We compared the *PSNR* expressed in decibels, the *MAE*, and the *NCD*. The video sequences *Miss America* (MA) and *Flowers* (F) in the QCIF format (176×144 pixels) were employed. The video sequences were distorted by impulsive noise of a different intensity and processed by various filters. The distortions in each image channel were independent of each other. Table 2 shows the test results for different standard filters. The table content confirms that the *3D-FCF* algorithm developed by us is the best for estimates made by the *MAE* criterion averaged over 100 frames of the *F* sequence within a wide noise intensity range. Thereby the problem of the efficient reconstruction of the edges and fine image features is successfully solved. At the same time, the values of the *PSNR* criterion show the superiority of the new algorithm compared to the others for intermediate intensity noise. In accordance with both the *PSNR* and *MAE* criteria, the new framework is the best in the case of the MA video sequence filtration for noise of low and intermediate intensities less than 20%. Table 3 shows the values of the *NCD* criterion for MA and F video sequences, which characterizes the chromatic properties of the filters. Here, the new *3D-FCF* algorithm again demonstrates the best quality within a wide range of noise intensity. The efficient filtration of the F sequence that contains a texture varying from frame to frame, as well as noticeable variations of colours, confirms the robustness of the method proposed. Subjective perception by human viewer can be observed in Fig. 4 showing better performance of the designed 3D framework in comparison with known methods in MA frame, where novel algorithm preserves better the edges, fine features, and chromaticity properties against other filters.

Real-Time analysis was realized on the DSP (TMS320DM642, Texas Instruments) and is based on Reference Framework defined as RF5 (Mullanix & Magdic et al., 2003, Gallegos-Funes, et al., 2009). Table 4 presents the processing times in some 2D and 3D algorithms,

which have been implemented on DSP, demonstrating reliability of the proposed approach against better algorithms found in literature.

Filter (%) noise	***3D-FCF***		*3D-MF*		*3D-VVMF*		*3D-VVDKNNVMF*		*3D-VGVDF*		*3D-VAVDATM*		*3D-VKNNF*	
	MAE	***PSNR***	***MAE***	***PSNR***	***MAE***	***PSNR***	***MAE***	***PSNR***	***MAE***	***PSNR***	***MAE***	***PSNR***	***MAE***	***PSNR***
F														
5	**2,13**	**29,52**	6,65	26,83	6,64	26,78	7,45	25,77	7,44	25,56	5,45	27,25	***3,98***	**31,30**
15	**3,38**	**27,76**	7,19	26,22	7,14	26,20	8,20	25,11	7,72	25,29	***6,28***	26,57	6,79	26,63
30	**6,08**	**25,04**	8,59	24,77	8,50	24,69	10,03	23,17	9,74	23,24	***8,13***	24,99	14,62	20,86
MA														
5	**0,37**	**39,59**	2,51	35,12	2,54	34,86	3,11	33,48	2,91	33,76	***1,11***	36,97	1,91	***37,22***
15	**1,18**	34,33	2,70	34,37	2,71	34,18	3,43	32,38	2,85	33,71	***1,71***	**35,38**	3,84	30,09
30	3,58	28,26	3,35	31,95	3,31	31,82	4,39	28,48	***3,28***	31,61	**2,85**	**32,33**	10,11	23,17

Table 2. Averaged values of criteria *MAE* and *PNSR* for video sequences F and MA.

Filter (%) Noise	***3D-FCF***		*3D-MF*		*3D-VVMF*		*3D-VVDKNNVMF*		*3D-VGVDF*		*3D-VAVDATM*		*3D-VATM*	
	F	**MA**	**F**	**MA**	**F**	**MA**	**F**	**MA**	**F**	**MA**	**F**	**MA**	**F**	**MA**
5	**0,006**	**0,002**	0,015	0,009	0,015	0,009	0,017	0,011	0,016	0,011	0,012	***0,004***	0,015	0,009
15	**0,009**	**0,005**	0,016	0,010	0,016	0,010	0,018	0,012	0,017	0,010	***0,014***	***0,006***	0,016	0,010
30	**0,012**	0,016	0,018	***0,012***	0,018	***0,012***	0,020	0,015	0,020	***0,012***	***0,017***	**0,010**	0,018	***0,012***

Table 3. Averaged values of *NCD* for video sequence F and MA.

Fig. 4. a) Zoomed image region of 10th *Miss America* frame contaminated by impulsive noise of 15% intensity, b) Designed ***3D-FCF***, c) *3D-MF*; d) *3D-VVMF*, e) *3D-VGVDF*, f) *3D-VAVDATM*; g) *3D-VATM*; h) *3D-VKNNF*.

Filters	Processing time in seconds		
	Maximum	**Average**	**Total**
3D-FCF	**7.533**	**7.440**	**148.806**
3D-VVMF	0.075	0.075	1.496
3D-VGVDF	28.52	25.6	512.02
3D-VAVDATM	25.551	24.867	497.356

3D-VKNNF	0.103	0.102	2.04
2D-FCF	**1.243**	**1.241**	**24.822**
VMF_FAS	2.093	2.055	41.116
2D-GVDF	5.887	5.869	117.382
2D-CWVDF	5.806	2.909	58.18

Table 4. Time processing for 20 frames of video sequence "Miss America" on DSP.

5. Conclusions

Several promising frameworks in suppression of noise of different nature in video sequences are presented in this chapter. It has been designed novel approach that employs the 3D fuzzy-vector order statistics frameworks based on the fuzzy-set theory and the directional angular information available as a result of processing multichannel still images and neighbouring frames in the video sequences contaminated by additive or impulsive noise. The designed fuzzy rules characterize the presence of motion and noise in processing area of the pixels in two neighbouring frames. Novel approach has appeared to demonstrate the essential improvement of the processing quality compared to all known filters. The method developed was successful in the suppression of a noise, as well as in the reconstruction of edges and fine details of the images. The excellent performance of the new filtering scheme has been tested during numerous simulations in terms commonly used objective criteria *PSNR*, *MAE*, *NCD*, and *MRCE*, as well as the subjective visual perception presented in form of the visual analysis by human visual system of filtered video sequences. The approach also turned out to be extremely efficient in the reproduction of chromatic characteristics of frame in video sequences. Real-Time analysis of several promising 2D and 3D algorithms was realized on the DSP presenting available processing performance.

Acknowledgements

The authors would thank National Polytechnic Institute of Mexico and CONACYT (project 81599) for their support to realize this work.

6. References

Amer, A. & Schröder, H. (1996). A new video noise reduction algorithm using spatial subbands. *ICECS*, vol. 1, 1996, pp. 45-48.

Astola J. and Kuosmanen P. (1997). *Fundamentals of Nonlinear Digital Filtering*, CRC Press, ISBN: ,Boca Raton-N.Y.

Balster E. J., Zheng Y. F., and Ewing R. L. (2006) Combined spatial and temporal domain wavelet shrinkage algorithm for video denoising. *IEEE Trans.Circuits Syst. Video Technol.*, vol. 16, no. 2, pp. 220–230, 2006. ISSN:1051-8215.

Bovik A. (2000). *Handbook of Image and Video Processing*, Academic Press, ISBN: , San Diego, CA.

Camarena, G., Gregori, J., Morillas, V., & Sapena, S. (2008). Fast Detection and Removal of Impulsive Noise Using Peer Groups and Fuzzy Metrics. *Journal of Visual Communication and Image Representation*, Vol. 19, No. 1, 2008, pp. 20-29. ISSN:1047-3203.

Camarena J.G., Gregori V., Morillas S., Sapena A. (2010). Some improvements for image filtering using peer group techniques. *Image and Vision Computing*. V.28, No.1, 2010, pp.188–201. ISSN: 0262-8856.

Franke, K.; Köppen, M. & Nickolay, B. (2000). Fuzzy Image Processing by using Dubois and Prade Fuzzy Norms. *Proceedi.ngs Pattern Recognition*, Vol. 3, 2000, pp. 514-517. ISSN: 1051-4651.

Gallegos-Funes F., Kravchenko V., Ponomaryov V., Rosales-Silva A. (2009) Video Denoising by Fuzzy Directional Filter Using the DSP EVM DM642. Lecture Notes in Computer Science, Vol.LNCS 5856, Springer. pp.997-1004, 2009. ISSN:0302-9743.

Jovanov L., Pizurica A., Schulte S., Schelkens P., et. Al. (2009) Combined Wavelet-Domain and Motion-Compensated Video Denoising Based on Video Codec Motion Estimation Methods" *IEEE Trans. On Circuits and Syst. For Video Techn*. Vol. 19, No. 3, 2009 pp.417-421. ISSN:1051-8215.

Kravchenko V., Perez-Meana H., Ponomaryov V. (2009) *Adaptive Digital Processing of Multidimensional Signals with Applications*, FizMatLit, ISBN:978-5-9221-110-6. Moscow, 2009 (available in http://www.posgrados.esimecu.ipn.mx/).

Lukac, R., Smolka, B., N. Plataniotis, K. & N. Venetsanopoulos, A. (2004). Selection Weighted Vector Directional Filters. *Comput. Vis. and Image Underst.*, Vol. 94, 2004, pp. 140-167. ISSN: 1077-3142.

Ma, Z. H., Wu, H. R. & Feng, D. (2007). Fuzzy vector partition filtering technique for color image restoration. *Comput. Vis. and Image Underst.*, Vol. 107, No. 1-2, 2007, pp. 26-37. ISSN: 1077-3142.

Mahbubur Rahman S. M., Omair Ahmad M., and Swamy M. N. S. (2007) Video Denoising Based on Inter-frame Statistical Modeling of Wavelet Coefficients" *IEEE Trans. On Circuits and Syst. For Video Techn.*, Vol. 17, No. 2, 2007, pp.187-198. ISSN:1051-8215.

Melange, T., Nachtegael, M., Kerre E.,E., Zlokolica V., Schulte S., De Witte V, Pižurica A., Philips W.(2008). Video denoising by fuzzy motion and detail adaptive averaging, *J. of Elect. Imag.* Vol. 7, No.4, 2008, pp.043005-1_19. ISSN: 1017-9909.

Morillas, S., Gregori, V., Peris-Fajarns, G. & Latorre, P. (2005). A fast impulsive noise color image filter using fuzzy metrics, *Real-Time Imaging*, Vol. 11, 2005, pp. 417-428. ISSN: 1077-2014.

Morillas, S., Gregori, V. & Sapena, A. (2006). Fuzzy bilateral filtering for color images, *Lect Not. in Comp. Science*, Vol. 4141, 2006, pp. 138-145. ISSN:0302-9743.

Morillas, S., Gregori, V., Peris-Fajarnes, G. & Sapena, A. (2008a). Local self-adaptive fuzzy filter for impulsive noise removal in color images. *Signal Processing*, Vol. 88, No. 2, 2008, pp. 390-398. ISSN:0165-1684.

Morillas, S., Gregori, V. & Peris-Fajarnes, G. (2008b). Isolating Impulsive Noise Pixels in Color Images by Peer Group techniques, *Comp. Vis. and Image Underst.*, Vol.110, No. 1, 2008, pp. 102-116. ISSN: 1077-3142.

Morillas S., Gregori V., and Antonio Hervás A. (2009). Fuzzy Peer Groups for Reducing Mixed Gaussian-Impulse Noise From Color Images. *IEEE Trans. on Image Proc.*, V.18, No.7, 2009, pp.1452-1466. ISSN: 1057-7149.

Mullanix, T., Magdic, D., Wan, V., Lee, B., Cruickshank, B., Campbell, A., DeGraw, Y. (2003). Reference Frameworks for eXpressDSP Software: RF5, An Extensive, High Density System (SPRA795A), Texas Instruments, 2003.

Nie, Y. & E. Barner, K. (2006). Fuzzy Rank LUM Filters. *IEEE Trans. on Image Proc.*, , Vol. 15, No. 12, 2006, pp. 3636-3654. ISSN: 1057-7149.

Plataniotis, K. N., Androutsos, D., Vinayagamoorthy, S. & Venetsanopoulos, A. N. (1997). Color Image Processing Using Adaptive Multichannel Filters. *IEEE Trans. on Image Proc.*, , Vol. 6, No. 7, 1997, pp. 933-949. ISSN: 1057-7149.

Plataniotis, K. N. & Venetsanopoulos, A. N. (2000). *Color Image Processing and Applications*, Springer Verlag, Berlin, 2000. ISBN: 3-540-66953-1.

Ponomaryov, VI., Gallegos-Funes, FJ. &Rosales-Silva, A. (2005). Real-time color imaging based on RM-filters for impulsive noise reduction. *J. of Imag. Science and Techn.*, Vol. 49, No. 3, 2005, pp. 205-219. ISSN 1062-3701.

Ponomaryov, V. I. (2007). Real-time 2D-3D filtering using order statistics based algorithms. *Journal of Real-Time Image Processing*, Vol. 1, No. 3, 2007, pp.173-194.

Ponomaryov V, Rosales-Silva A., Gallegos-Funes F (2009) 3D Filtering of Colour Video Sequences Using Fuzzy Logic and Vector Order Statistics, Lect. Not. in Artif. Intell., Vol. LNAI 5807, pp.210-221. ISSN:0302-9743.

Ponomaryov V., Rosales-Silva A., Gallegos-Funes A., Perez-Meana H. (2010). Fuzzy Directional (FD) filter to remove impulse noise from color images. *IEICE Trans. On Fundam. of Electron. Commun. and Comput. Sciences*, Vol.E-93A N.2, 2010, pp.570-572. ISSN:0916-8508.

Russo, F & Ramponi, G. (1996). A Fuzzy Filter for Images Corrupted by Impulse Noise, *Signal Processing Letters*, Vol. 3, No. 6, 1996, pp. 168-170. ISSN: 1000-9000.

Saeidi, M., Saeidi, B., Saeidi, Z. & Saeidi, K. (2006). A New Fuzzy Algorithm in Image Sequences Filtering. *Circuits, Signals, and Systems*, 2006, pp. 531-105.

Schulte, S., De Witte, V., Nachtegael, M., V. der Weken, D. & E. Kerre, E. (2006). Fuzzy Two-Step Filter for Impulse Noise Reduction From Color Images, *IEEE Trans. on Image Proc.*, Vol. 15, No. 11, 2006, pp. 3567-3578, ISSN: 1057-7149.

Schulte, S., De Witte, V., Nachtegael, M., Van Der Weken D. & E. Kerre, E. (2007a). Fuzzy Random Impulse Noise Reduction Method, *Fuzzy Sets and Systems*, Vol. 158, No. 3, 2007, pp. 270-283. ISSN: 0165-0114.

Schulte, S., De Witte, V. & E. Kerre, E. (2007b). A Fuzzy Noise Reduction Method for Color Images, *IEEE Trans. on Image Proc.*, , Vol. 16, No. 5, 2007, pp. 1425-1436. ISSN: 1057-7149.

Schulte, S., Morillas, S., Gregori, V. & Kerre, E. E. (2007c). A New Fuzzy Color Correlated Impulse Noise Reduction Method, *IEEE Trans. on Image Proc.*, Vol. 16, No. 10, 2007, pp. 2565-2575. ISSN: 1057-7149.

Shaomin, P. & Lucke, L. (1994). Fuzzy filtering for mixed noise removal during image processing, *IEEE Fuzzy Systems*, Vol. 1, 1994, pp. 89-93. ISSN: 0148-5598.

Smolka, B., Lukac, R., Chydzinski, A., N. Plataniotis, K. & Wojciechowski, W. (2003). Fast Adaptive Similarity Based Impulsive Noise Reduction Filter. *Real-Time Imaging*, Vol. 9, No. 4, 2003, pp. 261-276. ISSN: 1077-2014.

Trahanias, P. E., Venetsanopoulos, A. N. (1996). Directional Processing of Color Images: Theory and Experimental Results. *IEEE Trans. on Image Proc.*, Vo. 5, No. 6, 1996, pp. 868-880. ISSN: 1057-7149.

Xu Z., Wu H.R., Qiu B., and X Yu X. (2009) Geometric Features-Based Filtering for Suppression of Impulse Noise in Color Images, *IEEE Trans. on Image Proc.*, Vol.18, No.8, 2009, pp.1742-1759. ISSN: 1057-7149.

Yin H. B., Fang X. Z., Wei Z., and Yang X. K. (2007) An Improved Motion-Compensated 3-D LLMMSE Filter With Spatio–Temporal Adaptive Filtering Support. *IEEE Trans. On Circuits and Syst. For Video Techn*., Vol. 17, No. 12, 2007, pp.1714-1727. ISSN:1051-8215.

Zlokolica V., Pizurica A., and Philips W. (2003) Video denoising using multiple class averaging with multiresolution. *Lect. Not. Comp. Sci.* Vol. 2849, 2003, pp.172–179. ISSN:0302-9743.

Zlokolica, V., De Geyter, M., Schulte, S., Pizurica, A., Philips, W. & Kerre, E. (2005). Fuzzy Logic Recursive Motion Detection for Tracking and Denoising of Video Sequences, *IS&T/SPIE Symposium on Electronic Imaging, Video Communications and Processing,* Vol.5685, 2005, pp. 771-782. ISSN 0277-786X.

Zlokolica, V., Schulte, S., Pizurica, A., Philips, W. & Kerre E. (2006). Fuzzy Logic Recursive Motion Detection and denoising of Video Sequences, *Journal of Electronic Imaging*, Vol. 15, No. 2, 2006, pp. 023008. ISSN: 1017-9909.

8

A Novel Implicit Adaptive zero pole-placement PID Controller

*Ali Zayed and **Mahmoud ELFandi
*The 7th of Aperial University, Libya. **Alfateh University, Libya

1. Introduction

For most simple processes, PID control can provide satisfactory closed loop performance. However, in spite of the considerable advantages of conventional PID controllers (such as simplicity of their structures, robustness and ease of implementation), they still have a major drawback in that the controllers may need to be re-tuned, if the systems to be controlled are subjected to significant changes, in order to achieve satisfactory performance. For this reason, during the last two decades much work in linear control theory has been devoted to incorporating the flexibility of self-tuning control and the simplicity of PID structures. A lot of self-tuning methods have been developed and special attention is currently being paid to PID self-tuning controllers and their implementation, [e.g. Yusof R. & et al., (1994); Yusof, R.; (1993) and Tokuda M.; & Yamamoto T.; (2002)].

During the past three decades, a special attention has also been given to the problem of designing pole-placement controllers and self-tuning regulators. Various self-tuning controllers based on classical pole-placement ideas were developed and employed in real applications, [e.g. Sirisena H. & Teng F.,(1986); Zhu Q., & et al., (2002); Zayed A. & Hussain A., (2004); Astrom K.., & Wittenmark B., (1973)]. The popularity of pole-placement techniques may be attributed to the fact that in the regulator case they provide mechanisms to over-come the restriction to minimum-phase plants of the original minimum variance self-tuner of Astrom K., & Wittenmark B., (1973). In the servo case, they provide the ability to directly introduce bandwidth and damping ratio as tuning parameters. In addition, there is some improvement in robustness of pole-placement methods, as they simply modify the system dynamics as opposed to cancelling them as per the early optimal self-tuning controllers. Furthermore, unlike many of the self-tuning based PID control designs [see for example Yusof R. & et al., (1994); Yusof, R.; (1993)], in which the tuning parameters must be selected using a trial and error procedure, the tuning parameters for pole-placement controllers can be automatically set *on-line* by specifying the desired closed loop poles.

Comparatively, only little attention has been given to zeros since they are considered to be less crucial than poles. Most of the previous discussion on zeros are centred around the choice of the sampling time so that the resulting system is invertible. However, it is important to note that zeros may be used to achieve better set point tracking Zayed A. & Hussain A., (2004)., and they may also help reduce the magnitude of the control action Sirisena H. & Teng F.,(1986)..

Therefore, in order to achieve more effective control action and combine the advantages of the self-tuning controllers with those of the PID, and zero pole-placement controllers need to be integrated. However, the most of multivariable pole-placement controllers are explicit and have considerable drawbacks in that the control designs involve the solution of a Diophantine equations, which in some applications may lead to excessive computational and numerical instability problems and they are obtained as a right matrix-fraction and an additional transforming step from a right to left-matrix description is required in order to implement the control law Zhu Q., & et al., (2002).

In an attempt to avoid solving Diophantine equations and to obtain the control as a left matrix-fraction for direct implementation, a novel multivariable generalised minimum variance stochastic adaptive controller with PID pole placement structure is presented in this paper. It builds on the previous works Zhu Q., & et al., (2002); Zayed, A. & et al., (2004); and Zayed A. & Hussain A., (2004). The proposed design provides the designer with a choice of using either a self-tuning controller or an implicit PID controller.

The paper is organised as follows: the derivation of the control law is discussed in section 2. In section 3, a simulation case study is carried out in order to demonstrate the effectiveness of the proposed controller in the performance of the closed loop system. Finally, some concluding remarks are presented in section 4.

2. Derivation of control law

In deriving the multivariable self-tuning control law we assume that the process is described by the following Controlled Auto-Regressive Moving Average (CARMA) model Yusof R. & et al., (1994); Zayed, A. & et al., (2004); Astrom K.., & Wittenmark B., (1973):

$$\mathbf{A}(z^{-1})\mathbf{y}(t) = \mathbf{B}(z^{-1})\mathbf{u}(t-k) + \mathbf{C}(z^{-1})\xi(t) \tag{1}$$

where $\mathbf{y}(t)$ is the measured output vector with dimension $(n \times 1)$, $\mathbf{u}(t)$ is the measured control input vector $(n \times 1)$, $\xi(t)$ is an uncorrelated sequence of random variables with zero mean, k is the time delay in the integer sampling interval and (t) denotes the sampling instant, $t = 1, 2, 3, ..., .$

The polynomial matrices $\mathbf{A}(z^{-1})$, $\mathbf{B}(z^{-1})$ and $\mathbf{C}(z^{-1})$ are expressed in terms of the backwards shift operator, z^{-1} {i.e. $z^{-1}\mathbf{x}(t) = \mathbf{x}(t-1)$}, and are given as:

$$\mathbf{A}(z^{-1}) = \mathbf{I} + \mathbf{A}_1 z^{-1} + \mathbf{A}_2 z^{-2} + ... + \mathbf{A}_{n_a} z^{-n_a} \tag{2}$$

$$\mathbf{B}(z^{-1}) = \mathbf{B}_0 + \mathbf{B}_1 z^{-1} + ... + \mathbf{B}_{n_b} z^{-n_b} \quad , \mathbf{B}(0) \neq 0 \tag{3}$$

$$\mathbf{C}(z^{-1}) = \mathbf{I} + \mathbf{C}_1 z^{-1} + \mathbf{C}_2 z^{-2} + ... + \mathbf{C}_{n_c} z^{-n_c} \tag{4}$$

where n_a, n_b, and n_c are the degrees of the polynomials.

The coefficients of the above polynomials are $(n \times n)$ matrices and $\mathbf{I}$ is the $(n \times n)$ identity matrix.

It is assumed that the zeroes of the det $\mathbf{C}(z^{-1})$ lie inside the unit disc of the z-plane (that is, the polynomial $\mathbf{C}(z^{-1})$ is inverse stable). It is also assumed without any loss of generality, that the disturbance transfer function is proper (i.e. $n_c \le n_a$) [1, 4, 8]. No assumption concerning the polynomial $\mathbf{B}(z^{-1})$ is made implying that the process can be a minimum or non-minimum phase system.
The control law minimises the variance of an auxiliary output $\phi(t)$:

$$\varphi(t) = \mathbf{P}(z^{-1})\mathbf{y}(t) + \mathbf{Q}'(z^{-1})\mathbf{u}(t-k) - \mathbf{R}(z^{-1})\mathbf{w}(t-k) \tag{5}$$

Here $\mathbf{w}(t)$ is the $(n \times 1)$ set point vector and $\mathbf{P}(z^{-1})$, $\mathbf{Q}'(z^{-1})$ and $\mathbf{R}(z^{-1})$ are the user-defined transfer functions in the backward shift operator z^{-1}. $\mathbf{P}(z^{-1})$ is rational matrix which can be expressed as:

$$\mathbf{P}(z^{-1}) = \mathbf{P}_n(z^{-1})\mathbf{P}_d^{-1}(z^{-1}) \tag{6}$$

here $\mathbf{P}_n(z^{-1})$ and $\mathbf{P}_d(z^{-1})$ are respectively monic $(n \times n)$ numerator and denominator matrices with degrees n_{p_n} and n_{p_d}. The performance of closed loop system is determined by the selection of the polynomial matrices $\mathbf{P}(z^{-1})$, $\mathbf{Q}'(z^{-1})$ and $\mathbf{R}(z^{-1})$ which are important design decisions.
The control law which minimises the above cost function given by (5) can be expressed as(Zayed, A. & et al., (2004)):

$$\mathbf{Q}_s\mathbf{u}(t) = [\mathbf{H}_0\mathbf{w}(t) - \bar{\mathbf{F}}_s'\mathbf{y}(t)] \tag{7}$$

where $\mathbf{Q}_s$ and $\mathbf{H}_0$ are the user transfer functions matrices which they depend on $\mathbf{Q}$ and $\mathbf{R}$, respectively.
We further assume that $\mathbf{Q}_s$ and $\mathbf{H}_0$ can also be expressed as:

$$\left.\begin{aligned} \mathbf{H}_0 &= \bar{\mathbf{F}}_s' \\ \mathbf{Q}_s &= \Delta(\tilde{\mathbf{H}}'\mathbf{V})^{-1}\bar{\mathbf{Q}}_s' \\ \Delta &= \Delta\mathbf{I} = (1 - z^{-1})\mathbf{I} \end{aligned}\right\} \tag{8}$$

Δ and $\bar{\mathbf{Q}}_s'$ are the $(n \times n)$ polynomial matrices, and $\mathbf{V}$ is a user-defined polynomial diagonal gain matrix. Here $\tilde{\mathbf{H}}'$ and $\mathbf{I}$ are the $(n \times n)$ desired closed loop system zeros polynomial matrix and identity matrix, respectively.
Combining equations (7) and (8), gives:

$$\Delta\bar{\mathbf{Q}}_s'\mathbf{u}(t) = [\mathbf{V}\bar{\mathbf{F}}_s'\mathbf{w}(t) - \tilde{\mathbf{H}}'\mathbf{V}\bar{\mathbf{F}}_s'\mathbf{y}(t)] \tag{9}$$

In this case,

$$\bar{\mathbf{Q}}'_s(z^{-1}) = \bar{\mathbf{Q}}'_0 + \bar{\mathbf{Q}}'_1 z^{-1} + \ldots\ldots + \bar{\mathbf{Q}}'_{n_{\bar{Q}'_s}} z^{-n_{\bar{Q}'_s}} \tag{10}$$

It can clearly be seen from (8) that the polynomial $\bar{\mathbf{Q}}'_s$ and the gain $\mathbf{V}$ can be considered as user defined parameters since they depend on the user transfer function $\mathbf{Q}'(z^{-1})$.

We can see clearly from equations (8) and (9) that the controller denominator has now conveniently been split into two parts:

1. An integrator action part (Δ) required for PID design
2. An arbitrary compensator ($\bar{\mathbf{Q}}'_s(z^{-1})$) that may be used for pole-placement placement design.

2.1 Multivariable Self-tuning PID Controller design (mode 1)

In this mode, the generalised minimum variance controller operates as a conventional self-tuning PID controller, which can be expressed in the most commonly used velocity form [1, 2] as:

$$\Delta\mathbf{u}(t) = [\mathbf{K_P} + \mathbf{K_I} + \mathbf{K_D}]\mathbf{e}(t) - [\mathbf{K_P} + 2\mathbf{K_D}]\mathbf{e}(t-1) + [\mathbf{K_D}]\mathbf{e}(t-2) \tag{11}$$

$$\mathbf{e}(t) = \mathbf{w}(t) - \mathbf{y}(t) \tag{12}$$

where $\mathbf{K_P}$, $\mathbf{K_I}$ and $\mathbf{K_D}$ are $(n \times n)$ matrices denote the proportional gain, the integral gain and derivative gain respectively. Δ is the difference operator defined as:

In order to obtain a self-tuning controller with PID structure the control law in equation (9) must have the same form of as the PID controller in equation (11).

If we assume that the degree of polynomial $\bar{\mathbf{F}}'_s$ is equal to 2:

$$\bar{\mathbf{F}}'(z^{-1}) = \bar{\mathbf{F}}'_0 + \bar{\mathbf{F}}'_1 z^{-1} + \bar{\mathbf{F}}'_2 z^{-2} \tag{13a}$$

and if we set

$$\bar{\mathbf{Q}}'_s(z^{-1}) = \tilde{\mathbf{H}}' = \mathbf{I} \tag{13b}$$

and make use of equations (13a), (13b) and (9) a multivariable self-tuning controller with PID structure is obtained, where

$$\Delta\bar{\mathbf{Q}}'_s\mathbf{u}(t) = \mathbf{V}(\bar{\mathbf{F}}'_0 + \bar{\mathbf{F}}'_1 z^{-1} + \bar{\mathbf{F}}'_2 z^{-2})[\mathbf{w}(t) - \mathbf{y}(t)] \tag{14}$$

$$\mathbf{K_P} = -\mathbf{V}(\bar{\mathbf{F}}'_1 + 2\bar{\mathbf{F}}'_2) \tag{15a}$$

$$\mathbf{K_I} = \mathbf{V}(\bar{\mathbf{F}}'_0 + \bar{\mathbf{F}}'_1 + \bar{\mathbf{F}}'_2) \tag{15b}$$

$$\mathbf{K_D} = \mathbf{V}(\bar{\mathbf{F}}'_2) \tag{15c}$$

As can be seen from equations (5) and (8)-(15) that the PID controller is tuned by a selection of the polynomial $\mathbf{P}$ and the gain $\mathbf{V}$ which must be selected in trial and error procedure.

Alternatively, these tuning parameters can be automatically and implicitly set on line by specifying the desired closed loop poles Zhu Q., & et al., (2002); Zayed A. & Hussain A., (2004).

2.2 New Implicit Multivariable PID Pole-placement Controller (Mode 2)

The generalised minimum variance control law given by equation (9) was extended to achieve explicit PID pole-zero placement by Zayed A., (1997); Zayed A., (2005) Zayed A.,& et al., (2006) and Zhu and Zhu Q., & et al., (2002). However, these explicit PID controller designs have two drawbacks in that the controllers involve the solution of Diophantine equation. In addition, the explicit designs have the right fraction structure and an additional transforming step from a right to left-matrix description is required in order to implement the control law. For this reasons the generalised minimum variance control is modified such that solving Diophantine is not considered in the design and has a left fraction structure enables direct implementation.. The controller may then be considered as an implicit controller in the sense that the control design step is trivial.

If we set the desired closed loop zeros matrix $\tilde{\mathbf{H}}' = \mathbf{I}$, then the control law given by equation (9) can also expressed as follows:

$$\bar{\mathbf{q}}_s'\mathbf{u}(t) = [\mathbf{V}\bar{\mathbf{F}}_s'\mathbf{w}(t) - \mathbf{V}\bar{\mathbf{F}}_s'\mathbf{y}(t)] \tag{16}$$

where

$$\bar{\mathbf{q}}_s' = \Delta\bar{\mathbf{Q}}_s' \tag{17}$$

By combining equations (16) and (1), the closed loop transfer function is obtained as:

$$\mathbf{y}(t) = (\mathbf{A} + z^{-k}\mathbf{B}[\bar{\mathbf{q}}_s']^{-1}\mathbf{V}\bar{\mathbf{F}}_s')^{-1}[z^{-k}\mathbf{B}(\bar{\mathbf{q}}_s')^{-1}(\mathbf{V}\bar{\mathbf{F}}_s')\mathbf{w}(t) + \mathbf{C}\xi(t)] \tag{18}$$

If we set

$$\bar{\mathbf{F}}_s' = \mathbf{A} \tag{19}$$

then equation (18) becomes after some arrangement:

$$\mathbf{y}(t) = [\mathbf{A} + z^{-k}\mathbf{B}(\bar{\mathbf{q}}_s')^{-1}\mathbf{V}\mathbf{A}]^{-1}(z^{-k}\mathbf{B})(\bar{\mathbf{q}}_s')^{-1}[\mathbf{V}(\mathbf{A})\mathbf{w}(t) + \bar{\mathbf{q}}_s'(z^{-k}\mathbf{B})^{-1}\mathbf{C}\xi(t)] \tag{20}$$

Next, we can introduce the following relation [6]:

$$(\mathbf{A} + z^{-k}\mathbf{B}[\bar{\mathbf{q}}_s']^{-1}\mathbf{V}\mathbf{A})^{-1}z^{-k}\mathbf{B}(\bar{\mathbf{q}}_s')^{-1} = \mathbf{A}^{-1}z^{-k}\mathbf{B}(\bar{\mathbf{q}}_s' + z^{-k}\mathbf{V}\mathbf{B})^{-1} \tag{21}$$

Making use of equations (20) and (21), we obtain:

$$\mathbf{y}(t) = \mathbf{A}^{-1}z^{-k}\mathbf{B}(\bar{\mathbf{q}}_s' + z^{-k}\mathbf{V}\mathbf{B})^{-1}[\mathbf{V}(\mathbf{A})\mathbf{w}(t) + \bar{\mathbf{q}}_s'(z^{-k}\mathbf{B})^{-1}\mathbf{C}\xi(t)] \tag{22}$$

The desired closed loop configuration is achieved by setting:

$$(\bar{\mathbf{q}}'_s + z^{-k}\mathbf{VB}) = \mathbf{CT}'\mathbf{K}' \tag{23}$$

where $\mathbf{T}'$ represents the desired closed loop poles.
It is assumed, without loss of generality, that $\mathbf{T}'$ is normalised such that

$$\mathbf{T}'(1) = \mathbf{I} \tag{24}$$

The above equation can easily be satisfied by selecting $\mathbf{T}'$ such that [6]:

$$\mathbf{T}'(z^{-1}) = (\mathbf{I} + \mathbf{T}'_1 + ..., ... + \mathbf{T}'_{n_{T'}})^{-1}(\mathbf{I} + \mathbf{T}'_1 z^{-1} + \mathbf{T}_2 z^{-2} + ..., ... + \mathbf{T}'_{n_{T'}} z^{n_{T'}}) \tag{25}$$

Here $\mathbf{K}'$ is $(n \times n)$ user-defined gain matrix that has to be chosen such that the steady state error is zero. It can be seen from (17) and (23) that the user-defined gain matrix $\mathbf{K}'$ is employed to ensure the incorporation of the integral action into the design (i.e. $\bar{\mathbf{q}}'_s(1)$ in equation (23) equal to zero).
where $n_{T'}$ represents the degree of the polynomials $\mathbf{T}'$.
Using equations (19) and (23) and rearranging, we obtain:

$$\bar{\mathbf{q}}'_s = \mathbf{CT}'\mathbf{K}' - z^{-k}\mathbf{VB} \tag{26}$$

It can be seen from (17) and (26) that in order to ensure that $\bar{\mathbf{q}}'_s$ involves (Δ)(i.e. $\bar{\mathbf{q}}'_s(1)$ in equation (23) equal to zero), we set:

$$\bar{\mathbf{q}}'_s(1) = \mathbf{C}(1)\mathbf{T}'(1)\mathbf{K}' - \mathbf{VB}(1) = 0 \tag{27}$$

The above equation (27) can be satisfied by setting:

$$\mathbf{K}' = [\mathbf{C}(1)\mathbf{T}'(1)]^{-1}(\mathbf{VB}(1)) = [\mathbf{C}(1)]^{-1}\mathbf{VB}(1) \tag{28}$$

We can easily compute $\bar{\mathbf{Q}}'_s$ from equation (17) as follows:

$$\left.\begin{aligned} \bar{\mathbf{Q}}'_{s_0} &= \bar{\mathbf{q}}'_{s_0} \\ \bar{\mathbf{Q}}'_{s_i} &= \sum_{j=0}^{i} \bar{\mathbf{q}}'_{s_i} \end{aligned}\right\} \tag{29}$$

If we assume that the degree of $\bar{\mathbf{F}}'_s(z^{-1})$ is equal to 2, then equation (16) becomes:

$$\Delta\bar{\mathbf{Q}}'_s\mathbf{u}(t) = [\mathbf{V}(\bar{\mathbf{F}}'_{s0} + \bar{\mathbf{F}}'_{s1}z^{-1} + \bar{\mathbf{F}}'_{s2}z^{-2})][\mathbf{w}(t) - \mathbf{y}(t)] \tag{30}$$

However, the zeros may be used to achieve better set point tracking or they may also help reduce the magnitude of the control action [5, 7, 11]. In the following section (2.3) a new implicit zero pole-placement is derived.

2.3 New Implicit Multivariable PID Zero Pole-placement Controller (Mode 3)

If we set the desired closed loop zeros matrix $\tilde{\mathbf{H}}' \neq \mathbf{I}$, then the control law given by equation (9) can be expressed as follows:

$$\bar{\mathbf{q}}_s'\mathbf{u}(t) = [\mathbf{V}\bar{\mathbf{F}}_s'\mathbf{w}(t) - \tilde{\mathbf{H}}'\mathbf{V}\bar{\mathbf{F}}_s'\mathbf{y}(t)] \tag{31}$$

By combining equations (31) and (1), the closed loop transfer function is obtained as:

$$\mathbf{y}(t) = (\mathbf{A} + z^{-k}\mathbf{B}[\bar{\mathbf{q}}_s']^{-1}\tilde{\mathbf{H}}'\mathbf{V}\bar{\mathbf{F}}_s')^{-1}[z^{-k}\mathbf{B}(\bar{\mathbf{q}}_s')^{-1}(\tilde{\mathbf{H}}'\mathbf{V}\bar{\mathbf{F}}_s')\mathbf{w}(t) + \mathbf{C}\boldsymbol{\xi}(t)] \tag{32}$$

If we assume, without loss of generality, at steady state that:

$$\left.\begin{aligned} \bar{\mathbf{F}}_s' &= \mathbf{A}\mathbf{K}_0 \\ \mathbf{H}_s' &= \tilde{\mathbf{H}}'\mathbf{V} \end{aligned}\right\} \tag{33}$$

then equation (32) becomes after some arrangement:

$$\begin{aligned} \mathbf{y}(t) = [\mathbf{A} + z^{-k}\mathbf{B}(\bar{\mathbf{q}}_s')^{-1}\tilde{\mathbf{H}}_s'\mathbf{A}]^{-1}(z^{-k}\mathbf{B})(\bar{\mathbf{q}}_s')^{-1}[(\tilde{\mathbf{H}}_s'\mathbf{A}(1)\mathbf{K}_0)\mathbf{w}(t) + \\ \bar{\mathbf{q}}_s'(z^{-k}\mathbf{B})^{-1}\mathbf{C}\boldsymbol{\xi}(t)] \end{aligned} \tag{34}$$

Next, we can introduce the following relation [5]:

$$(\mathbf{A} + z^{-k}\mathbf{B}[\bar{\mathbf{q}}_s']^{-1}\tilde{\mathbf{H}}_s'\mathbf{A})^{-1}z^{-k}\mathbf{B}(\bar{\mathbf{q}}_s')^{-1} = \mathbf{A}^{-1}z^{-k}\mathbf{B}(\bar{\mathbf{q}}_s' + z^{-k}\tilde{\mathbf{H}}_s'\mathbf{B})^{-1} \tag{35}$$

Making use of equations (35) and (34), we obtain:

$$\mathbf{y}(t) = \mathbf{A}^{-1}z^{-k}\mathbf{B}(\bar{\mathbf{q}}_s' + z^{-k}\tilde{\mathbf{H}}_s'\mathbf{B})^{-1}[(\tilde{\mathbf{H}}_s'\mathbf{A}\mathbf{K}_0)\mathbf{w}(t) + \bar{\mathbf{q}}_s'(z^{-k}\mathbf{B})^{-1}\mathbf{C}\boldsymbol{\xi}(t)] \tag{36}$$

The desired closed loop configuration is achieved by setting:

$$(\bar{\mathbf{q}}_s' + z^{-k}\tilde{\mathbf{H}}_s'\mathbf{B}) = \mathbf{C}\mathbf{T}'\mathbf{K}' \tag{37}$$

where $\mathbf{T}'$ represents the desired closed loop poles and $\tilde{\mathbf{H}}'$ represents the desired closed loop zeros.

It is assumed, without loss of generality, that $\mathbf{T}'$ and $\tilde{\mathbf{H}}'$ are normalised such that

$$\mathbf{T}'(1) = \tilde{\mathbf{H}}'(1) \tag{38}$$

The above equation can easily be satisfied by selecting $\mathbf{T}'$ and $\tilde{\mathbf{H}}'$ such that [5]:

$$\tilde{\mathbf{H}}'(z^{-1}) = (\mathbf{I} + \tilde{\mathbf{h}}' + ... + \tilde{\mathbf{h}}_{n_{\tilde{h}'}})^{-1}(\mathbf{I} + \tilde{\mathbf{h}}'_1 z^{-1} + ... + \tilde{\mathbf{h}}_{n_{\tilde{h}'}} z^{-n_{\tilde{h}'}}) \tag{39}$$

$$\mathbf{T}'(z^{-1}) = (\mathbf{I} + \mathbf{T}'_1 + ..., ... + \mathbf{T}'_{n_{T'}})^{-1}(\mathbf{I} + \mathbf{T}'_1 z^{-1} + \mathbf{T}_2 z^{-2} + ..., ... + \mathbf{T}'_{n_{T'}} z^{n_{T'}}) \tag{40}$$

Here $\mathbf{K}'$ is the $(n \times n)$ user-defined gain matrix that has to be chosen such that the steady state error is zero. It can be seen from (17) and (37) that the user-defined gain matrix $\mathbf{K}'$ is employed to ensure the incorporation of the integral action into the design (i.e. $\bar{\mathbf{q}}'_s(1)$ in equation (37) equal to zero) [5].

where $n_{T'}$ and $n_{\tilde{h}'}$ represents the degree of the polynomials $\mathbf{T}'$ and $\tilde{\mathbf{H}}'$, respectively.

Using equations(33) and (37) and rearranging, we obtain:

$$\bar{\mathbf{q}}'_s = \mathbf{C}\mathbf{T}'\mathbf{K}' - z^{-k}\tilde{\mathbf{H}}'\mathbf{V}\mathbf{B} \tag{41}$$

It can be seen from (17) and (41) that in order to ensure that $\bar{\mathbf{q}}'_s$ involves (Δ)(i.e. $\bar{\mathbf{q}}'_s(1)$ in equation (37) equal to zero), we set:

$$\bar{\mathbf{q}}'_s(1) = \mathbf{C}(1)\mathbf{T}'(1)\mathbf{K}' - \tilde{\mathbf{H}}'(1)\mathbf{V}\mathbf{B}(1) = 0 \tag{42}$$

The above equation (42) can be satisfied by setting:

$$\mathbf{K}' = [\mathbf{C}(1)\mathbf{T}'(1)]^{-1}(\tilde{\mathbf{H}}'(1)\mathbf{V}\mathbf{B}(1)) = [\mathbf{C}(1)]^{-1}\mathbf{V}\mathbf{B}(1) \tag{43}$$

It can be seen from equation (36) that the closed loop poles will be placed in the desired locations if we assume the following:

$$\left.\begin{aligned} \mathbf{X}_s &= \tilde{\mathbf{H}}'_s\mathbf{A} \\ \mathbf{K}_0 &= \mathbf{K}'_0[\mathbf{K}'_0(1)]^{-1} \\ \tilde{\mathbf{K}}'_0\tilde{\mathbf{X}}_s &= \mathbf{X}_s\mathbf{K}'_0 \\ \tilde{\mathbf{K}}'_0 &= \mathbf{C} \end{aligned}\right\} \tag{44}$$

The new implicit multivariable pole-zero placement controller block diagram is shown in Figure (1a).

The implicit pole-zero placement controller illustrated in Figure (1a) is now extended to combine the advantages of both PID control and pole-zero placement control. In order to show the inherent incorporation of the PID control explicitly in our design, the polynomial $\bar{\mathbf{q}}'_s$ in equation (17) must be split into an integral action (Δ) part and a pole-placement compensator $\bar{\mathbf{Q}}'_s$.

We can easily compute $\bar{\mathbf{q}}'_s$ from equation (17) as follows:

$$\left.\begin{aligned}\bar{\mathbf{Q}}'_{s_0} &= \bar{\mathbf{q}}'_{s_0}\\ \bar{\mathbf{Q}}'_{s_i} &= \sum_{j=0}^{i}\bar{\mathbf{q}}'_{s_i}\end{aligned}\right\} \tag{45}$$

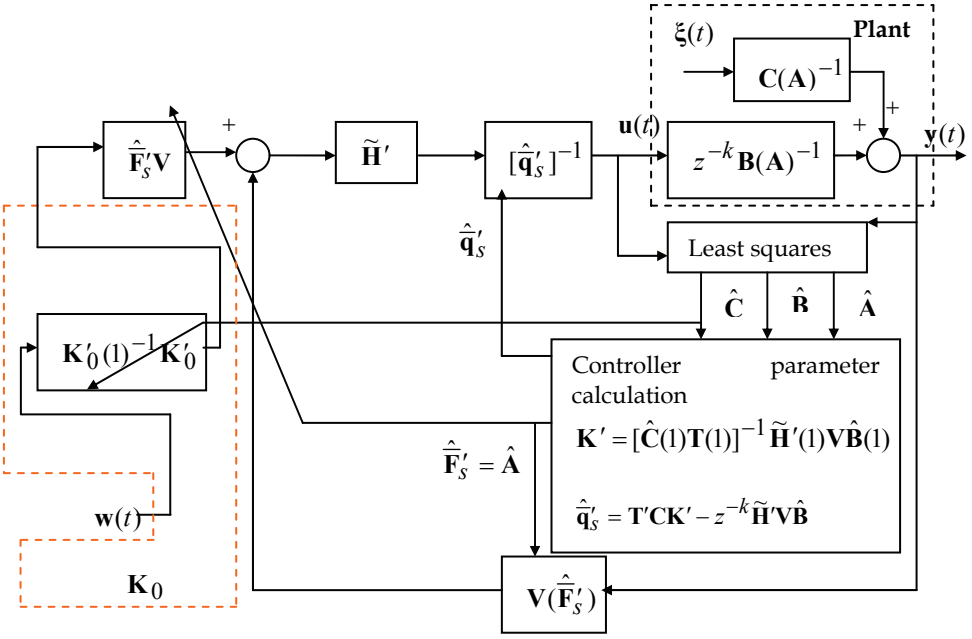


Fig. (1a). New implicit multivariable pole-zero placement controller.

If we assume that the degree of $\bar{\mathbf{F}}'_s(z^{-1})$ is equal to 2, then equation (31) becomes:

$$\begin{aligned}\Delta\bar{\mathbf{Q}}'_s\mathbf{u}(t) = \tilde{\mathbf{H}}[&\mathbf{V}(\bar{\mathbf{F}}'_{s0} + \bar{\mathbf{F}}'_{s1}z^{-1} + \bar{\mathbf{F}}'_{s2}z^{-2})\mathbf{K}_0\mathbf{w}(t) -\\ &\mathbf{V}(\bar{\mathbf{F}}'_{s0} + \bar{\mathbf{F}}'_{s1}z^{-1} + \bar{\mathbf{F}}'_{s2}z^{-2})\mathbf{y}(t)]\end{aligned} \tag{46}$$

The controller parameters $\mathbf{K}_0$, $\bar{\mathbf{Q}}'_s$ and $\bar{\mathbf{F}}'_s(z^{-1})$ in the above equation (46) are obtained from the equations (44), (46) and (33) respectively.

The implicit pole-zero placement control law given by equation (46) is shown in Figure (1b). It can be seen from the above equation (46) and Figure (1b) that the pole-zero placement controller can be represented by an equivalent controller consisting of a PID controller plus three compensators labelled as compensator 1, compensator 2 and compensator 3 in the Figure (1b). The first compensator is used to ensure that at steady state, the output signal tracks the set point. The compensator 2 is used to achieve pole-placement control and compensator 3 is used to achieve zero-placement.

From equation (33) it is clear that a PI controller is achieved if the polynomial $\mathbf{A}(z^{-1})$ is of order 1, whereas, a PID controller is achieved if $\mathbf{A}(z^{-1})$ is of order 2.
A pure PI/PID control is achieved if these three compensators are switched off.
The algorithm for the pole-zero placement controller can then be summarised as follows:

Step 1. Select the desired closed-loop system poles and zeros polynomial matrices, $\mathbf{T}'(z^{-1})$ and $\tilde{\mathbf{H}}'(z^{-1})$ respectively, and select the user-defined gain matrix $\mathbf{V}$.

Step 2. Read the new values of the output $\mathbf{y}(t)$, the control input $\mathbf{u}(t)$ and reference signal $\mathbf{w}(t)$

Step 3. Estimate the process parameters $\hat{\mathbf{A}}$, $\hat{\mathbf{B}}$, and $\hat{\mathbf{C}}$ using the linear least squares algorithm.

Step 4. Set $\overline{\mathbf{F}}_s'(z^{-1}) = \hat{\mathbf{A}}(z^{-1})$.

Step 5. Compute $\mathbf{K}_0$ and $\mathbf{K}'$ using equations (44) and (43), respectively.

Step 6. Compute $\hat{\overline{\mathbf{q}}}_s'$ using equation (41).

Step 7. Apply the control law using equation (31).

Steps 2 to 7 are repeated for every sampling instant.

3 Simulation results

The objective of this section is to study the performance and the robustness of the proposed multivariable pole-zero placement controller.
Two simulation examples are carried out in order to demonstrate the ability of the proposed algorithm to locate the closed loop poles and zeros at their pre-specified locations under set point changes. The simulation study also includes an investigation of the influence of the load disturbances and stochastic disturbances on the system. In all performed simulations the least squares estimator has been employed and 800 samples are used with a set point change every 100 sampling instants.
In order to demonstrate the closed loop performance of the implicit controller we arrange manually (for reason of comparison) the controller to work in three control modes, namely as a PID pole-placement controller, a PID pole-zero placement controller and as a PID self-tuning controller as described below:
a) From 0th up to 250th sampling time, the implicit PID pole-placement controller is selected to operate on-line.
b) The Implicit PID pole-zero placement controller is switched on from 251st to 550th sampling times.
c) The conventional PID self-tuning controller is switched on from 551st to 800th sampling time.
Two case studies are considered in this section: a two-input two-output water bath system and a simulated non-minimum phase system.

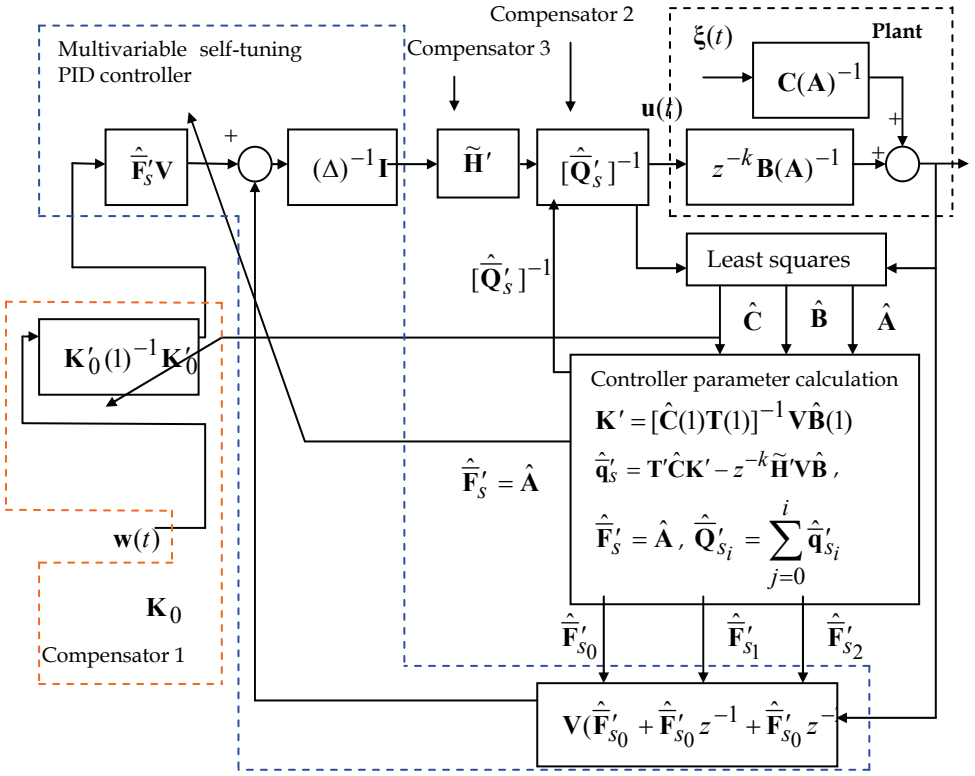


Fig. (1b). Novel implicit multivariable PID pole-zero placement controller.

3.1 Case study 1: Two-input Two-output Water Bath System Simulation results

The algorithm proposed in sections (2.2) was applied to a two-input two-output water bath treated previously by Yusof et al. [2, 13] and Zayed et al. [4, 10, 11]. The water bath system is shown in Figure (2). The water bath is an example of an important component in many industrial chemical processes. The control objective is to bring the temperature of the water or some chemical product in the bath to the desired set-points as accurately as possible.
The discrete model of the water bath system can be written as [2, 10, 11]:

$$(\mathbf{I} + \mathbf{A}_1 z^{-1})\mathbf{y}(t) = \mathbf{B}_0\mathbf{u}(t-k) + \xi(t) \tag{47}$$

where

$$\mathbf{A}_1 = \begin{bmatrix} -0.411 & -0.634 \\ -0.103 & -0.885 \end{bmatrix},\ \mathbf{B}_0 = \begin{bmatrix} 0.492 & 0.085 \\ 0.041 & 0.237 \end{bmatrix}$$

and the sample time = 30 sec.
The simulation was performed over 800 samples (400 minutes) under set point $\mathbf{w}(t) = \begin{bmatrix} \mathbf{w}_1(t) \\ \mathbf{w}_2(t) \end{bmatrix}$ changes every 100 sampling instants as follows:

1) $\mathbf{w}_1(t)$ changes from 60^0C to 80^0C and from 80^0C to 60^0C.

2) $\mathbf{w}_2(t)$ changes from 35^0C to 55^0C and from 55^0C to 35^0C.

In each sampling instant the parameter estimations $\hat{\mathbf{A}}_1$ and $\hat{\mathbf{B}}_0$ are estimated using the least squares estimator and the steps summarised in section (2) are followed.

Note that, by selecting the pre-filter polynomial matrix $\mathbf{P}_d(z^{-1})$ to be of order one, a PI self-tuning controller is obtained.

The user-defined gain and the pre-filter polynomial matrices were respectively selected as:

$$\mathbf{V} = \begin{bmatrix} V_1 & 0 \\ 0 & V_2 \end{bmatrix} = \begin{bmatrix} 0.7 & 0 \\ 0 & 0.8 \end{bmatrix},\ \mathbf{P}_d(z^{-1}) = \mathbf{I} + \mathbf{P}_{d_1} z^{-1} \text{ and } \mathbf{P}_n(z^{-1}) = \mathbf{I} + \mathbf{P}_{n_1} z^{-1}.$$

where,

$$\mathbf{P}_{d_1}(z^{-1}) = \begin{bmatrix} -0.8 & 0 \\ 0 & -0.9 \end{bmatrix} \text{ and } \mathbf{P}_{n_1}(z^{-1}) = \begin{bmatrix} -0.3 & 0 \\ 0 & -0.4 \end{bmatrix}.$$

The desired closed loop poles polynomial matrix ($\mathbf{T}$) and the desired zero-placement polynomial matrix ($\tilde{\mathbf{H}}$) were selected as follows:

$$\left.\begin{aligned} \mathbf{T}(z^{-1}) &= \mathbf{I} + \mathbf{T}_1 z^{-1} + \mathbf{T}_2 z^{-2} \\ \tilde{\mathbf{H}}(z^{-1}) &= [\tilde{\mathbf{h}}(1)]^{-1}(\mathbf{I} + \tilde{\mathbf{h}}_1 z^{-1} + \tilde{\mathbf{h}}_2) z^{-2} \end{aligned}\right\}$$

where

$$\mathbf{T}_1 = \begin{bmatrix} -0.5 & 0 \\ 0 & -0.6 \end{bmatrix},\ \mathbf{T}_2 = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix},\ \tilde{\mathbf{h}}_1 = \begin{bmatrix} 0.9 & 0 \\ 0 & 0.8 \end{bmatrix} \text{ and } \tilde{\mathbf{h}}_2 = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}.$$

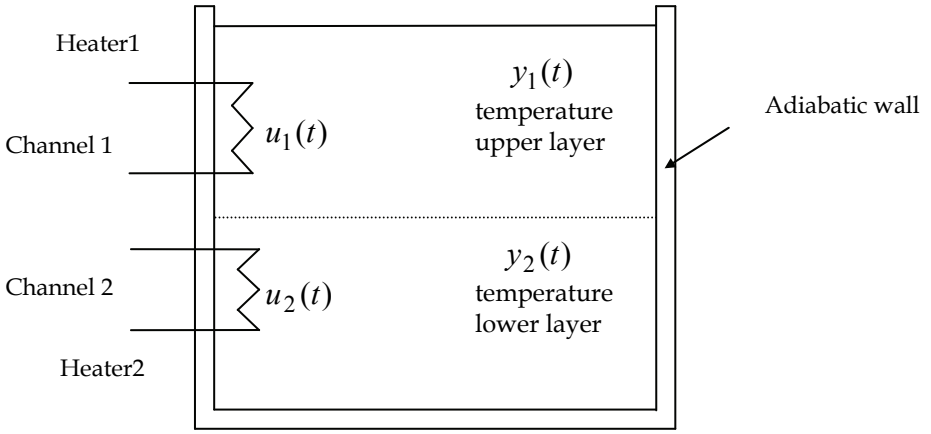
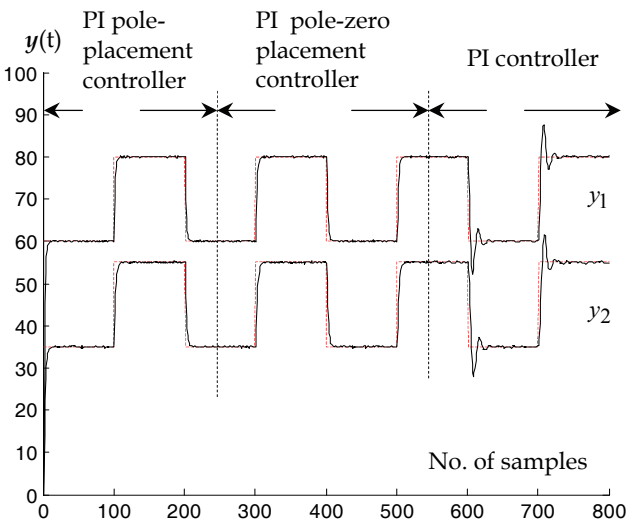


Fig. (2). A two channel water bath system

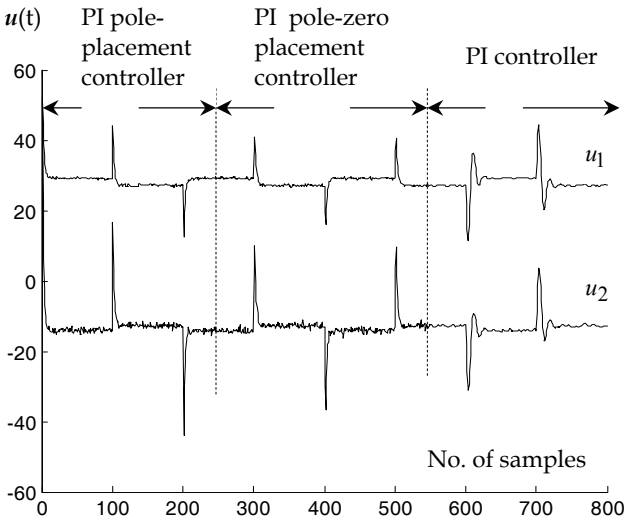


$$\mathbf{V}=\begin{bmatrix}0.7 & 0\\ 0 & 0.8\end{bmatrix},\quad \mathbf{P}_{n_1}=\begin{bmatrix}-0.3 & 0\\ 0 & -0.4\end{bmatrix},\quad \mathbf{P}_{d_1}=\begin{bmatrix}-0.8 & 0\\ 0 & -0.9\end{bmatrix},\quad \mathbf{T}_1=\begin{bmatrix}-0.5 & 0\\ 0 & -0.6\end{bmatrix}$$

and $\mathbf{T}_2=\begin{bmatrix}0 & 0\\ 0 & 0\end{bmatrix}$. $\tilde{\mathbf{h}}_1=\begin{bmatrix}0.9 & 0\\ 0 & 0.8\end{bmatrix}$

and $\tilde{\mathbf{h}}_2=\begin{bmatrix}0 & 0\\ 0 & 0\end{bmatrix}$.

Fig. (3a). The outputs



$$\mathbf{V}=\begin{bmatrix}0.7 & 0\\ 0 & 0.8\end{bmatrix},\quad \mathbf{P}_{n_1}=\begin{bmatrix}-0.3 & 0\\ 0 & -0.4\end{bmatrix},\quad \mathbf{P}_{d_1}=\begin{bmatrix}-0.8 & 0\\ 0 & -0.9\end{bmatrix},$$

$$\mathbf{T}_1=\begin{bmatrix}-0.5 & 0\\ 0 & -0.6\end{bmatrix}\quad \text{and}\quad \mathbf{T}_2=\begin{bmatrix}0 & 0\\ 0 & 0\end{bmatrix}.\quad \tilde{\mathbf{h}}_1=\begin{bmatrix}0.9 & 0\\ 0 & 0.8\end{bmatrix}\quad \text{and}\quad \tilde{\mathbf{h}}_2=\begin{bmatrix}0 & 0\\ 0 & 0\end{bmatrix}.$$

Fig. (3b). The control inputs

The outputs and the control inputs of the multiple controller are respectively shown in the Figures (3a) and (3b).

It is clear from these Figures (3a) and (3b) that, the transient response is significantly shaped by the choice of the polynomial $\mathbf{T}$ when either a PI pole-placement controller or a PI pole-zero placement controller is used. It can also clearly be seen from Figure (3b) that excessive control action, which resulted from set-point changes, is tuned most effectively when the new implicit PI zero-pole placement controller is on-line (during the sampling interval

251-550). Also note that during the last 250 samples (551-800 sampling times), where the conventional self tuning PI is operating, small oscillations can be seen in the control input and closed loop output, hence exhibiting the worst performance as expected, due to its inherent limitations. The other disadvantage of the self-tuning PID controller is that the tuning parameters must be selected using a trial and error procedure. The performance of the conventional PI controller can be improved by fine adjusting the user defined polynomial matrices $\mathbf{P}_n(z^{-1})$, $\mathbf{P}_d(z^{-1})$ and gain matrix $\mathbf{V}$.

The following simulation experiment investigates the effect of the user-defined parameter **V** on the response of the closed loop system when the PI multiple-controller is used.

3.1.1 Investigating the Influence of the Gain $\mathbf{V}$ on the Closed-Loop Performance

In order to see the effect of the user-defined gain on all controllers (the PI controller, PI pole-placement controller and PI pole-zero placement controller), the gain matrix **V** was changed from $\mathbf{V}=\begin{bmatrix}0.7 & 0\\ 0 & 0.8\end{bmatrix}$ to $\mathbf{V}=\begin{bmatrix}2 & 0\\ 0 & 0.8\end{bmatrix}$ (only V_1 was increased). The outputs and the control inputs are respectively shown in the Figures (4a) and (4b).

It is clear from these Figures (4a) and (4b) that increasing only the gain V_1 influences the outputs $y_1(t)$ and $y_2(t)$, when the PI controller is used, whereas the desired outputs are obtained (as expected) when the implicit PI pole-placement or the implicit PI pole-zero placement controllers is turned on. It can clearly be seen from Figure (4a) that the control action u_1 is increased.

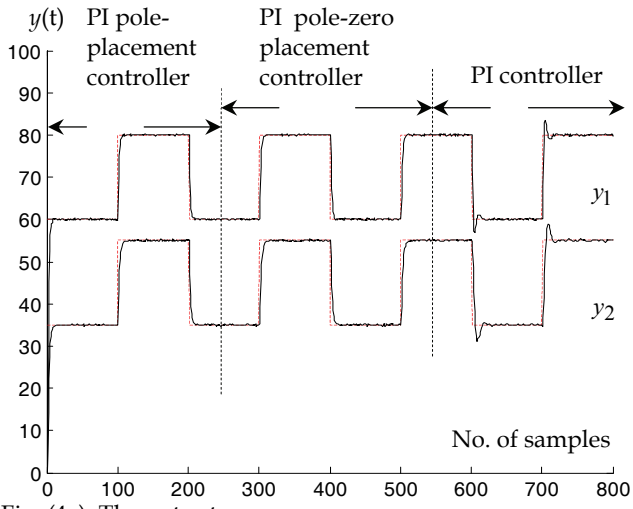


Fig. (4a). The outputs

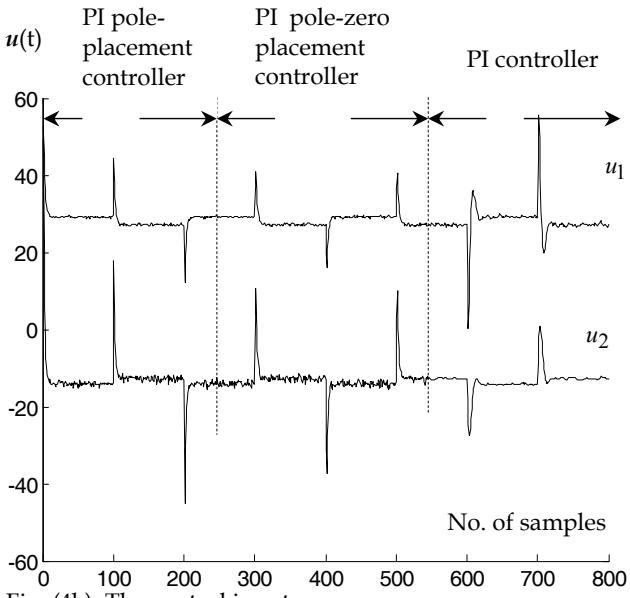
$$\mathbf{V}=\begin{bmatrix}2 & 0\\ 0 & 0.8\end{bmatrix},\ \mathbf{P}_{n_1}=\begin{bmatrix}-0.3 & 0\\ 0 & -0.4\end{bmatrix},$$

$$\mathbf{P}_{d_1}=\begin{bmatrix}-0.8 & 0\\ 0 & -0.9\end{bmatrix},$$

$$\mathbf{T}_1=\begin{bmatrix}-0.5 & 0\\ 0 & -0.6\end{bmatrix} \quad \text{and}$$

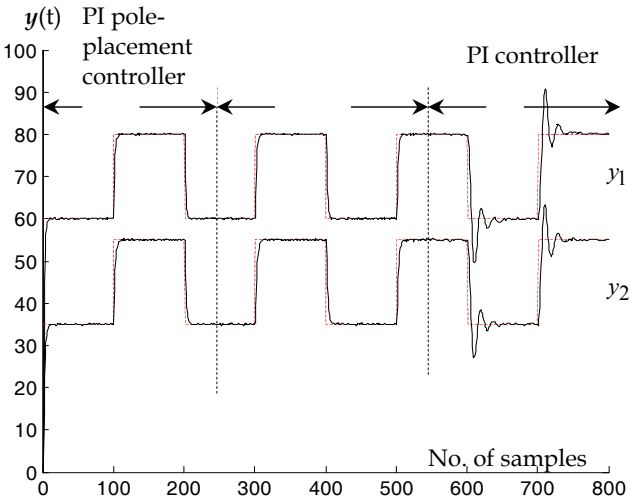
$$\mathbf{T}_2=\begin{bmatrix}0 & 0\\ 0 & 0\end{bmatrix},\ \tilde{\mathbf{h}}_1=\begin{bmatrix}0.9 & 0\\ 0 & 0.8\end{bmatrix} \quad \text{and}$$

$$\tilde{\mathbf{h}}_2=\begin{bmatrix}0 & 0\\ 0 & 0\end{bmatrix}.$$



$$\mathbf{V}=\begin{bmatrix}2 & 0\\ 0 & 0.8\end{bmatrix},\ \mathbf{P}_{n_1}=\begin{bmatrix}-0.3 & 0\\ 0 & -0.4\end{bmatrix},\ \mathbf{P}_{d_1}=\begin{bmatrix}-0.8 & 0\\ 0 & -0.9\end{bmatrix},\ \mathbf{T}_1=\begin{bmatrix}-0.5 & 0\\ 0 & -0.6\end{bmatrix}\ \text{and}\ \mathbf{T}_2=\begin{bmatrix}0 & 0\\ 0 & 0\end{bmatrix}.\ \tilde{\mathbf{h}}_1=\begin{bmatrix}0.9 & 0\\ 0 & 0.8\end{bmatrix}\ \text{and}\ \tilde{\mathbf{h}}_2=\begin{bmatrix}0 & 0\\ 0 & 0\end{bmatrix}.$$

Fig. (4b). The control inputs



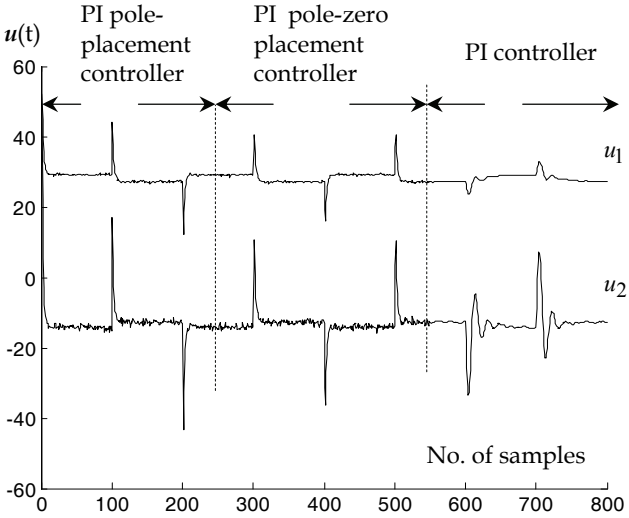
$$\mathbf{V}=\begin{bmatrix}0.1 & 0\\ 0 & 0.8\end{bmatrix},\ \mathbf{P}_{n_1}=\begin{bmatrix}-0.3 & 0\\ 0 & -0.4\end{bmatrix},\ \mathbf{P}_{d_1}=\begin{bmatrix}-0.8 & 0\\ 0 & -0.9\end{bmatrix},\ \mathbf{T}_1=\begin{bmatrix}-0.5 & 0\\ 0 & -0.6\end{bmatrix}\ \text{and}\ \mathbf{T}_2=\begin{bmatrix}0 & 0\\ 0 & 0\end{bmatrix}.\ \tilde{\mathbf{h}}_1=\begin{bmatrix}0.9 & 0\\ 0 & 0.8\end{bmatrix}\ \text{and}\ \tilde{\mathbf{h}}_2=\begin{bmatrix}0 & 0\\ 0 & 0\end{bmatrix}.$$

Fig. (5a). The outputs

The gain $\mathbf{V}$ was again changed from $\mathbf{V}=\begin{bmatrix}2 & 0\\ 0 & 0.8\end{bmatrix}$ to $\mathbf{V}=\begin{bmatrix}0.1 & 0\\ 0 & 0.8\end{bmatrix}$ (only V_1 is decreased).

The outputs and the control inputs are respectively shown in the Figures (5a) and (5b).

It can obviously be seen from the above Figures (5a) and (5b) that decreasing only the gain V_1 influences the outputs $y_1(t)$ and $y_2(t)$, when the PI controller is used, whereas the demanded outputs are achieved when the PI pole-placement or PI pole-zero placement is turned on.



$$\mathbf{V}=\begin{bmatrix}0.1 & 0\\ 0 & 0.8\end{bmatrix},\ \mathbf{P}_{n_1}=\begin{bmatrix}-0.3 & 0\\ 0 & -0.4\end{bmatrix},$$

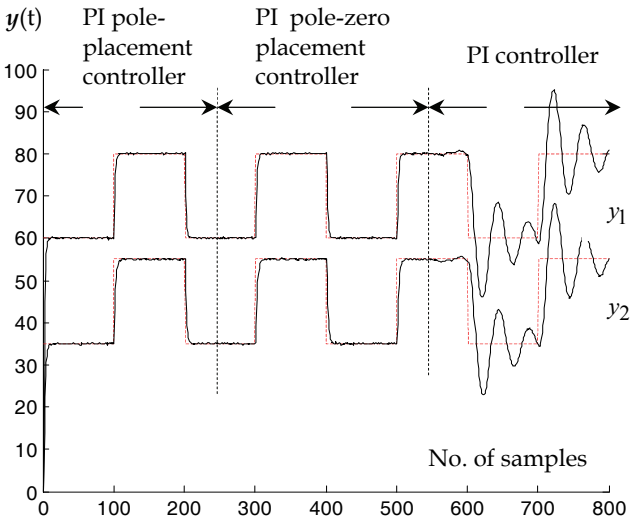
$$\mathbf{P}_{d_1}=\begin{bmatrix}-0.8 & 0\\ 0 & -0.9\end{bmatrix},$$

$$\mathbf{T}_1=\begin{bmatrix}-0.5 & 0\\ 0 & -0.6\end{bmatrix}$$

and $\mathbf{T}_2=\begin{bmatrix}0 & 0\\ 0 & 0\end{bmatrix}$. $\tilde{\mathbf{h}}_1=\begin{bmatrix}0.9 & 0\\ 0 & 0.8\end{bmatrix}$

and $\tilde{\mathbf{h}}_2=\begin{bmatrix}0 & 0\\ 0 & 0\end{bmatrix}$.

Fig. (5b). The control inputs



$$\mathbf{V}=\begin{bmatrix}0.1 & 0\\ 0 & 0.1\end{bmatrix},$$

$$\mathbf{P}_{n_1}=\begin{bmatrix}-0.3 & 0\\ 0 & -0.4\end{bmatrix},$$

$$\mathbf{P}_{d_1}=\begin{bmatrix}-0.8 & 0\\ 0 & -0.9\end{bmatrix},$$

$$\mathbf{T}_1=\begin{bmatrix}-0.5 & 0\\ 0 & -0.6\end{bmatrix}$$

and $\mathbf{T}_2=\begin{bmatrix}0 & 0\\ 0 & 0\end{bmatrix}$. $\tilde{\mathbf{h}}_1=\begin{bmatrix}0.9 & 0\\ 0 & 0.8\end{bmatrix}$

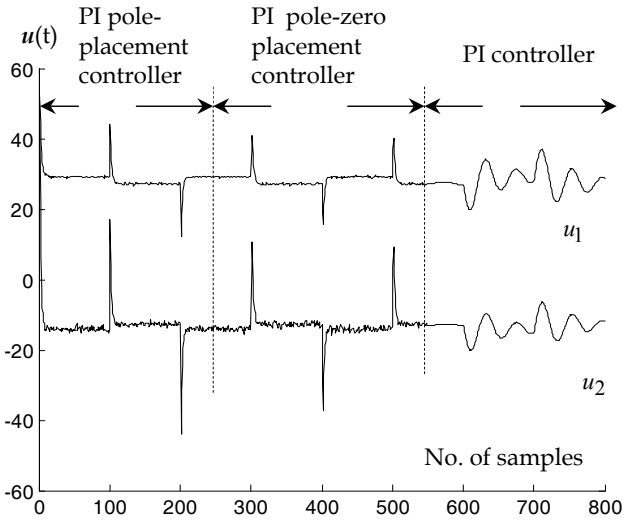
and $\tilde{\mathbf{h}}_2=\begin{bmatrix}0 & 0\\ 0 & 0\end{bmatrix}$.

Fig. (6a). The outputs

The gain **V** was further changed from $\mathbf{V}=\begin{bmatrix}0.1 & 0\\ 0 & 0.8\end{bmatrix}$ to $\mathbf{V}=\begin{bmatrix}0.1 & 0\\ 0 & 0.1\end{bmatrix}$ (only V_2 was changed).

The outputs and the control inputs are respectively shown in the Figures (6a) and (6b).

We can clearly see from the Figures (6a) and (6b) that changing the gain V_2 affects the outputs $y_1(t)$ and $y_2(t)$ only if the PI controller is used, whereas the desired outputs are obtained when either the implicit PI pole-placement controller or implicit PI pole-zero placement controller is used.



$$\mathbf{V}=\begin{bmatrix}0.1 & 0\\ 0 & 0.1\end{bmatrix},\ \mathbf{P}_{n_1}=\begin{bmatrix}-0.3 & 0\\ 0 & -0.4\end{bmatrix},$$

$$\mathbf{P}_{d_1}=\begin{bmatrix}-0.8 & 0\\ 0 & -0.9\end{bmatrix},$$

$$\mathbf{T}_1=\begin{bmatrix}-0.5 & 0\\ 0 & -0.6\end{bmatrix}\ \text{and}\ \mathbf{T}_2=\begin{bmatrix}0 & 0\\ 0 & 0\end{bmatrix}.$$

$$\tilde{\mathbf{h}}_1=\begin{bmatrix}0.9 & 0\\ 0 & 0.8\end{bmatrix}\ \text{and}\ \tilde{\mathbf{h}}_2=\begin{bmatrix}0 & 0\\ 0 & 0\end{bmatrix}.$$

Fig. (6b). The control inputs

It is clear from the previous figures that changing either V_1 or V_2 influences the closed loop system responses if the conventional self-tuning PI controller is used.

Excessive changes of either V_1 or V_2 will produce an unstable closed loop system if the PI self-tuning is used.

The gain matrix **V** has no influence on the closed system if the PI based Pole-zero placement controller is used, since the controller parameters change automatically in response to the change of the gain matrix **V** in order to place the closed loop system poles at pre-specified locations.

3.1.2 Investigating the Influence of the load disturbances on the Closed Loop Performance Using implicit Controller

The next task is to see the effect of the load disturbances on the closed system when the implicit PI pole zero placement for MIMO case is used. Artificial load disturbances of values 8°C and 5.5°C (10% of set point values) were added respectively to the outputs $y_1(t)$ and $y_2(t)$, from the 350th sampling interval to 800th sampling interval.

The two controller set points were both kept constant at values of 55°C and 80°C throughout. The outputs and the control inputs for PI pole-zero placement are shown in the Figures (7a) and (7b) respectively.

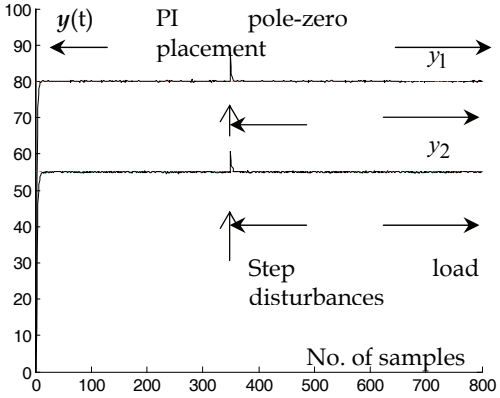


Fig. (7a). The outputs

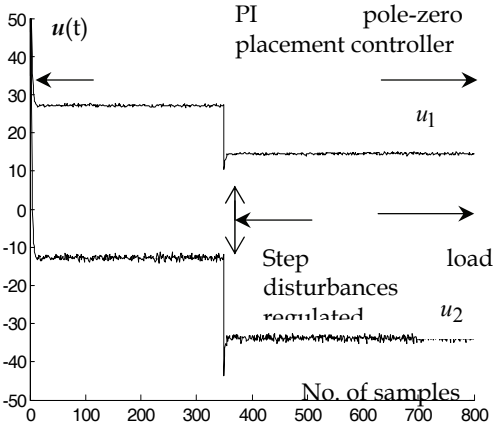


Fig. (7b). The control inputs

It can clearly be seen from all the figures (7a) and (7b) that at steady state, the proposed PI implicit controller has the ability to effectively regulate constant load disturbances to zero.
It can clearly be seen from all the figures (7a) and (7b) that at steady state, the proposed PI implicit controller has the ability to effectively regulate constant load disturbances to zero.

3.1.3 Investigating the Influence of the Polynomial $\tilde{H}$ on the Closed Loop Performance Using Implicit Controller

In order to see the influence of the zeros on the performance of the closed loop system, the implicit PI pole-zero placement was switched on from $t = 0$ up to $t = 800$ and the

polynomial matrix $\mathbf{T}$ was fixed as before and only the closed loop zeros polynomial $\tilde{\mathbf{H}}(z^{-1}) = [\tilde{\mathbf{h}}(1)]^{-1}(\mathbf{I} + \tilde{\mathbf{h}}_1 z^{-1} + \tilde{\mathbf{h}}_2 z^{-2})$ was changed three times as follows:

$$
\begin{aligned}
&0 \le t < 250 && \tilde{\mathbf{h}}(z^{-1}) = \mathbf{I} + \begin{bmatrix} 0.9 & 0 \\ 0 & 0.9 \end{bmatrix} z^{-1} + \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} z^{-2} \\
&250 \le t < 550 && \tilde{\mathbf{h}}(z^{-1}) = \mathbf{I} + \begin{bmatrix} -0.7 & 0 \\ 0 & -0.8 \end{bmatrix} z^{-1} + \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} z^{-2} \qquad (48) \\
&550 \le t < 800 && \tilde{\mathbf{h}}(z^{-1}) = \mathbf{I} + \begin{bmatrix} 0.9 & 0 \\ 0 & 1.4 \end{bmatrix} z^{-1} + \begin{bmatrix} 0.2 & 0 \\ 0 & 0.49 \end{bmatrix} z^{-2}
\end{aligned}
$$

The outputs and the control inputs are shown in the Figures (8a), (8b) and (8c), respectively. It is clear from the Figures (8a), (8b) and (8c) that changing the polynomial $\tilde{\mathbf{H}}$ affects the closed loop system performance. Excessive control input results from selecting unsuitable $\tilde{\mathbf{H}}$.

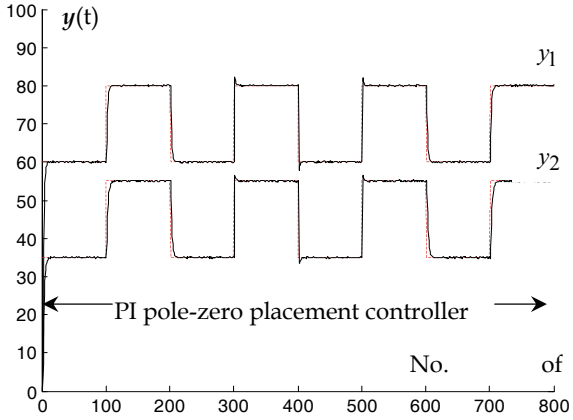


Fig. (8a). The outputs

$\mathbf{V} = \begin{bmatrix} 0.1 & 0 \\ 0 & 1.2 \end{bmatrix}$, $\mathbf{P}_{n_1} = \begin{bmatrix} -0.3 & 0 \\ 0 & -0.4 \end{bmatrix}$, $\mathbf{P}_{d_1} = \begin{bmatrix} -0.8 & 0 \\ 0 & -0.9 \end{bmatrix}$, $\mathbf{T}_1 = \begin{bmatrix} -0.5 & 0 \\ 0 & -0.6 \end{bmatrix}$ and $\mathbf{T}_2 = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$.

The desired zero- placement polynomial matrix $\tilde{\mathbf{h}}$ changes according to equation (48).

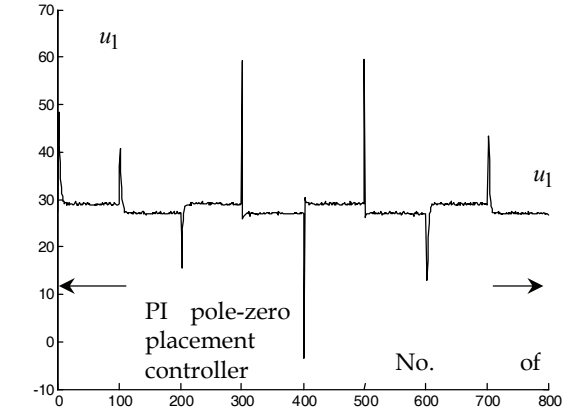


Fig. (8b) The control input

$\mathbf{V} = \begin{bmatrix} 0.1 & 0 \\ 0 & 1.2 \end{bmatrix}$, $\mathbf{P}_{n_1} = \begin{bmatrix} -0.3 & 0 \\ 0 & -0.4 \end{bmatrix}$, $\mathbf{P}_{d_1} = \begin{bmatrix} -0.8 & 0 \\ 0 & -0.9 \end{bmatrix}$, $\mathbf{T}_1 = \begin{bmatrix} -0.5 & 0 \\ 0 & -0.6 \end{bmatrix}$ and $\mathbf{T}_2 = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$.

The desired zero- placement polynomial matrix $\tilde{\mathbf{h}}$ changes according to equation (48).

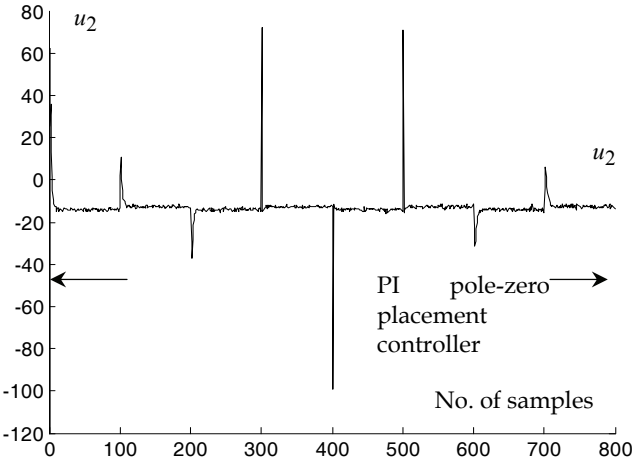


Fig. (8c). The control input

$$\mathbf{V}=\begin{bmatrix}0.1 & 0\\ 0 & 1.2\end{bmatrix},\ \mathbf{P}_{n_1}=\begin{bmatrix}-0.3 & 0\\ 0 & -0.4\end{bmatrix},$$

$$\mathbf{P}_{d_1}=\begin{bmatrix}-0.8 & 0\\ 0 & -0.9\end{bmatrix},$$

$$\mathbf{T}_1=\begin{bmatrix}-0.5 & 0\\ 0 & -0.6\end{bmatrix}$$

and $\mathbf{T}_2=\begin{bmatrix}0 & 0\\ 0 & 0\end{bmatrix}$.

The desired zero- placement polynomial matrix $\tilde{\mathbf{h}}$ changes according to equation (46).

3.1.4 Investigating the Influence of the Polynomial $\mathbf{T}$ on the Closed loop Performance Using Implicit Controller

In order to see the effect of the desired closed loop poles polynomial, on the closed loop system performance, the implicit PI pole-zero placement was switched on from $t=0$ to $t=800$ and the zero placement polynomial matrix $\tilde{\mathbf{H}}=[\mathbf{I}+\tilde{\mathbf{h}}]^{-1}(\mathbf{I}+\tilde{\mathbf{h}}_1 z^{-1})$ was fixed, whereas, the polynomial matrix $\mathbf{T}(z^{-1})=\mathbf{I}+\mathbf{T}_1 z^{-1}+\mathbf{T}_2 z^{-2}$ was changed three times as follows:

$$\begin{aligned}
&0\le t<250 && \mathbf{T}=\mathbf{I}+\begin{bmatrix}-.7 & 0\\ 0 & -1.65\end{bmatrix}z^{-1}+\begin{bmatrix}0 & 0\\ 0 & 0.7870\end{bmatrix}z^{-2}\\
&250\le t<550 && \mathbf{T}=\mathbf{I}+\begin{bmatrix}-0.8 & 0\\ 0 & -0.95\end{bmatrix}z^{-1}+\begin{bmatrix}0 & 0\\ 0 & 0\end{bmatrix}z^{-2}\\
&550\le t<800 && \mathbf{T}=\mathbf{I}+\begin{bmatrix}-1.85 & 0\\ 0 & -0.9\end{bmatrix}z^{-1}+\begin{bmatrix}0.887 & 0\\ 0 & 0\end{bmatrix}z^{-2}
\end{aligned} \tag{49}$$

The polynomial $\tilde{\mathbf{h}}$ is selected as follows:

$$\tilde{\mathbf{h}}(z^{-1})=\mathbf{I}+\begin{bmatrix}0.9 & 0\\ 0 & 0.9\end{bmatrix}z^{-1}$$

The outputs and the control inputs are shown in the Figures (9a) and (9b) respectively.
It clear from the Figures (9a) and (9b) that the performance of the closed loop system is affected by the changes in the polynomial $\mathbf{T}$.

3.2 Case study 2: Non-minimum Phase System

The implicit PID based pole-zero placement for MIMO systems proposed in section (2.2) is applied to the following MIMO plant, originally introduced by Prager and Wellstead (1980) and treated previously by Zayed et al.(2004) :

$$(\mathbf{I}+\mathbf{A}_1 z^{-1}+\mathbf{A}_2 z^{-2})\mathbf{y}(t)=z^{-1}(\mathbf{B}_0+\mathbf{B}_1 z^{-1})\mathbf{u}(t)+(\mathbf{I}+\mathbf{C}_1 z^{-1})\xi(t) \tag{50}$$

where:

$\mathbf{A}_1=\begin{bmatrix}-1.4 & -0.2\\ -0.1 & -0.9\end{bmatrix}$, $\mathbf{A}_2=\begin{bmatrix}0.48 & 0.1\\ 0 & 0.2\end{bmatrix}$, $\mathbf{B}_0=\begin{bmatrix}1 & 0\\ 0 & 0\end{bmatrix}$, $\mathbf{B}_1=\begin{bmatrix}1.5 & 1\\ 0 & 1\end{bmatrix}$, $\mathbf{C}_1=\begin{bmatrix}-0.5 & 0\\ 0.1 & -0.3\end{bmatrix}$

and $\xi(t)$ is a white noise vector sequence with zero mean and variance $\mathbf{R}'=\begin{bmatrix}0.1 & 0\\ 0 & 0.1\end{bmatrix}$.

Notice that the plant is a non-minimum phase system and also has different time delays in the two channels.

The set point $\mathbf{w}(t)$ changes every 100 samples as follows:

1. $w_1(t)$ changes from 5 to 10 and from 10 to 5.
2. $w_2(t)$ changes from 15 to 20 and from 20 to 15.

The user-defined gain and the pre-filter polynomials were respectively selected as:

$\mathbf{V}=\begin{bmatrix}V_1 & 0\\ 0 & V_2\end{bmatrix}=\begin{bmatrix}0.02 & 0\\ 0 & 0.04\end{bmatrix}$, $\mathbf{P}_d(z^{-1})=\mathbf{I}+\mathbf{P}_{d_1}z^{-1}$ and $\mathbf{P}_n(z^{-1})=\mathbf{I}+\mathbf{P}_{n_1}z^{-1}$

where $\mathbf{P}_{d_1}(z^{-1})=\begin{bmatrix}-0.5 & 0\\ 0 & -0.6\end{bmatrix}$ and $\mathbf{P}_{n_1}(z^{-1})=\begin{bmatrix}-0.4 & 0\\ 0 & -0.4\end{bmatrix}$.

It can clearly be seen from equations (13a), (15a), (15b) and (15c), that a PID controller is obtained if the polynomial matrix $\bar{\mathbf{F}}_s'$ is of second order. This can be achieved by selecting the pre-filter polynomial matrix $\mathbf{P}_d(z^{-1})$ to be of order one.

The desired closed loop poles and zeros polynomial matrices are respectively selected as follows:

$$\mathbf{T}=\mathbf{I}+\begin{bmatrix}-0.5 & 0\\ 0 & -0.6\end{bmatrix}z^{-1} \text{ and } \tilde{\mathbf{h}}=\mathbf{I}+\begin{bmatrix}0.8 & 0\\ 0 & 0.85\end{bmatrix}z^{-1}.$$

The outputs and the control inputs are, respectively, shown in Figures (10a) and (10b). We can see clearly from these Figures (10a) and (10b) that the excessive control actions resulting from set point changes are further reduced (i.e. more effectively tuned) when the new PID pole-zero placement controller is on line (during thesampling interval 251-550). Small oscillations can also be seen in the control inputs and closed loop system outputs during the last 250 samples (551-800 sampling times), where the conventional self-tuning PID is operating. The performance of the conventional PID controller can be further improved by adjusting the gain matrix V and the user defined polynomial matrices $\mathbf{P}_d$ and $\mathbf{P}_n$. However these tuning parameters must be selected using a trial and error procedure

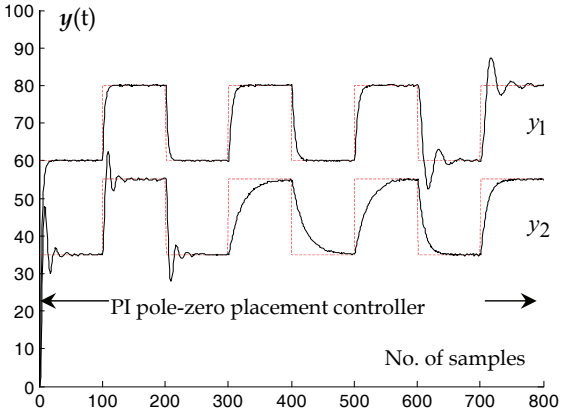


Fig. (9a). the outputs

$\mathbf{V} = \begin{bmatrix} 0.1 & 0 \\ 0 & 1.2 \end{bmatrix}$, $\tilde{\mathbf{h}}_1 = \begin{bmatrix} 0.9 & 0 \\ 0 & 0.8 \end{bmatrix}$ and $\tilde{\mathbf{h}}_2 = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$. The polynomial matrix $\mathbf{T}$ changes according to equation (49).

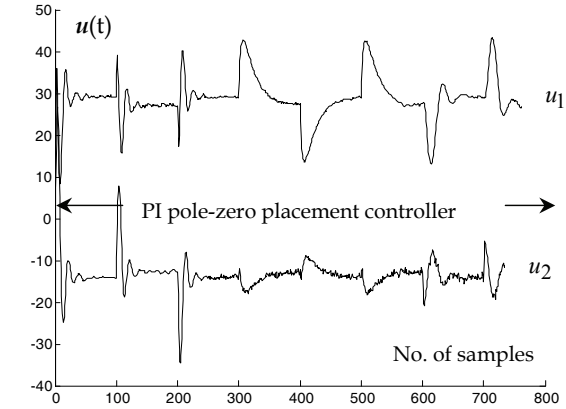


Fig. (9b). the control inputs

$\tilde{\mathbf{h}}_1 = \begin{bmatrix} 0.9 & 0 \\ 0 & 0.9 \end{bmatrix}$ and $\tilde{\mathbf{h}}_2 = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$. The polynomial matrix $\mathbf{T}$ changes according to equation (49).

–

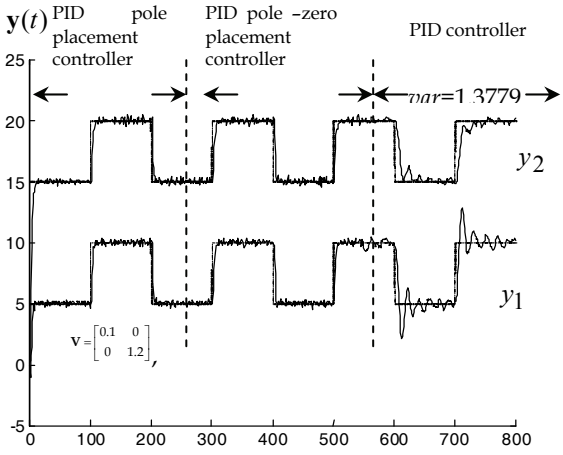


Fig. (10a). The outputs

$\mathbf{V} = \begin{bmatrix} 0.02 & 0 \\ 0 & 0.04 \end{bmatrix}$, $\mathbf{P}_{n_1} = \begin{bmatrix} -0.4 & 0 \\ 0 & -0.4 \end{bmatrix}$, $\mathbf{P}_{d_1} = \begin{bmatrix} -0.5 & 0 \\ 0 & -0.6 \end{bmatrix}$, $\mathbf{T}_1 = \begin{bmatrix} -0.5 & 0 \\ 0 & -0.6 \end{bmatrix}$ and $\mathbf{T}_2 = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$. $\tilde{\mathbf{h}}_1 = \begin{bmatrix} 0.8 & 0 \\ 0 & 0.85 \end{bmatrix}$ and $\tilde{\mathbf{h}}_2 = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$.

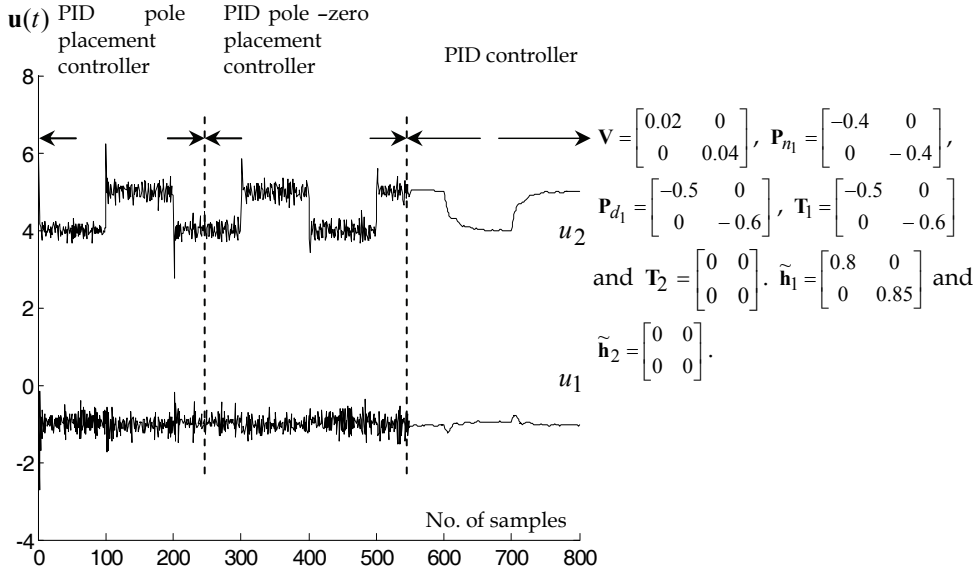


Fig. (10b). The control inputs

3.2.1 Investigating the Influence of the load disturbances on the Closed Loop Performance Using the Implicit Controller

The next task is to investigate the influence of the load disturbances on the closed loop system. Constant load disturbances of value $\begin{bmatrix}1\\1\end{bmatrix}$ were added to the outputs from the 350th instant to 800th sampling time instant. The two controller set points were both kept constant at values of 10 and 20 throughout.

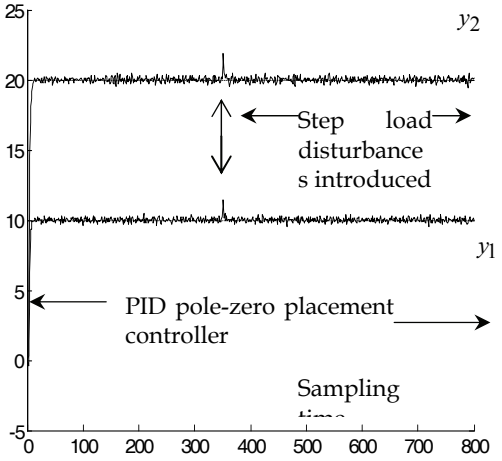


Fig. (11a). The outputs

The outputs and the control inputs for each of the three controller modes (namely PID pole-placement, PID pole-zero placement and the PID controller modes) are shown in the Figures (11a) to (11b) respectively.
It can clearly be seen from all the figures (11a) and (11b) that at steady state, the proposed PID based pole-zero placement controller has the ability to effectively regulate constant load disturbances to zero.

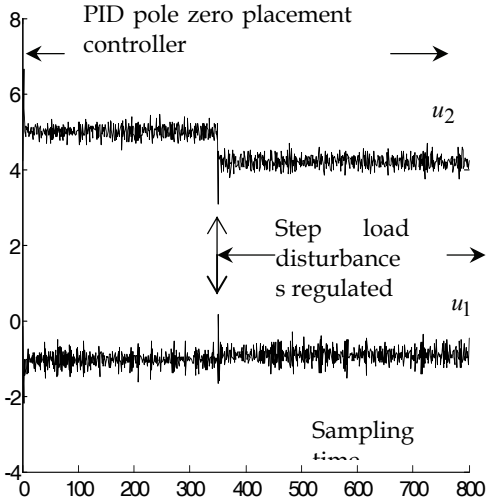


Fig. (11b). The control inputs

4. Conclusions

In this chapter, a new computationally efficient algorithm to incorporate the robustness of PID control and classical pole-zero placement into the generalised minimum variance stochastic self-tuning controller for multivariable systems has been proposed. The resulting PID self-tuning controller provides an adaptive mechanism, which ensures that the closed loop poles and zeros are located at their pre-specified positions. It is effectively an implicit algorithm in the sense that the controller design step is trivial (solving Diophantine equation at each sampling instant is avoided). Furthermore, the results presented here indicate that the controller tracks set point changes with the desired speed of response, penalises the excessive control action, and can deal with non-minimum phase systems. The transient response is shaped by the choice of the pole polynomial $\mathbf{T}(z^{-1})$, while the zeropolynomial $\tilde{\mathbf{H}}'(z^{-1})$ can be used to reduce the magnitude of control action or to achieve better set point tracking. In addition, the controller has the ability to ensure zero steady state error. Moreover, the controller is obtained as a left matrix-fraction and so can be immediately implemented. It is clear from sections (2.1), (2.2) and (3.3) that the proposed control design can be extended to a novel implicit multiple PID control. In this case the controller can then be operated in three modes, as either a conventional PID self-tuning controller, an implicit pole placement self-tuning control or a newly proposed implicit PID pole-zero placement

controller through the flick of a switch. The switching decision between the different PID controllers can be done manually or by using stochastic learning Automata.

5. References

Yusof R. & et al., (1994). Self-tuning PID control: a multivariable derivation and application, *Automatica*, *30*, pp. 1975-1981.

Yusof, R.; (1993). A multivariable self-tuning PID controller, *Int. J. Control, 57*, pp.1387-1403.

Tokuda M.; & Yamamoto T.; (2002) A neural-Net Based Controller supplementing a Multiloop PID Control System, *IEICE Trans. Fundamentals, Vol. E85-A*, 1, pp256-261.

Zayed, A. & et al., (2004). A New Multivariable Generalised Minimum-variance Stochastic Self-tuning with Pole-zero Placement, *International Journal of Control and Intelligent Systems*, *32* (1), pp2004, 35-44.

Sirisena H. & Teng F.,(1986). Multivariable pole-zero placement self-tuning controller, *Int. J. Systems Sci*., *17*(2), pp. 345-352.

Zhu Q., & et al., (2002)., A neural network enhanced generalised minimum variance self-tuning proportional, integral and derivative control algorithm for complex dynamic systems, *Journal of systems and Control Engineering, 216, part 1*, pp. 265-273.

Zayed A. & Hussain A., (2004). A New multivariable Non-linear Multiple-Controller Incorporating a Neural Network Learning Sub-model, *The first International Conference on Brain Inspired Cognitive Systems*, Stirling, Scotland, Uk., 29 Aug.-1 Sep., paper in CD.

Astrom K.., & Wittenmark B., (1973) On self-tuning regulators, *Automatica*, *9*, pp.185-199.

Prager D., & wellstead P., Multivariable pole-placement Self-tuning regulators, *Proc. Inst. Electr. Engineering*, *Part D*, *128*, 1980, 9-18.

Zayed A., (1997) *Minimum Variance Based Adaptive PID Control Design*, M.Phil Thesis, Industrial Control Centre, University of Strathclyde, Glasgow, U.K.

Zayed A., (2005) *Novel linear and nonlinear minimum variance control techniques for adaptive control engineering*, PhD Thesis , Department of computing science and mathematics, University of Stirling, U.K.

Zayed A.,& et al., (2006). A novel multiple-controller incorporating a radial basis function neural networks based generalised learning model, *Neurocomputing,69,* pp.1868-1881.

Yusof R.,& et al (1994), Temperature control of a water bath by self-tuning PI controller, *INT. J. SYSTEMS SCI*., *25*, pp1391-1404.

9

Nonlinear System Identification through Local Model Approaches: Partitioning Strategies and Parameter Estimation

Christoph Hametner and Stefan Jakubek
Vienna University of Technology
Austria

1. Introduction

In this chapter nonlinear system identification utilising Total Least Squares (TLS) and Generalised Total Least Squares (GTLS) methodologies in local model networks is addressed. These Neuro-Fuzzy networks are based on the identification of subdomains of the system that can reasonably accurately be described by local models. The aggregation of such subdomains in a so-called local model network then yields a versatile description of the overall system. One of the main challenges in the design of local model networks is the determination of the region of validity for the local models. Some of the related methodologies completely leave the partitioning to the user so that a solid idea about the nature of the nonlinearity of the system is required, cf. Johansen et al. (2000). Other methods make use of the input/output data of the system to identify suitable subdomains, Jakubek & Keuth (2006); Nelles (2002).
The identification and partitioning algorithm presented in this contribution is based on an iterative decomposition that works in the partition space rather than in the input- or product space. Thereby, in each step an axis-oblique partitioning is performed by multi-objective optimisation using an Expectation-Maximisation (EM) algorithm, Hametner & Jakubek (2007). The discrimination between input space and partition space allows the incorporation of prior knowledge into the partitioning process and reduces the complexity of the optimisation problem dramatically. In contrast to conventional clustering algorithms the number of rules is determined during the training by adding new models to the hierarchical tree until no statistically significant improvement is achieved.
The second important problem addressed in this work is the estimation of the local model parameters in the presence of noise in measured input *and* output data. In this situation conventional parameter estimation methods generally do not yield consistent estimates. Methods that are designed to cope with the case of noisy input and output channels are generally called "errors-in-variables methods", Soderstrom (2007). The corresponding block diagram is depicted in Fig. 1.
Various approaches for the identification of linear systems in a complex noisy environment can be found in literature, e. g. Han et al. (1996); Vandersteen (1998). A well-suited estimation procedure when all inputs are subject to noise is the Total Least Squares method (TLS). This method has been studied in many different areas over the last 25 years, e. g. Golub & Loan (1980); Huffel & Zha (1991); Markovsky et al. (2005). The main drawback of TLS is that

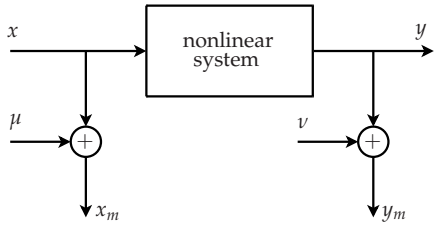


Fig. 1. Errors-in-variables model

strongly different noise levels easily result in ill-posed problems. A proper extension to the TLS method for such cases are Generalized Total Least Squares methods (GTLS). Generally, GTLS algorithms are methods for *linear* parameter estimation, when some or all inputs are subject to noise. There are various analyses and solution approaches to solve the GTLS problem, Huffel & Zha (1991); Nayak et al. (2006); Paige & Wei (1993); Van Huffel & J. (1989). For the integration of the above mentioned methodologies in a local model network weighted TLS and GTLS parameter estimation algorithms are presented in this work that allow for individual weighting of data records.
Recent local model network approaches utilise the prediction error for partitioning, Abonyi et al. (2002); Hametner & Jakubek (2007); Jakubek & Keuth (2006). If some or all signals involved in the parameter estimation process are corrupted by noise this approach is no longer feasible. In this paper a more general residual is defined to determine the region of validity of the local models. As a basis for the partitioning procedure mentioned above a suitable formulation of the *GTLS residual* will be introduced.
This chapter is organised as follows: Section 2 describes the architecture and the construction of the local model network. In section 3 weighted TLS and GTLS parameter estimation algorithms and the associated residual are presented. In the last section 4 the applicability and benefits of the proposed concepts are demonstrated by means of an illustrative example.

2. Construction of the local model network

Local model networks offer a versatile structure for the identification of nonlinear static and dynamic systems. The construction of these Neuro-Fuzzy networks is based on the identification of subdomains of the system that can reasonably accurate be described by local models. The aggregation of such subdomains then yields the description of the overall system.
The architecture of a dynamic local model networks is depicted in Fig. 2. The *physical inputs* are denoted by $\boldsymbol{u} = \begin{bmatrix} u_1 & u_2 & \dots & u_q \end{bmatrix}^T$ and the output by y, respectively. Each local model (indicated by subscript i) consists of two parts: The validity function Φ_i and its model parameters $\boldsymbol{\theta}_i$.
The *local* estimate for the output $y(k)$ is obtained by

$$\hat{y}_i(k) = \boldsymbol{x}^T(k)\boldsymbol{\theta}_i, \tag{1}$$

where $\boldsymbol{x}(k)$ denotes the input vector for the rule consequents at time k. For *dynamic* models typically a local affine model structure is implemented and $\boldsymbol{x}(k)$ contains past inputs and outputs and the offset term:

$$\boldsymbol{x}(k) = \begin{bmatrix} u_1(k-1) \\ \cdots \\ u_q(k-m) \\ y(k-1) \\ \cdots \\ y(k-n) \\ 1 \end{bmatrix}. \tag{2}$$

In (2) m and n denote the system order of the numerator and denominator respectively. For ease of demonstration the time delay of the system is neglected in (2). In case of *static* modelling the regressor vector $\boldsymbol{x}(k)$ may contain arbitrary functions (e. g. polynomials) of the elements of $\boldsymbol{u}$.

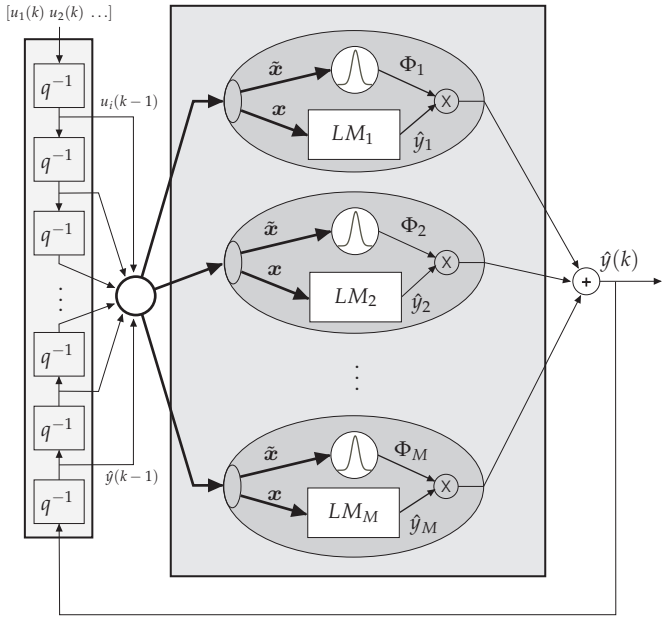


Fig. 2. Architecture of a dynamic local model network

All local estimations $\hat{y}_i(k)$ are used to form the global model output $\hat{y}(k)$ by weighted aggregation

$$\hat{y}(k) = \sum_{i=1}^{M} \Phi_i(k)\hat{y}_i(k), \tag{3}$$

where

$$\Phi_i(k) = \Phi_i(\tilde{\boldsymbol{x}}(k)) \quad , \quad \hat{y}_i(k) = \boldsymbol{x}^T(k)\boldsymbol{\theta}_i \tag{4}$$

and M denotes the number of local models.
Obviously, the input vector for the membership functions $\tilde{\boldsymbol{x}}(k)$ can be chosen differently to the input vector for the local model, cf. Fig. 2. The discrimination between input arguments of the

consequents and the premises is particularly useful if information about the structure of the nonlinearity is available, Nelles (2002). Especially in dynamic identification the dimension of the partition space and furthermore the complexity of the optimisation problem can be reduced dramatically. Therefore, the elements of the partition space $\tilde{\boldsymbol{x}}(k)$ are chosen according to the nonlinear behaviour of the system only.
The algorithm outlined here is based on the concept of decision trees, Breiman et al. (1984); Theodoridis & Koutroumbas (1999). The growing tree can be described by a binary tree where each node corresponds to a split of the partition space into two parts, e. g. Fig. 3. The free ends of the branches represent the actual local models with their validity functions Φ_i and the parameter vectors $\boldsymbol{\theta}_i$. The validity functions for the layout in Fig. 3 are obtained by

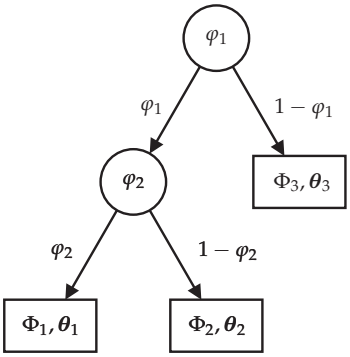


Fig. 3. Hierarchical discriminant tree

$$\Phi_1(\tilde{\boldsymbol{x}}) = \varphi_1(\tilde{\boldsymbol{x}})\varphi_2(\tilde{\boldsymbol{x}}), \tag{5}$$

$$\Phi_2(\tilde{\boldsymbol{x}}) = \varphi_1(\tilde{\boldsymbol{x}})(1-\varphi_2(\tilde{\boldsymbol{x}})), \tag{6}$$

$$\Phi_3(\tilde{\boldsymbol{x}}) = 1-\varphi_1(\tilde{\boldsymbol{x}}). \tag{7}$$

The partitioning of these new models is that a partition of unity ($\sum_{i=1}^{M} \Phi_i(\tilde{\boldsymbol{x}}) = 1$) throughout the partition space (for all $\tilde{\boldsymbol{x}}$) is guaranteed and no normalisation side effects like reactivations can occur cf. Nelles (2002).
The main challenge with local model networks is to determine Φ_i in such a way that the local estimate (1) sufficiently accurate describes the true process within the region of validity.
In each iteration step one local model of the growing tree is replaced by a new node and two new models attached to this node. The selection of the particular local model to be replaced is based on local accuracy which is assessed by the mean squared GTLS residual (see section 3.3). The partitioning of this new models is obtained considering a two-category classification problem. The main advantage of this concept is that with each iteration step the number of training data involved and thus the computational effort decreases.

2.1 Discriminant optimisation

In each step of the recursive decomposition of the partition space a nonlinear optimisation is necessary. Considering the two-category classification problem each data point has to be assigned to either class η_1 or η_2 depending on the value of the corresponding discriminants $\varphi_i(\tilde{\boldsymbol{x}}(k), \boldsymbol{\psi})$:

$$\varphi_1(\tilde{\boldsymbol{x}}(k), \boldsymbol{\psi}) = \frac{1}{1 + \exp(-\begin{bmatrix}1 & \tilde{\boldsymbol{x}}^T(k)\end{bmatrix}\boldsymbol{\psi})} \text{ for class } \eta_1 \tag{8}$$

$$\varphi_2(\tilde{\boldsymbol{x}}(k), \boldsymbol{\psi}) = 1 - \varphi_1(\tilde{\boldsymbol{x}}(k), \boldsymbol{\psi}) \text{ for class } \eta_2. \tag{9}$$

Here, $\boldsymbol{\psi} = \begin{bmatrix}\psi_0 & \psi_1 & \dots & \psi_p\end{bmatrix}^T$ denotes the weight vector. Each node of the model tree (see Fig. 3) corresponds to a logistic sigmoid discriminant function which is depicted in Fig. 4.

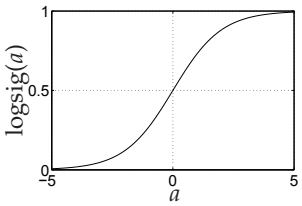


Fig. 4. Logistic sigmoid discriminant function

Since the input argument into (8) is linear in its input $\tilde{\boldsymbol{x}}(k)$ the decision boundary is also linear in the partition space, e. g. Fig. 5.

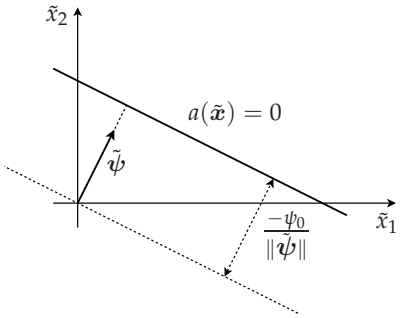


Fig. 5. Linear decision boundary of the nonlinear discriminant function

In order to find an optimal weight vector $\boldsymbol{\psi}$ for the two-category classification problem, an Expectation-Maximisation based algorithm is applied, e. g. Dempster et al. (1977). The goal of the iterative EM algorithm is to classify each data point in a way as to minimise the probability of misclassification. The EM algorithm proceeds in two steps (see also Hametner & Jakubek (2007)):

E-step: Based on the current estimate of the weight vector $\boldsymbol{\psi}_n$ the posterior probabilities $p(\eta_i|\boldsymbol{w}^T(k), \boldsymbol{\psi}_n)$ are calculated: The parameters of the two classes (local models) are

computed using weighted TLS or GTLS. For the evaluation of the model error a normal density function based on the (G)TLS residual, which is defined in section 3.3, is calculated. With the class-conditional residual distribution function (43), the posterior probabilities are obtained using Bayes' theorem.

M-step: The posterior probabilities are used to compute the new weight vector for the discriminant function and parameters of the two classes. The maximisation aims at determining the optimal weight parameters for the logistic discriminant function. Using e. g. a Levenberg-Marquardt learning algorithm the parameter vector $\boldsymbol{\psi}_n$ is adjusted such that the discriminant function (8) is optimally fit to $p(\eta_i|\boldsymbol{w}^T(k), \boldsymbol{\psi}_n)$.

Following, the steps of the iterative optimisation algorithm are reformulated:

2.1.1 Iterative optimisation algorithm

- *Initial choice* for $\boldsymbol{\psi}$: $\boldsymbol{\psi}_{ini}$ is ideally chosen such that the discriminant function bisects the data to be modeled. Thus the robustness to outliers is improved because the generation of two models with very unequal dimensions is avoided.
- *Step 1:* Compute the posterior probabilities $p(\eta_i|\boldsymbol{w}^T(k), \boldsymbol{\psi}_n)$.
- *Step 2:* Calculate/optimise the local model parameters $\boldsymbol{\theta}_{1,2}$ of the two classes by weighted (G)TLS.
- *Step 3:* Adjust $\boldsymbol{\psi}$ by Levenberg-Marquardt.
- *Repeat steps 1-3* until a certain termination criterion is reached, e. g. $\| \boldsymbol{\psi}_n - \boldsymbol{\psi}_{n-1} \|_2 < \epsilon$, where n denotes the iteration step and ϵ is a certain termination tolerance.

3. Local model parameter estimation

In many applications the inputs and outputs of a system are taken from measurements and are thus subject to noise. In this situation conventional parameter estimation methods suffer from the drawback that the parameter estimates are not consistent.
In this section weighted TLS and GTLS algorithms for the estimation of the local model parameters are presented. The application of Total Least Squares for parameter estimation from noisy inputs and outputs has been suggested repeatedly in recent years, Heij C. (1999); Markovsky et al. (2005); Roorda (1995). Demonstrative introductions and derivations of TLS are given in e. g. De Groen (1996); Golub & Loan (1980); Markovsky & Huffel (2007). While TLS is limited to situations where *all* channels are subject to noise, the GTLS algorithm yields consistent parameter estimates in the more general case when *some* (or all) channels are corrupted by noise, see also Huffel & Zha (1991); Nayak et al. (2006); Paige & Wei (1993); Van Huffel & J. (1989).
For the integration of the above mentioned methodologies in a local model network the algorithms presented in this section allow for individual weighting of data records.

3.1 Weighted Total Least Squares

The estimation of the local model parameters by Least Squares (LS) is based on the minimisation of the prediction error at the training data:

$$J = \frac{1}{2N} \sum_{k=1}^{N} [y(k) - \hat{y}(k)]^2. \tag{10}$$

Here N denotes the number of data records used for the local model.
In the case that only target data are affected by noise the minimisation of (10) yields a bias-free estimation of $\boldsymbol{\theta}_i$. Fig. 6 illustrates the problem by means of a simple linear map $y(x) = \beta_1 x + \beta_0$ where both x and y are corrupted by noise. For the parameter estimation only measured data $y_m = y + \nu$ and $x_m = x + \mu$ are available. For the ease of demonstration the noise variances $\sigma^2(\nu)$ and $\sigma^2(\mu)$ are assumed to be equal and uncorrelated. Fig. 6 compares the model obtained by minimisation of (10) to the true process.
In order to obtain a bias-free parameter estimation in the case of noisy inputs and outputs it is necessary to reconstruct both outputs *and* inputs. This means that instead of (10) the following criterion has to be minimised:

$$J = \frac{1}{2N} \left\{ \sum_{k=1}^{N} [x(k) - \hat{x}(k)]^2 + \sum_{k=1}^{N} [y(k) - \hat{y}(k)]^2 \right\} \tag{11}$$

Since (11) entails that both inputs and outputs have to be reconstructed the underlying optimisation is called *Total Least Squares* (TLS). From a geometric point of view the optimisation of (11) requires that the euclidean distances between data points and the model are minimised. Fig. 7 shows that this approach significantly improves the accuracy of the model. It can be shown that for $N \to \infty$ TLS delivers bias-free parameter estimates, cf. Heij C. (1999).

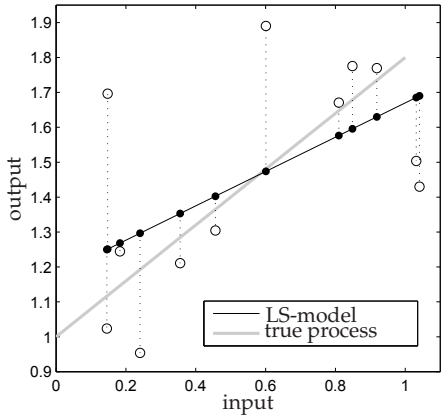


Fig. 6. Linear model, optimised by Least Squares

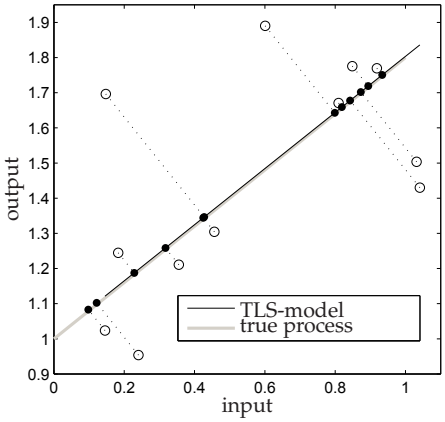


Fig. 7. Linear model, optimised by Total Least Squares

Following, an exemplary explanation of Total Least-Squares for a linear dynamic system is presented:
Let $\boldsymbol{X} \in \mathbb{R}^{N \times M}$ be the regressor matrix where every row contains all elements of $\boldsymbol{x}^T(k)$ except the constant offset term and let $\boldsymbol{y} \in \mathbb{R}^{N \times 1}$ be the observation vector. TLS aims at modifying both $\boldsymbol{y}$ and $\boldsymbol{X}$ in such a way that the following condition is satisfied:

$$\hat{\boldsymbol{y}} \in \text{Image}(\hat{\boldsymbol{X}}) \;\text{ and }\; || \boldsymbol{y} - \hat{\boldsymbol{y}} \mid \boldsymbol{X} - \hat{\boldsymbol{X}} ||_F = \min . \tag{12}$$

For the reconstructions $\hat{\boldsymbol{y}}, \hat{\boldsymbol{X}}$ a linear affine model structure is chosen. For that purpose an augmented regressor matrix is defined:

$$\boldsymbol{W} = [\, \boldsymbol{y} \mid \boldsymbol{X} \,] \tag{13}$$

A linear affine reconstruction of $\boldsymbol{w}^T$ is obtained from

$$\hat{\boldsymbol{w}}^T = \boldsymbol{w}^T - \left[(\boldsymbol{w}-\boldsymbol{m})^T\boldsymbol{b}\right]\boldsymbol{b}^T \;\text{ with } ||\boldsymbol{b}||_2 = 1 \tag{14}$$

Here $\boldsymbol{b}$ denotes the unit normal vector to the affine hyperplane and $\boldsymbol{m}$ is a point the hyperplane passes through. This linear model structure also ensures that $\hat{\boldsymbol{y}} \in \text{Image}(\hat{\boldsymbol{X}})$ holds. A reconstruction of all N data records yields

$$\hat{\boldsymbol{W}} = \boldsymbol{W} - \left[(\boldsymbol{W}-\mathbf{1}\boldsymbol{m}^T)\boldsymbol{b}\right]\boldsymbol{b}^T \tag{15}$$

with the $N \times 1$-vector $\mathbf{1} = [\,1\,,\,1\,,\,\ldots\,,\,1]^T$. The difference between original data and their reconstructions according to (12) is then given by

$$\boldsymbol{W} - \hat{\boldsymbol{W}} = \left[(\boldsymbol{W}-\mathbf{1}\boldsymbol{m}^T)\boldsymbol{b}\right]\boldsymbol{b}^T. \tag{16}$$

For *weighted* parameter estimation the estimation errors are weighted through the validity function Φ_i. Let $\boldsymbol{Q}_i$ denote a diagonal weighting matrix for the i-th local model. Its diagonal elements $q_i(k)$ represent the values of the validity function at the training data points. Compared to (11) a modified criterion is defined:

$$J_i = \frac{1}{2N}\left\{\sum_{k=1}^{N} q_i(k)[x(k)-\hat{x}(k)]^2 + \sum_{k=1}^{N} q_i(k)[y(k)-\hat{y}(k)]^2\right\} \tag{17}$$

Minimisation of (17) corresponds to a weighted version of TLS. Instead of the Frobenius norm in (12) one now has to minimise the following norm:

$$||\boldsymbol{Q}_i^{1/2}(\boldsymbol{W}-\hat{\boldsymbol{W}})||_F^2.$$

Since $||\boldsymbol{b}||_2 = 1$ the Frobenius norm becomes

$$||\boldsymbol{Q}_i^{1/2}(\boldsymbol{W}-\hat{\boldsymbol{W}})||_F^2 =$$

$$= \boldsymbol{b}^T(\boldsymbol{W}^T - \boldsymbol{m}\mathbf{1}^T)\boldsymbol{Q}_i(\boldsymbol{W}-\mathbf{1}\boldsymbol{m}^T)\boldsymbol{b}. \tag{18}$$

For the determination of $\boldsymbol{b}$ and $\boldsymbol{m}$ a weighted centroid vector $\boldsymbol{\mu}_W$ of all data records is defined:

$$\boldsymbol{\mu}_{Wi} = \boldsymbol{W}^T\boldsymbol{q}_i/s_q. \tag{19}$$

Here $\boldsymbol{q}_i$ denotes the vector composed from the main diagonal of $\boldsymbol{Q}_i$ and s_q is the sum of its elements: $s_q = \mathbf{1}^T\boldsymbol{q}_i$.
If all data records are referenced to the centroid (19) according to $\tilde{\boldsymbol{W}} = \boldsymbol{W} - \mathbf{1}\boldsymbol{\mu}_W^T$ and if one further observes that referencing to the centroid yields $\boldsymbol{q}_i^T\tilde{\boldsymbol{W}} = \mathbf{0}$ the minimisation of (18) yields

$$||\boldsymbol{Q}_i^{1/2}(\boldsymbol{W}-\hat{\boldsymbol{W}})||_F^2 = \boldsymbol{b}^T(\tilde{\boldsymbol{W}}^T\boldsymbol{Q}_i\tilde{\boldsymbol{W}})\boldsymbol{b} +$$

$$+ s_q[(\boldsymbol{m}-\boldsymbol{\mu}_{Wi})^T\boldsymbol{b}]^2 = \min. \tag{20}$$

The optimal choice for $\boldsymbol{m}$ is apparently $\boldsymbol{m} = \boldsymbol{\mu}_{Wi}$. The matrix $\tilde{\boldsymbol{W}}^T \boldsymbol{Q}_i \tilde{\boldsymbol{W}}$ is symmetric and positive semidefinite, the unit normal vector $\boldsymbol{b}$ is consequently obtained as the eigenvector associated to the smallest eigenvalue of $\tilde{\boldsymbol{W}}^T \boldsymbol{Q}_i \tilde{\boldsymbol{W}}$.
After partitioning of $\boldsymbol{b}$ according to $\boldsymbol{b} = [b_1\,,\,\boldsymbol{\beta}]^T$ TLS offers the following difference equation:

$$\hat{y}_i(k) = \frac{1}{b_1}\left[-\boldsymbol{x}^T(k)\boldsymbol{\beta} + \boldsymbol{m}^T\boldsymbol{b}\right]. \tag{21}$$

3.1.1 Decorrelation of identification data

In the minimisation of (11) it was assumed that all measurements in $\boldsymbol{w}^T$ are equally corrupted with noise and that the individual noise sources are uncorrelated. In practical applications these prerequisites are almost never fulfilled which can lead to a misinterpretation by the TLS parameter estimation concept, cf. Abonyi et al. (2002). In these cases the identification data must be decorrelated prior to parameter identification.
Let $\nu(k)$ denote the noise signal that is superimposed to the true output $y_0(k)$ and let $\mu(k)$ be the noise signal belonging to the true input $u_0(k)$. The data record $\boldsymbol{w}(k)$ is the obtained from its unperturbed equivalent $\boldsymbol{w}_0(k)$ from

$$\boldsymbol{w}^T(k) = \boldsymbol{w}_0^T(k) + \boldsymbol{n}^T(k),$$

with the noise vector

$$\begin{aligned}\boldsymbol{n}(k) = [\,&\nu(k)\,,\,\nu(k-1)\,,\,\ldots\\ |\,&\mu(k-d-1)\,,\,\mu(k-d-2)\,,\,\ldots]^T.\end{aligned} \tag{22}$$

The covariance matrix $\boldsymbol{R}_n = E\{\boldsymbol{n}(k)\boldsymbol{n}^T(k)\}$ contains all the above mentioned correlations and is assumed to be known. In practical applications it can be determined from data records $\tilde{\boldsymbol{W}}_s$ from steady-state phases according to

$$\boldsymbol{R}_n \approx \frac{1}{N-1}(\tilde{\boldsymbol{W}}_s^T\tilde{\boldsymbol{W}}_s). \tag{23}$$

For $N \to \infty$ the approximation (23) converges to the expectation $E\{\boldsymbol{n}(k)\boldsymbol{n}^T(k)\}$. In the practical tests conducted in connection with the presented method it turned out that (23) always led to good results. For a correct application of TLS the following statistical property must hold:

$$\frac{1}{N-1}E\{(\tilde{\boldsymbol{W}} - \boldsymbol{W}_0)^T(\tilde{\boldsymbol{W}} - \boldsymbol{W}_0)] = \boldsymbol{I},$$

where

$$\boldsymbol{W}_0 = \begin{bmatrix}\boldsymbol{w}_0^T(1)\\ \boldsymbol{w}_0^T(2)\\ \vdots\\ \boldsymbol{w}_0^T(N)\end{bmatrix}. \tag{24}$$

In order to accomplish this $\boldsymbol{b}$ is substituted by $\boldsymbol{b} = \boldsymbol{T}\tilde{\boldsymbol{b}}$. The new relevant noise covariance matrix then becomes

$$\tilde{\boldsymbol{R}} = \boldsymbol{T}^T\boldsymbol{R}_n\boldsymbol{T}. \tag{25}$$

A correct optimisation of $\tilde{\boldsymbol{b}}$ through TLS can thus be assured if the transformation matrix $\boldsymbol{T}$ is chosen such that

$$\boldsymbol{T}^T\boldsymbol{R}_n\boldsymbol{T} = \boldsymbol{I} \tag{26}$$

holds. The transformed unit normal vector $\tilde{\boldsymbol{b}}$ then turns out as the eigenvector belonging to the smallest eigenvalue of the matrix $\boldsymbol{T}^T\tilde{\boldsymbol{W}}^T\boldsymbol{Q}_i\tilde{\boldsymbol{W}}\boldsymbol{T}$.
In the case of pure measurement noise $\boldsymbol{R}_n$ is a diagonal matrix so that $\boldsymbol{T} = \boldsymbol{R}_n^{-1/2}$ is a solution to (26). Otherwise $\boldsymbol{T}$ can be obtained from the inverse of the Cholesky factorization of $\boldsymbol{R}_n$.

3.2 Weighted Generalised Total Least Squares

The GTLS algorithm described in this paper is based on the idea that conventional TLS inherently reconstructs all noisy inputs and the output.
As stated above the minimisation of (11) only leads to a consistent parameter estimate if all signals are corrupted by noise with equal variance and zero cross-correlations. If single variances are very different or if some signals are noise free the necessary decorrelation (see section 3.1.1) can easily result in an ill-posed problem. The GTLS algorithm overcomes these difficulties by excluding such signals from the reconstruction and treating them as noise-free. It is worth mentioning that the algorithm does not discriminate between inputs (regressors) and the output (observation). Consequently even the observation vector itself can be regarded as noise-free in the parameter estimation process.
As opposed to TLS, for GTLS parameter estimation noisy components in the augmented regressor $\boldsymbol{W}$ (see equation (13)) are indicated by a subscript "n" and noise-free components by an "o". Accordingly, $\boldsymbol{W}$ is partitioned into a part containing only noisy components $\boldsymbol{W}_n$ and another part containing only noise-free components $\boldsymbol{W}_o$:

$$\boldsymbol{W} = \begin{bmatrix}\boldsymbol{W}_n & \boldsymbol{W}_o\end{bmatrix}. \tag{27}$$

Without loss of generality, (27) is obtained from (13) by appropriately reordering its columns. In the sequel, $\boldsymbol{W}$ (and its estimates) always refer to the definition given by (27).
Similarly to TLS it is assumed that the noise signal $\boldsymbol{n}^T(k)$ that corrupts the rows of $\boldsymbol{W}_n$ is Gaussian with unity variance and zero cross-correlations. In most practical cases this has to be ensured by decorrelation of the noisy regressors, see section 3.1.1.
As opposed to TLS, GTLS only reconstructs noisy components:

$$\hat{\boldsymbol{W}} = \hat{\boldsymbol{W}}_n. \tag{28}$$

In the algorithm presented here, the reconstruction is based on ordinary TLS estimation where the original noisy data are projected on the TLS hyperplane using a reference point $\boldsymbol{m}$ and a unit normal vector $\boldsymbol{b}$, see equation (14). GTLS uses as an augmentation a linear combination of the noise-free regressors, given by $\boldsymbol{C}^T\boldsymbol{w}_o$. Here $\boldsymbol{w}_o^T$ denotes the noise-free regressor which is one row vector of $\boldsymbol{W}_o$ in (27).
The number of rows in $\boldsymbol{C}$ corresponds to the number of noise-free signals whereas the number of columns is the number of noisy signals.
Thus the GTLS reconstruction becomes

$$\hat{\boldsymbol{w}}_n^T(k) = \boldsymbol{w}_n^T(k) - \left[(\boldsymbol{w}_n(k) - \boldsymbol{m} - \boldsymbol{C}^T\boldsymbol{w}_o(k))^T\boldsymbol{b}\right]\boldsymbol{b}^T \tag{29}$$

and re-written for all data records ($k = 1, 2, \ldots N$)

$$\boldsymbol{W}_n - \hat{\boldsymbol{W}}_n = \left[(\boldsymbol{W}_n - \mathbf{1}\boldsymbol{m}^T - \boldsymbol{W}_o\boldsymbol{C})\boldsymbol{b}\right]\boldsymbol{b}^T, \tag{30}$$

with the $(N \times 1)$-vector $\mathbf{1} = \begin{bmatrix}1 & 1 & \ldots & 1\end{bmatrix}^T$. Optimisation of parameters yields

$$\|\boldsymbol{W}_n - \hat{\boldsymbol{W}}_n\|_F^2 = \min \quad \text{subject to} \quad \hat{\boldsymbol{y}} \in \text{Image}(\hat{\boldsymbol{X}}). \tag{31}$$

According to (31) the *weighted* optimisation of parameters yields

$$\|\boldsymbol{Q}_i^{1/2}(\boldsymbol{W}_n - \hat{\boldsymbol{W}}_n)\|_F^2 = \min \quad \text{subject to} \quad \hat{\boldsymbol{y}} \in \text{Image}(\hat{\boldsymbol{X}}), \tag{32}$$

with the weighting matrix $\boldsymbol{Q}_i$. In the sequel, the index i of the weighting matrix is dropped for the ease of notation. Evaluating (32) using (30) then yields

$$\begin{gathered}\|\boldsymbol{Q}^{1/2}(\boldsymbol{W}_n - \hat{\boldsymbol{W}}_n)\|_F^2 = \\ = \boldsymbol{b}^T(\boldsymbol{W}_n - \boldsymbol{1}\boldsymbol{m}^T - \boldsymbol{W}_o\boldsymbol{C})^T\boldsymbol{Q}(\boldsymbol{W}_n - \boldsymbol{1}\boldsymbol{m}^T - \boldsymbol{W}_o\boldsymbol{C})\boldsymbol{b}.\end{gathered} \tag{33}$$

Next, both $\boldsymbol{W}_n$ and $\boldsymbol{W}_o$ are referenced to their weighted centroids:

$$\boldsymbol{W}_n = \tilde{\boldsymbol{W}}_n + \boldsymbol{1}\boldsymbol{\mu}_n^T \tag{34}$$

with

$$\boldsymbol{\mu}_n^T = \frac{1}{s_q}\boldsymbol{q}^T\boldsymbol{W}_n. \tag{35}$$

and

$$\boldsymbol{W}_o = \tilde{\boldsymbol{W}}_o + \boldsymbol{1}\bar{\boldsymbol{w}}_o^T \tag{36}$$

with

$$\bar{\boldsymbol{w}}_o^T = \frac{1}{s_q}\boldsymbol{q}^T\boldsymbol{W}_o. \tag{37}$$

Thereby $\boldsymbol{q}$ is defined as the main diagonal of $\boldsymbol{Q}$ and s_q is the sum of all weights:

$$\boldsymbol{q} = \text{diag}(\boldsymbol{Q}), \quad s_q = \boldsymbol{1}^T\boldsymbol{q}.$$

This makes (33)

$$\begin{gathered}\|\boldsymbol{Q}^{1/2}(\boldsymbol{W}_n - \hat{\boldsymbol{W}}_n)\|_F^2 = \boldsymbol{b}^T\{\tilde{\boldsymbol{W}}_n^T\boldsymbol{Q}\tilde{\boldsymbol{W}}_n + s_q(\boldsymbol{\mu}_n - \boldsymbol{m})(\boldsymbol{\mu}_n - \boldsymbol{m})^T - \\ -2\boldsymbol{C}^T(\tilde{\boldsymbol{W}}_o^T\boldsymbol{Q}\tilde{\boldsymbol{W}}_n - s_q\bar{\boldsymbol{w}}_o\boldsymbol{m}^T + s_q\bar{\boldsymbol{w}}_o\boldsymbol{\mu}_n^T) + \boldsymbol{C}^T(\tilde{\boldsymbol{W}}_o^T\boldsymbol{Q}\tilde{\boldsymbol{W}}_o + s_q\bar{\boldsymbol{w}}_o\bar{\boldsymbol{w}}_o^T)\boldsymbol{C}\}\boldsymbol{b} = \min\end{gathered} \tag{38}$$

Collecting terms ($\boldsymbol{q}^T\tilde{\boldsymbol{W}}_o = \boldsymbol{0}$ and $\boldsymbol{q}^T\tilde{\boldsymbol{W}}_n = \boldsymbol{0}$) and using the abbreviation $\boldsymbol{c}_o = \boldsymbol{C}^T\bar{\boldsymbol{w}}_o$ leads to

$$\begin{gathered}\|\boldsymbol{Q}^{1/2}(\boldsymbol{W}_n - \hat{\boldsymbol{W}}_n)\|_F^2 = \boldsymbol{b}^T\{(\tilde{\boldsymbol{W}}_n - \tilde{\boldsymbol{W}}_o\boldsymbol{C})^T\boldsymbol{Q}(\tilde{\boldsymbol{W}}_n - \tilde{\boldsymbol{W}}_o\boldsymbol{C}) + \\ + s_q[(\boldsymbol{\mu}_n - \boldsymbol{m} - \boldsymbol{c}_o)(\boldsymbol{\mu}_n - \boldsymbol{m} - \boldsymbol{c}_o)^T]\}\boldsymbol{b} = \min\end{gathered} \tag{39}$$

The minimisation of (39) now has to be carried out with respect to $\boldsymbol{C}$, $\boldsymbol{m}$ and $\boldsymbol{b}$. The centroid $\boldsymbol{m}$ only appears in the second term of (39) which is a positive semidefinite expression. Therefore, $\boldsymbol{m}$ has to be chosen such that the second term in (39) vanishes:

$$\boldsymbol{m} = \boldsymbol{\mu}_n - \boldsymbol{C}^T\bar{\boldsymbol{w}}_o. \tag{40}$$

The first term in (39) is independent of $\boldsymbol{m}$ and has to be minimised separately. Its minimisation can be re-written as

$$\begin{gathered}\boldsymbol{b}^T(\tilde{\boldsymbol{W}}_n - \tilde{\boldsymbol{W}}_o\boldsymbol{C})^T\boldsymbol{Q}(\tilde{\boldsymbol{W}}_n - \tilde{\boldsymbol{W}}_o\boldsymbol{C})\boldsymbol{b} = \\ \|\boldsymbol{Q}^{1/2}(\tilde{\boldsymbol{W}}_n - \tilde{\boldsymbol{W}}_o\boldsymbol{C})\boldsymbol{b}\|_F^2 = \min_{\boldsymbol{C},\boldsymbol{b}}.\end{gathered}$$

The above Frobenius norm can be expanded in the following way:

$$\|\boldsymbol{Q}^{1/2}(\tilde{\boldsymbol{W}}_n - \tilde{\boldsymbol{W}}_o\boldsymbol{C})\boldsymbol{b}\|_F^2 =$$

$$\|\boldsymbol{Q}^{1/2}(\tilde{\boldsymbol{w}}_{n,1} - \tilde{\boldsymbol{W}}_o\boldsymbol{c}_1)b_1\|_2^2 + \|\boldsymbol{Q}^{1/2}(\tilde{\boldsymbol{w}}_{n,2} - \tilde{\boldsymbol{W}}_o\boldsymbol{c}_2)b_2\|_2^2 + \dots$$

where $\tilde{\boldsymbol{w}}_{n,j}$ and $\boldsymbol{c}_j$ denote the j-th column vectors of $\tilde{\boldsymbol{W}}_n$ and $\boldsymbol{C}$, respectively and b_j denotes the j-th element of the vector $\boldsymbol{b}$.
The minimisation of the Frobenius norm with respect to the two arguments $\boldsymbol{C}, \boldsymbol{b}$ can now be split into two subtasks:

1. The norms $\|\boldsymbol{Q}^{1/2}(\tilde{\boldsymbol{w}}_{n,j} - \tilde{\boldsymbol{W}}_o\boldsymbol{c}_j)\|_2^2$ of every single column vector have to be minimised which determines every column vector $\boldsymbol{c}_j$ of $\boldsymbol{C}$ independently. This can be summarised in one matrix operation:

$$Tr\{(\tilde{\boldsymbol{W}}_n - \tilde{\boldsymbol{W}}_o\boldsymbol{C})^T\boldsymbol{Q}(\tilde{\boldsymbol{W}}_n - \tilde{\boldsymbol{W}}_o\boldsymbol{C})\} = \min_{\boldsymbol{C}}.$$

$$\frac{\partial Tr}{\partial \boldsymbol{C}} = -2\tilde{\boldsymbol{W}}_o^T\boldsymbol{Q}(\tilde{\boldsymbol{W}}_n - \tilde{\boldsymbol{W}}_o\boldsymbol{C}) = 0 \tag{41}$$

$$\boldsymbol{C} = (\tilde{\boldsymbol{W}}_o^T\boldsymbol{Q}\tilde{\boldsymbol{W}}_o)^{-1}\tilde{\boldsymbol{W}}_o^T\boldsymbol{Q}\tilde{\boldsymbol{W}}_n \tag{42}$$

   Note that (42) is the solution of the overdetermined system $\tilde{\boldsymbol{W}}_o\boldsymbol{C} = \tilde{\boldsymbol{W}}_n$ where every row is weighted individually by the elements of $\boldsymbol{q}$.

2. Vector $\boldsymbol{b}$ becomes the unit eigenvector corresponding to the minimal eigenvalue of $(\tilde{\boldsymbol{W}}_n - \tilde{\boldsymbol{W}}_o\boldsymbol{C})^T\boldsymbol{Q}(\tilde{\boldsymbol{W}}_n - \tilde{\boldsymbol{W}}_o\boldsymbol{C})$.

Remark: The special case when only $y(k)$ is noisy leads to a partitioning $\boldsymbol{W}_n = [\boldsymbol{y}]$ and $\boldsymbol{W}_o = [\boldsymbol{X}]$ which essentially makes (42) a conventional LS solution. If, on the other hand all components are noisy then $\boldsymbol{C}$ vanishes and one obtains the TLS solution.

3.3 The GTLS residual

In this section a statistical criterion is derived that enables the statistical assessment of the residual error. This is an important prerequisite in a local model network with data-based partitioning. The GTLS residual is used to discriminate between unsystematic errors from measurement noise and systematic errors from truncation.
If the weights contained in $\boldsymbol{Q}_i$ are referred to a local model η_i a *class-conditional* GTLS-residual can be defined:

$$r(\boldsymbol{w}^T(k), \eta_i) = [\tilde{\boldsymbol{w}}_n^T(k) - \tilde{\boldsymbol{w}}_o^T(k)\boldsymbol{C}]\boldsymbol{b} \tag{43}$$

The argument η_i for r was instanced in order to emphasize that for the given model (indexed by i) all parameters ($\boldsymbol{b}$, $\boldsymbol{m}$ and $\boldsymbol{C}$) and consequently also the residual r essentially depend on the weight $q_i(k)$ at which every single training data record influences the parameters. Given the weights $q_i(k)$ the weighted GTLS parameter estimation (32) can be formally written as

$$\sum_{k=1}^{N} q_i(k)\, r(\boldsymbol{w}^T(i), \eta_i)\frac{\partial r}{\partial \boldsymbol{\theta}_i} = 0. \tag{44}$$

Note that for the special case that only $y(k)$ is noisy (43) simply reduces to the prediction error and (44) results in WLS parameter estimation.
If the noise signals in $\boldsymbol{W}_n$ are Gaussian with variance one and zero cross-correlations then the GTLS-residual r also follows a Gaussian distribution with zero mean and unit variance. Consequently, a *class-conditional residual distribution function* can be defined:

$$p(\boldsymbol{w}^T(k), \eta_i) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{r^2(\boldsymbol{w}^T(k), \eta_i)}{2}\right), \tag{45}$$

with r from (43).
The residual distribution function always refers to a linear model with parameters obtained from weighted GTLS with the weights chosen according to the class η_i.
It describes the probability density of a data record $(y(k), \boldsymbol{x}(k))$, belonging to model η_i and having a GTLS residual r. For an increasing number of data records used for training the empirical distribution of the residual should follow (45) if the model is actually linear or - more generally - if the data can be described by the chosen model structure. If there are nonlinearities that cannot be described by the model then the actual residual distribution will deviate from (45) in size and shape.
For illustration Fig. 8 compares the histograms of actual GTLS residual distributions to the ideal distribution (45). The upper figure shows an adverse situation which is caused by nonlinearities which cannot be properly modeled by the GTLS model structure (note the outlier at $r = 8$) whereas the lower figure shows a result, where the training data originate from a suitable (linear) process.

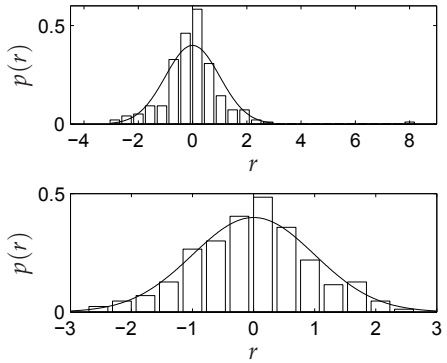


Fig. 8. Residual distribution for an adverse (top) and proper (bottom) model structure

4. Illustrative example

A simulation example is chosen to demonstrate the applicability and the benefits of the proposed concepts. For that purpose a nonlinear dynamic MISO system with three inputs was chosen from Treichl et al. (2002), where it is used to validate another MISO identification scheme. The block diagram is depicted in Fig. 9.
The transfer functions $F_1(z^{-1})$, $F_2(z^{-1})$ and $F_3(z^{-1})$ and the nonlinear static characteristics are chosen very similar to Treichl et al. (2002):

$$\begin{aligned} F_1(z^{-1}) &= \tfrac{0.1092z^{-1}+0.09552z^{-2}}{1-1.605z^{-1}+0.6703z^{-2}} \\ F_2(z^{-1}) &= \tfrac{0.2183z^{-1}+0.191z^{-2}}{1-1.605z^{-1}+0.6703z^{-2}} \\ F_3(z^{-1}) &= \tfrac{0.1092z^{-1}+0.09552z^{-2}}{1-1.605z^{-1}+0.6703z^{-2}} \end{aligned}$$

$$\begin{aligned} \mathcal{NL}_1: \quad v_1 &= -\tfrac{1}{2}u_1 + \tfrac{1}{2}u_1^2 - u_1^3 \\ \mathcal{NL}_2: \quad v_2 &= -\tfrac{1}{2}u_2 - u_2^2 + \tfrac{1}{2}u_3^3 \\ \mathcal{NL}_3: \quad v_3 &= \tfrac{1}{2} - \tfrac{1}{2}u_3 - u_2^3 \end{aligned}$$

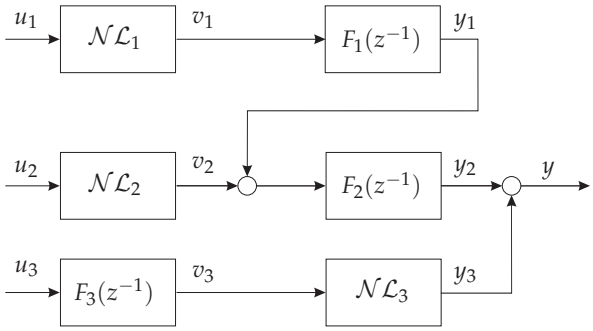


Fig. 9. Structure of the nonlinear MISO model

	LS		GTLS	
Data	R^2	MSE	R^2	MSE
identification	0.97924	0.13289	0.99193	0.051672
validation	0.97706	0.15699	0.99331	0.045785

Table 1. Simulation results

In this example, the input u_1 and the output are corrupted by Gaussian noise. For the excitation of the system, i. e. for the simulation of the training and generalisation data record, respectively, APRB-signals are selected. Their amplitudes and bandwidths are designed such that the whole static and dynamic operating range of the MISO system is covered.
In table 1 a comparison of the simulation results with LS and GTLS parameter estimation with eight local models is presented. It is clearly visible that the performance model with GTLS parameter estimation and partitioning based on the GTLS residual is considerably better than the LS model. This result is reflected in Fig. 10 where the autocorrelation function of the prediction error of the validation data record for LS and GTLS parameter estimates is depicted. The validation data record is chosen to be noise-free in this example to separate the model error from measurement error.
In Fig. 11 the simulation results with validation data are presented.

5. Conclusion

In this chapter the problem of noise in measured data in nonlinear system identification is addressed. First, a robust and efficient partitioning strategy using an EM algorithm is proposed. Second, weighted TLS and GTLS algorithms are presented which yield consistent parameter estimates of the local model parameters when some (GTLS) or all (TLS) input channels are turn out as special case of GTLS.
The GTLS residual is defined, which allows the statistical assessment of the residual error. This is an important prerequisite in a local model network with data-based partitioning.
The benefits of the proposed concepts are demonstrated by means of a simulation example. The performance of the resulting nonlinear model with local parameters estimated by

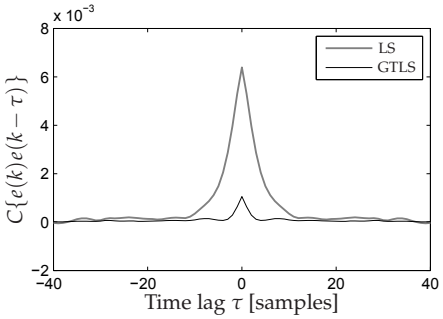


Fig. 10. Autocorrelation function of the prediction error for validation data with LS and GTLS parameter estimates, respectively.

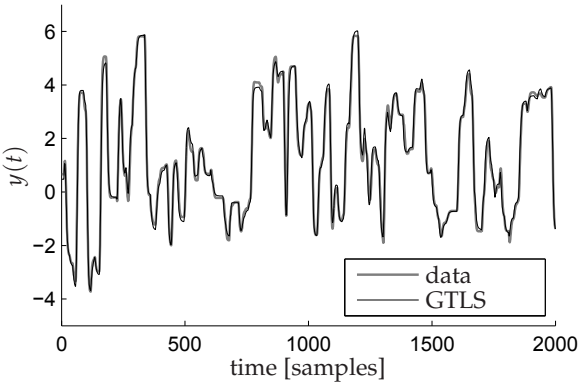


Fig. 11. Comparison of validation data to nonlinear models with GTLS parameter estimates

weighted GTLS is a product both of the parameter estimation itself and the associated residual used for the partitioning process.

6. References

Abonyi, J., Babuska, R. & Szeifert, F. (2002). Modified Gath-Geva Fuzzy Clustering for Identification of Takagi-Sugeno Fuzzy Models, *IEEE Transactions on Systems, Man and Cybernetics*, Vol. 32, IEEE, pp. 612–621.

Breiman, L., Friedman, J., Olshen, R. & Stone, C. (1984). *Classification and Regression Trees*, Wadsworth and Brooks, Monterey, CA.

De Groen, P. (1996). An introduction to total least squares, *Nieuw Archief Voor Wiskunde* **14**(2): 237–254.
URL: *http://www.citebase.org/cgi-bin/citations?id=oai:arXiv.org:math/9805076*

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977). Maximum Likelihood From Incomplete Data Via The EM Algorithm, *Journal of the Royal Statistical Society, Series B* **39**(1): 1–38.

Golub, G. & Loan, C. V. (1980). An analysis of the Total Least Squares problem, *SIAM J. Numer. Anal.* **17**(3): 883–893.

Hametner, C. & Jakubek, S. (2007). Neuro-fuzzy modelling using a logistic discriminant tree, *American Control Conference, 2007. ACC '07* pp. 864–869.

Han, S., Kim, J. & Sung, K. (1996). Extended Generalized Total Least Squares Method for the Identification of Bilinear Systems, *IEEE Transactions on Signal Processing*, Vol. 44, IEEE, pp. 1015 – 1018.

Heij C., S. W. (1999). Consistency Of System Identification By Global Total Least Squares, *Automatica* **35**: 993–1008.

Huffel, S. V. & Zha, H. (1991). The Restricted Total Least Squares Problem: Formulation, Algorithm, and Properties, *SIAM Journal on Matrix Analysis and Applications* **12**(2): 292–309. **URL:** *http://link.aip.org/link/?SML/12/292/1*

Jakubek, S. & Keuth, N. (2006). A Local Neuro-Fuzzy Network for High-Dimensional Models and Optimisation, *Engineering Applications of Artificial Intelligence* **19**: 705–717.

Johansen, T. A., Shorten, R. & Murray-Smith, R. (2000). On the Interpretation and Identification of Dynamic Takagi-Sugeno Fuzzy Models, *IEEE Transactions on Fuzzy Systems* **8**(3): 297–313.

Markovsky, I. & Huffel, S. V. (2007). Overview of total least-squares methods, *Signal Process.* **87**(10): 2283–2302.

Markovsky, I., Willems, J., Van Huffel, S., De Moor, B. & Pintelon, R. (2005). Application of Structured Total Least Squares for System Identification and Model Reduction, *IEEE Transactions On Automatic Control* **50**: 1490–1500.

Nayak, A., Trucco, E. & Thacker, N. A. (2006). When are Simple LS Estimators Enough? An Empirical Study of LS, TLS, and GTLS, *Int. J. Comput. Vision* **68**(2): 203–216.

Nelles, O. (2002). *Nonlinear System Identification*, 1st edn, Springer Verlag.

Paige, C. C. & Wei, M. (1993). Analysis of the generalized Total Least Squares Problem $AX = B$ when some columns of A are free of error, *Numerical Mathematics* **65**: 177–202.

Roorda, B. (1995). Algorithms for Global Total Least Squares Modelling of Finite Multivariable Time Series, *Automatica* **31**(3): 391 – 404.

Soderstrom, T. (2007). Errors-in-variables methods in system identification, *Automatica* **43**(6): 939–958.
URL: *http://www.sciencedirect.com/science/article/B6V21-4NH6NBT-2/2/74cb4630fabb1a3fbf568dbe52a8f8b6*

Theodoridis, S. & Koutroumbas, K. (1999). *Pattern Recognition*, Academic Press, San Diego, CA, USA.

Treichl, T., Hofmann, S. & Schroder, D. (2002). Identification of nonlinear dynamic MISO systems with orthonormal base function models, *Proceedings of the 2002 IEEE International Symposium on Industrial Electronics*, Vol. 1, pp. 337– 342.

Van Huffel, S. & J., V. (1989). Analysis and Propoerties of the generalized Total Least Squares Problem $AX = B$ when some or all columns of A are subject to Error, *SIAM Journal on Matrix Analysis and Applications* **10**(3): 294 – 315.

Vandersteen, G. (1998). On the Use of Compensated Total Least Squares in System Identification, *IEEE Transactions on Automatic Control*, Vol. 43, IEEE, pp. 1436 – 1441.

10

Utilising Virtual Environments To Research Ways To Improve Manipulation Task

Faieza Abdul Aziz, On Chee Leong and Lai Jian Ming
Department of Mechanical and Manufacturing Engineering, Universiti Putra Malaysia, Malaysia

Abstract

Real-time feedback systems has been widely used and become very important in many fields. In this project Microsoft Visual C++ together with OpenGL programming software were employed to create a Tower of Hanoi which was used as the experiment task. The real-time system has been studied by adding a real-time visual feedback into a simulation task. Two types of real-time visual feedback were discussed in this work, which were colour feedback and text feedback. Visual feedback techniques were design to give the cues to the users about the reasons for errors occurrence. Four different methods were compared and contrasted which are no feedback, with feedback, with text feedback and with colour feedback. The Tower of Hanoi is also programmed to provide different feedback in real time for the purpose of investigating the effect of auditory feedback to the user. Moreover, the Tower of Hanoi is programmed in stereoscopic for virtual reality manipulation task. Three types of feedback were evaluated. It consists of non-speech, speech auditory feedback and without feedback. The goals of this study were to explore real-time visual feedback technique and compare the effectiveness of different types of real-time visual feedback technique. For the purpose of investigating the effect of auditory feedback to the user, the result of the project showed the participant performance in solving the Tower of Hanoi is better in the non-speech auditory feedback. Beside that, speech auditory feedback provides greatest understanding to the user throughout the experiment, but the drawback is the participant cannot complete the task in shorter time.

1. Introduction

Nowadays, real-time feedback application is become more important in many fields such as communication products, consumer appliances, telephone switch and others (Michael, 1992). Real-time feedback from a system is crucial to determine the success of a process or activity. The meaning of real time refers to the immediate response of the system. The term is used to describe a number of different computer features. Real-time system is a computer-based system which can resolve a number of difficult issues simultaneously with rapid response (Philip, 2004). For example, real-time operating systems are systems that respond to input immediately. They are used for such tasks as navigation, in which the computer

must react to a steady flow of new information without interruption. Most general-purpose operating systems are not real-time because they normally take a few seconds, or even minutes, to react. It has become a common practice to use digital computer for the real-time application such as computer-integrated manufacturing (CIM), industrial process control, defenced systems and, electric power distribution and monitoring. All these applications usually require getting the response from operating system instantly or as fast as possible (Sang, 1995). Real-time feedback application allows the user can straight away know the error while any mistake of input has been given to the real-time feedback system. Instead of telling the error to the user, real-time feedback system also can react as the medium for training the new user to become familiar with new system and to train new user not to make any mistake that may caused error to the system.

Virtual reality system is used to create a virtual world for the user for various applications in training users by giving real-time feedback to the users. Flight simulators for training airplane pilots and astronauts were the first form of this technology, which provided a very realistic and very expensive simulation (www.techweb.com). Lai et al. (Lai et al., 2008) conducted a study in solving the task of "Tower of Hanoi" by adding auditory feedback and the results showed that the users improved their performance and completed the manipulation task in shorter time.

A number of studies have shown that audio contributed to the interaction process in order to provide richer, more robust environment than with mere graphic feedback (Heeter and Gomes, 1992). Auditory feedback can present further information when the bandwidth of graphic information has been exhausted, as is often the case with the present emphasis on graphic presentation (Rauterberg, 1999). By expanding conventional interfaces in another dimension, sounds make tasks easier and more productive. Other studies have even shown certain types of information to be represented better by sound than through graphics or text. Additionally, audio feedback may complement graphics and text to create valuable redundancy, reinforcing or reconfirming a concept in the user's mind (Winberg and Bowers, 2004).

In this work, a study on a real-time visual feedback manipulation task as well as real time auditory feedback technique in manipulation task will be carried out to solve three discs Tower of Hanoi, invented by the French mathematician Edouard Lucas in 1883. The user needs to move all the discs from the left peg to the right peg. The rules for the task are only one disc can be moved at one time and the larger disc may not be placed on top of the smaller disc. The objective of this work is to create the task of 3 discs 'Tower of Hanoi' using OpenGL together with Microsoft Visual C++, apply real time visual feedback and investigate the effectiveness on various type of real time visual feedback technique for solving the task of Tower of Hanoi. The best method of real-time visual feedback in simulation task will be evaluated. A group of user has been chosen and test on the "Tower Of Hanoi" game that has different type of visual feedback such as "colour" or "text" feedback and the time to complete the task has been plotted and analysed. The second objective is to investigate the effect on real time auditory feedback technique and solving the Tower of Hanoi manipulation task in virtual reality. Three auditory feedback techniques will be tested which include task with speech auditory feedback, task with non-speech auditory feedback and task without auditory feedback.

2. Methodology

Since there will be two different type of feedbacks being investigated in this work, the scope of the methodology taken to perform those tasks are divided into two parts. The first part explained the procedures taken to perform a real-time visual feedback manipulation task to solve the 'Tower Of Hanoi' problem. The following part explained on the methods taken to solve the Tower of Hanoi with real time auditory feedback technique. Finally, the Tower of Hanoi manipulation task is solved in virtual reality using visual and auditory feedbacks. The OpenGL has been utilized in programming Tower of Hanoi to be solved using visual and auditory feedback.

2.1 Tower of Hanoi OpenGL programming

The Towers of Hanoi utilized many features and special functions found in the three OpenGL libraries, enabling endless possibilities. In this project, the author is required to combine these elements into a single application (Segal and Akeley):

• Display Lists: were used to quickly and efficiently render the pole objects and disc objects. The disc objects were created by glut library toruses, which being a solid torus. This created a specific visual effect desired by the programmer. The poles were solid cone, transformed and placed within the display list to automatically come in the correct size and object placement when initialized requiring a programmer to merely place a pole object.

• Reshape change subcomponents: Handles the rendering of the viewing space, and the reshaping of the window.

• Transformation Functions: The program relied heavily on the Translate and Scale functions to accurately place and sizes the discs, as the discs themselves were at the same size and location at initialization. Rotation was used to properly orient the pole cube to the base cube of the pole object.

• Geometric Objects: Advanced geometric objects such as, glutSolidTorus(), and glutSolidCone() were used to render the objects in the scene.

• Lighting: Many material and lighting functions of OpenGL was used to create the look of the Towers of Hanoi. The materials of the scene were given ambient, diffuse, specular and shininess settings to make the solid objects in the scene appear shiny yet dull to their reflective properties.

• Keyboard Function: Keyboard was used to handle the various keyboard function and data input.

• Mouse Function: The mouse was used for simple controls and interface for the user where the users are able to move the disk to their desired location to complete the task that is to solve the tower of Hanoi.

2.2 Real-time visual feedbacks to solve Tower of Hanoi

The scope of the methodology applied for a study on a real-time visual feedback manipulation task is mainly divided into several stages. The author needs to program the Tower of Hanoi using OpenGL together with Microsoft Visual C++ with visual feedback which will be used as the experiment task. Then, the author needs to conduct the experiment on different visual feedback technique to investigate the performance of solving the task of Tower of Hanoi.

Each of the real-time feedback system has the ability to let users to input the data before it calculates the available results within an interval time. Thus, in this stage, the ability of the manipulation task to receive the incoming data has been tested before interpret it. At the beginning of this stage, the author tried to give some input by using keyboard. The input can be the steps of completing the manipulation task, the hints to complete the manipulation task. Then, the author also enabled the ability of manipulation task to get the input by clicking or dragging by using mouse.
The author tried to put some warning message when a user (a person who manipulates the task) has keyed in the illegal move that is not allowed by the rules for the task. Later in advance, the author also tried to put some animation when the errors occurred.

2.3 Real time auditory feedback to solve Tower of Hanoi

The scope of the methodology applied for technique in manipulation task is mainly divided into four major stages. Firstly, the OpenGL programming language has been explored and tested by practicing various examples available form the internet and also programmed by the authors. The code has been programmed specifically to be used in solving the Tower of Hanoi problem and enhance the understanding of auditory feedback and virtual reality. Secondly, to provide auditory feedback to enhance the manipulation task. Thirdly, to conduct the experiment on different auditory feedback technique and investigated the performance of solving the task as well as the experiment done in virtual reality for Tower of Hanoi manipulation task. Finally, analyze the result and identify the cause and effect for the different feedback. Here are the detailed steps taken to perform the auditory feedback simulation.

2.3.1. Tower of Hanoi feedback simulation

This experiment examined whether different auditory feedback techniques may affect the participants performance in solving the task. Three conditions were programmed separately as the experiment tasks, which were the Tower of Hanoi without auditory feedback, Tower of Hanoi with non-speech auditory feedback and the Tower of Hanoi with speech auditory feedback. The auditory feedback will be provided whenever there is any mouse-clicking occurred, error movement, complete or fail to solve the task in a given time and pre-start game sound as illustrated below:

2.3.1.1 Start game

The game starting sound will be provided when the user press ‘S’ to start the game.

2.3.1.2 Mouse clicking

The mouse-clicking sound will be provided when the user click to move and place the disc on the pole (as shown in Figure 1).

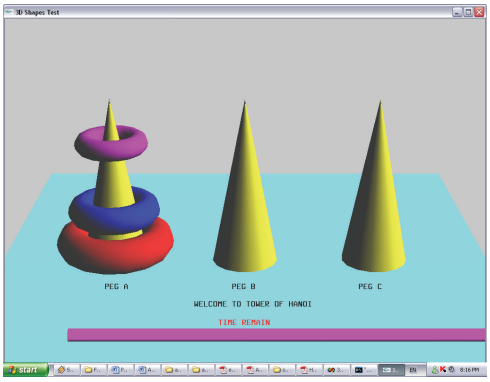


Fig. 1. Mouse-clicking sound

2.3.1.3 Error movement

The sound feedback will be provided when the error movement was made, such error include:

• Put the bigger disc on top of the smaller disc (as indicated in Figure 2).

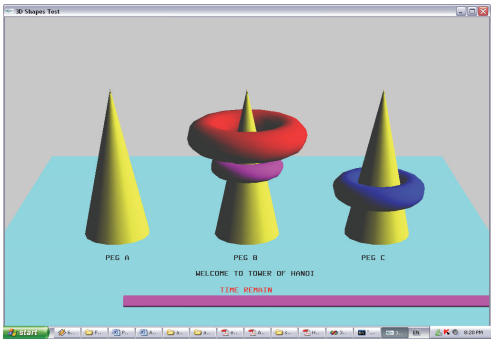


Fig. 2. Sound for error movement

• Did not put the disc at a proper location (as indicated in Figure 3).

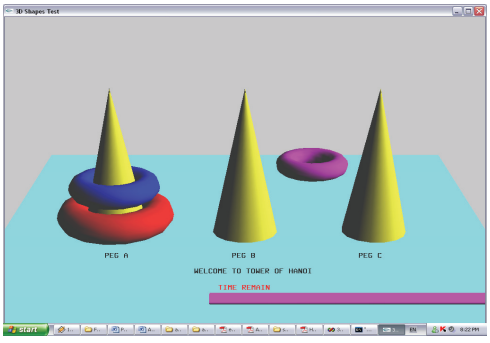


Fig. 3. Improper location

• Try to move a disc in the position to the other pole (as indicated in Figure 4).

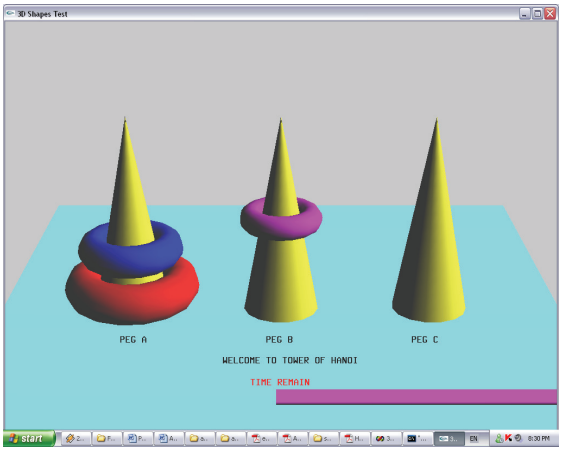


Fig. 4. Putting in the other pole

2.3.1.4 Game over

The sound feedback will be provided when the times is run out. It indicates that the game is over (as shown in Figure 5).

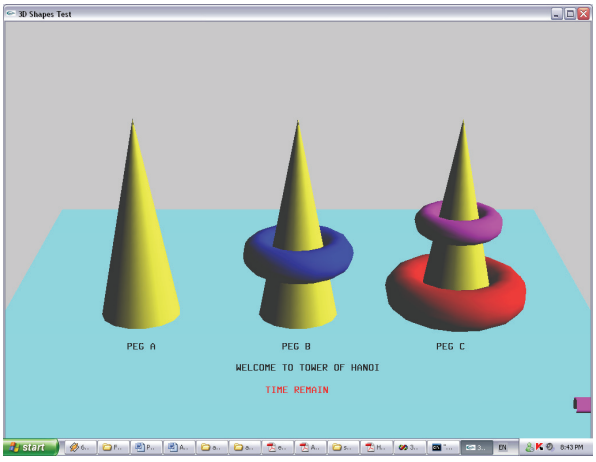


Fig. 5. Game over

2.4 Virtual reality for Tower of Hanoi

An experiment will be carried out in virtual reality for the Tower of Hanoi manipulation task at Virtual Reality Laboratory, Faculty of Computer Science in University Malaya.

3. Experiment

Experiments were conducted to evaluate the effectiveness of real time feedbacks in solving the Tower of Hanoi simulation task. The first part of the experiment evaluated the effectiveness of real-time visual feedback system while the second part of the experiment studied on the effect of auditory feedback technique in solving the Tower of Hanoi. In the first part of the experiment, there are two types of visual feedbacks being experimented. They were simulation with real-time visual feedback to be compared with simulation without real-time feedback and simulation with real-time text feedback which is compared to simulation without real-time feedback. The second part of the experiment involved three different sections. They are Tower of Hanoi with speech auditory feedback, Tower of Hanoi with non-speech auditory feedback and Tower of Hanoi without auditory feedback. The results and discussion for all these experiments are presented in the following section.

3.1 Effectiveness of real-time visual feedback system

An experiment was carried out to investigate the effect of visual feedback technique. The experiment result will determine whether the visual feedback from one system is encouraging or giving a negative effect on the user performance.

3.1.1 Participants

Sixteen males and four females between the ages of 23 and 24, with the mean age of 23.5 years old with a standard deviation of 0.5 years old have participated in this study. The entire users have been told about the goal of the game. The group of participant has been divided into 2, where group A is for the evaluation on the effective of occurrence of real-time visual feedback while group B is for the evaluation on the effectiveness of the type of real-time visual feedback. None of the subjects had an experience on manipulating the "Tower of Hanoi" puzzle.

3.1.2 Experimental procedure

Each of the experiment has been conducted separately by each user and each user has the time limit of 100 seconds to complete the task for the "3 discs Tower of Hanoi". Before they start to do the experiment, the user has been briefly taught by the author on the goal of the simulated task. The goal for the task is to move all the discs from peg A to peg C.

Before the experiment begin, the users have been divided into 2 group where group A is doing the evaluation on the effective of occurrence of real-time visual feedback and group B is for the evaluation on the effectiveness of the type of real-time visual feedback. The users need to key in their information and the data has been save automatically in a notepad as .txt file. Then, the experiment of "3 discs Tower of Hanoi" has been conducted by the user. The user must finish the task within the time frame given for the experiments. After they finish the task, the time, the steps and the error steps in completing the task have been save automatically in the notepad as .txt file.

3.1.3 Results and Discussion

The results for each types of experiment have been saved in different file name. Figure 6 shows the graphic that created from the programming code. This is "Tower Of Hanoi" with

3 different colour discs which are magentas, blue and red. The bar below the peg shows the time indicated for the simulation task.

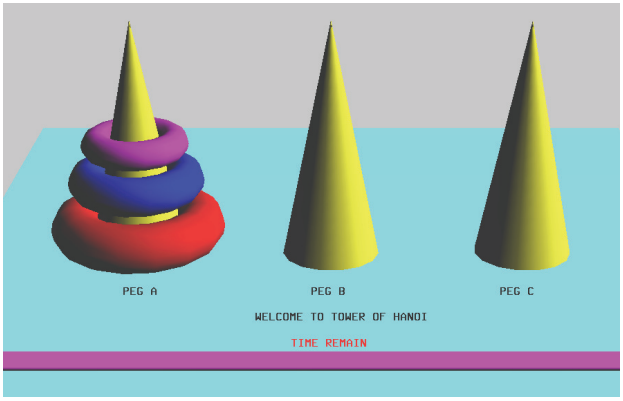


Fig. 6. Simulation task on "3 discs tower of Hanoi"

Table 1 showed the results for the real-time visual feedback. Table 2 showed the results for the simulation without real-time visual feedback. Table 3 showed the results for the simulation real-time text feedback. Table 4 shows the results for simulation with real-time colour feedback.

Participant	1	2	3	4	5	Average
Task Completion Time(s)	18.98	24.66	41.47	15.72	56.83	31.53
Number of steps	7	13	8	7	12	9.4
Error Steps	3	1	3	0	1	1.6

Table 1. Results for simulation with real-time visual feedback

Participant	1	2	3	4	5	Average
Task Completion Time(s)	23.56	13.42	23.42	21.32	56.83	22.88
Number of steps	10	8	13	9	14	10.8
Error Steps	1	0	2	4	3	2

Table 2. Results for simulation without real-time feedback

Participant	1	2	3	4	5	Average
Task Completion Time(s)	7.42	32.41	23.89	45.97	16.52	25.24
Number of steps	7	18	15	20	8	13.6
Error Steps	0	3	2	2	1	1.6

Table 3. Results for simulation with real-time text feedback

Participant	1	2	3	4	5	Average
Task Completion Time(s)	28.9	12.88	15.32	8.92	25.21	18.24
Number of steps	16	25	8	7	32	17.6
Error Steps	3	10	2	0	7	4.4

Table 4. Results for simulation with real-time colour feedback

Figure 7, 8 and 9 showed the results on task completion time, number of steps and error steps on 4 simulation tasks which are simulation task with real-time visual feedback, simulation task without real-time visual feedback, simulation task with real-time text feedback and simulation task with real-time colour feedback.
Average task completion time for all the experiments conducted is shown in figure 7. The experiment with colour feedback showed the shortest time, followed by experiment without feedback, experiment with text feedback and finally experiment with feedback.

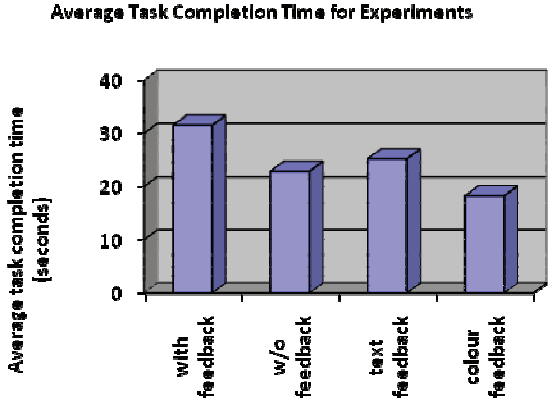


Fig. 7. Average Task Completion Time

Average number of steps for all the experiments conducted is shown in figure 8. The experiment with feedback showed the shortest time, followed by experiment without feedback, experiment with text feedback and finally experiment with colour feedback.

Average number of error steps for all the experiments conducted is shown in figure 9. The experiment with text feedback showed the shortest time, followed by experiment with feedback, experiment without feedback and finally experiment with colour feedback.

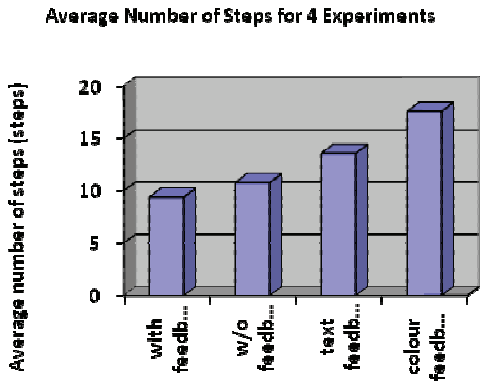


Fig. 8. Average number of steps

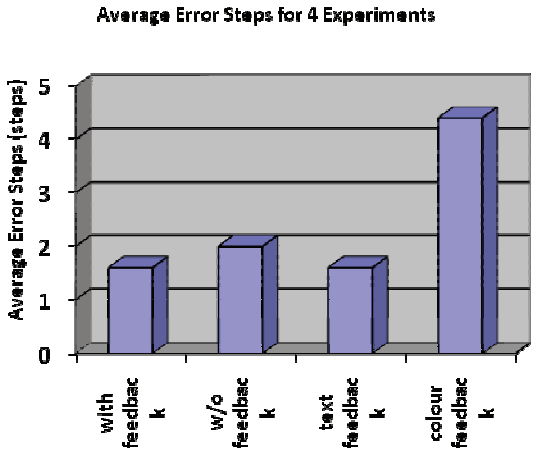


Fig. 9. Average error steps

The average task completion time for the simulation task with real-time visual feedback is slightly longer than the simulation task without real-time visual feedback. However, the average error steps for the simulation task with real-time visual feedback are lower than the simulation task without real-time visual feedback.

This is because while the visual feedback is added to the simulation, the user need sometime to figure out what is the error steps that has been make through out the text of error that shown on the screen as in figure 10. After knowing the reasons for the errors, the user will avoid doing the same errors and the errors steps for them are lower when comparing to the

user that using the simulation task without visual feedback. The main reason is because they did not know the reasons of the errors.
As a conclusion, the simulation with real-time visual feedback is better than the simulation without real-time visual feedback. The simulation with real-time visual feedback has taught user knowing the reasons of errors and do not repeat the errors.

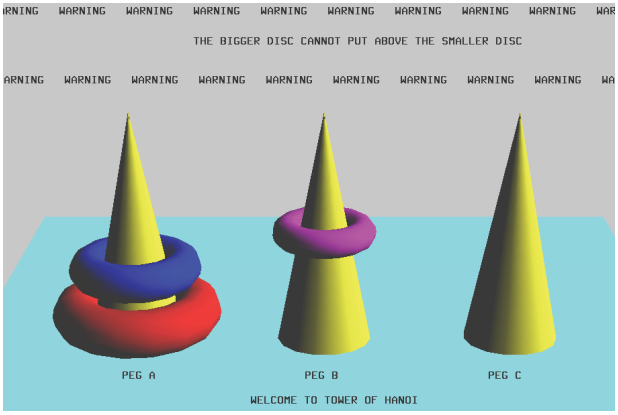


Fig. 10. Figure show the text notice for errors.

Next experiment has been carried out to evaluate the best method of real-time visual feedback. These experiments were conducted to evaluate whether a system with real-time text feedback is better than system with real-time colour feedback or vice versa. From the experiment, we can see that the average task completion time for the simulation task with real-time text feedback is slightly longer than the simulation task with real-time colour feedback. However, the average error steps for the simulation task with real-time text feedback are lower than the simulation task with real-time colour feedback.
This is because while the text feedback is added to the simulation, the user need sometime to read the text and figure out what is the reasons for error steps that has been made through out the text of error that shown on the screen as in figure 5 while for the colour feedback, the user only know the error has occurred but they may be do not know what causes of the error and does not need time to read or understand the reasons of an error step. After knowing the reasons for that particular error, the user will avoid for doing the same errors again, and the errors steps for them will become lower than the user that using the simulation task without visual feedback.
It has been found that the simulation with real-time text feedback is better than the simulation with real-time colour feedback. The simulation with real-time text feedback has taught the user understand the reasons of errors and do not repeat the errors.

3.2 Effectiveness of real time auditory feedback

An experiment will be carried out to investigate the effect of auditory feedback technique. The experiment result will determine whether the auditory feedback from one system is encouraging or giving a negative effect on the user performance. Fifteen participants were instructed to solve the three discs Tower of Hanoi without auditory feedback and the Tower

of Hanoi with non-speech and speech auditory feedback created by the programmer using the OpenGL software. The performance in solving the task will be analyzed.

3.2.1 Participant

There were 15 participant from the age group 23-25 had been instructed to solve the Tower of Hanoi with speech/non-speech and without auditory feedback. None of them have experienced to solve the task before and suffering any hearing illness.

3.2.2 Material

The Tower of Hanoi created using Microsoft Visual C++ together with OpenGL is used as the experiment tasks. The program used was divided with speech/non-speech and without auditory feedback to provide real time feedback to the participant during the process for solving the task.

The computer mouse and keyboard were used as input devices to move the disc and keyed in the participants' personal detail. The output device such as CRT or LCD monitor used for graphic display, and a speaker is used as a sound output device.

3.2.3 Procedure

All participants are required to complete a questionnaire for general information, read and sign a health consent form prior Tower of Hanoi solving experience.

Five participants were instructed to solve the 3 disc Tower of Hanoi with speech auditory feedback and five participants were instructed to solve the 3 discs Tower of Hanoi with non-speech auditory feedback. Another five participants were instructed to solve the 3 discs Tower of Hanoi without auditory feedback.

3.2.4 Measure

In this project, the task solving performance will be measure where the time for the participants for solving the task will be recorded for all cases with speech/non-speech and without auditory feedback and the number of error step made by the participants on the process of solving the task is recorded.

3.2.5 Results and Discussion

- Participant

The Tower of Hanoi experiments were performed by 15 adult participants, with a mean age of 24.2 years. The participants were in good health with no history of hearing illness. None of the participants has experienced in solving Tower of Hanoi problem. Trials were carried out by each participant individually.

- Task solving performance

The task solving times and the number of error made by the participant in the experiment is recorded for three types of auditory feedback technique which is Tower of Hanoi with speech auditory feedback as in Table 1, Tower of Hanoi with non-speech auditory feedback as in Table 2 and Tower of Hanoi without auditory feedback as in Table 3. Beside that, a graph is plotted for the comparison between these three different auditory feedbacks technique in term of task solving time as shown in Figure 11.

No.	Time (s)	No. of error
1	20.78	0
2	26.46	2
3	20.48	2
4	30.52	2
5	23.74	1
Average	24.40	1.4

Table 1. Speech auditory feedback

No.	Time (s)	No. of error
1	16.65	2
2	15.83	1
3	11.21	0
4	18.23	1
5	15.90	2
Average	15.56	1.2

Table 2. Non- Speech auditory feedback

No.	Time (s)	No. of error
1	21.12	3
2	25.21	4
3	18.03	2
4	20.90	2
5	14.23	1
Average	19.90	2.4

Table 3. Without auditory feedback

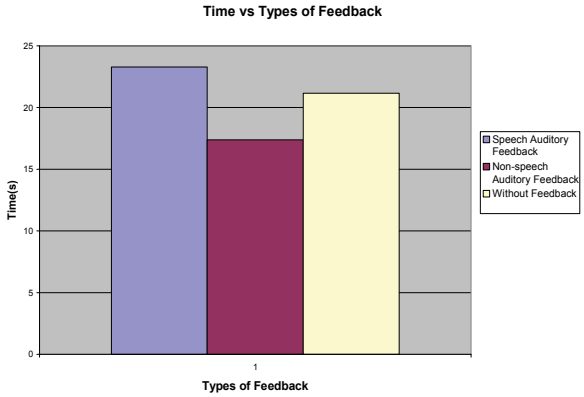


Fig. 11. Task Solving Time

- Discussion

From the result obtained throughout the experiment, it showed that the time taken to solved the task is shortest for the cases where non-speech auditory feedback used as task feedback,

where the average of 15.56s, followed by without feedback: 19.90s, and speech auditory feedback: 24.40s.
The number of errors recorded was found best (least error) which performed by non-speech auditory feedback; 1.2, followed by speech auditory feedback: 1.4 and the worst was without feedback: 2.4.
Throughout the experiment process, the auditory feedback was observed effective in the following way. Auditory feedback helped participants keep track of the ongoing process. For example, auditory alarm (non-speech) feedback alerts the participant for any errors disc movement therefore reduce the possibility for participant of making the same error repeatedly, besides that, a sound generated before the game is going to start, make the participants to be more prepare for the up coming task. It helps on reduce the lagged starting time to the participants. The same situation is happen in which the speech auditory has selected as task feedback. In this case, a more detail feedback message is conveyed to the participant, the participants are more understand what is going on for the occurrence of all that error instead of just an alarm indicated there is an error. For example, a human speech of "error, error, the bigger disc cannot put above the smaller disc", inform the participant the reason of such error.

Auditory Feedback		
Criteria	Speech	Non-speech
Presenting information	Slow	Fast
To assimilate information	Hear from beginning to end	Messages are shorter
	Need many word to be understood	None
	Messages are straight forward	Need to think
	No learning necessary	Required learning to understand
	Not universal (different language)	More universal
Presenting continuous information	Good	Better
Rapid feedback	Good	Better
Convey instruction	Better	Good

Table 4. A comparison between Speech and Non-speech Feedback.

However, the drawback of using speech auditory feedback is presenting information was much slower because of it serial nature, to assimilate information, the participant must typically hear it form beginning to end and many words have to be comprehend before a message can be understood. Therefore, it is time consuming for a participant in solving the task if compared to the others two methods used in these experiments.
For the experiment with no audio feedback technique, the participant seem to solved the task a bit faster than speech auditory feedback but the error made was the most frequent compared to others. This is because the participants had observed to carried a 'try and error' style in solving the task, the drawback is some of the participant made the same error twice, and some of them are not even know why their disc movement were not allowed if performed wrongly. Table 4 shows a comparison between Speech and Non-speech Feedback.

4. Virtual reality for Tower of Hanoi

In this project, an experiment had been carried out in virtual reality for Tower of Hanoi manipulation task at Virtual Reality Laboratory, Faculty of Computer Science in University Malaya. The virtual reality lab is equipped with many virtual reality advanced equipments. By having a stereoscopic view of Tower of Hanoi task, the user will feel the depth of the virtual object, therefore it will enhance the realism of conducting a real experiment.
The experiment is done by compiling and executing the stereoscopic Tower of Hanoi programming source code on the computer which had connected to stereoscopic display device. The user is able to solve the Tower of Hanoi manipulation task in virtual reality environment by using a clicking device (as an input device) that work together with the tracker and sensory device (as shown in Figure 12).

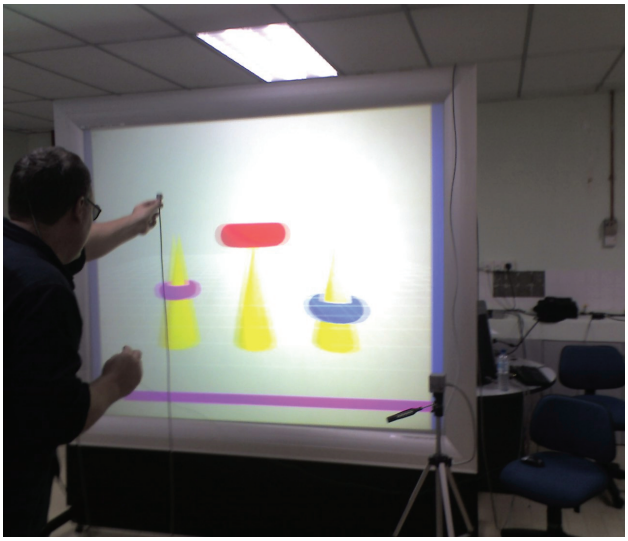


Fig. 12. Manipulating Tower of Hanoi in VR

5. Conclusion

From the results and analysis presented, there are a few conclusions which can be drawn according to experiments.
In general, simulation with real-time feedback is better than simulation without real-time feedback. With real-time feedback, the system can be used to train the user to conduct a proper technique or method, and at the same time avoid making any mistake which may cause errors to the system. For experiment of the effectiveness of visual feedback, it can be concluded that text feedback is better than the colour feedback since the colour feedback only tell the errors but the text feedback has give the description on errors instead of showing the user about the occurred errors only. Colour feedback only can use if the user already been told which colour represent which error message. However, the colour feedback can be used to notify a user on error quicker than the text feedback. Since both of colour and text feedback have their own advantages on errors notification to the user, the combined colour and text feedback will be the best real-time visual feedback among the application with colour feedback only and the application with text feedback only. For the experiments of the Effectiveness of real time auditory feedback, the results of these experiments showed that, the performance of the participants for solving the experiment tasks of Tower of Hanoi could be significantly improved when non-speech auditory feedback is provided throughout the task. It can also been observed that a decrease in error movement made by the participants in solving the task with auditory feedback than without feedback. Using sounds to provide system information is important for several reasons. First, by adding sound to the interface the bandwidth of communication can be significantly increased. Second, the information conveyed by sounds is complementary to that available visually.

6. Future Work

The study was developed to provide real-time visual feedback to the systems to tell the users about the error steps. Future works may include other real-time feedback techniques.Simulation on stereoscopic vision on real-time feedback may also be considered to study the stereoscopic effects on the time taken to complete the task.

7. Acknowledgements

The authors would like to thank Dr. Adam Eppendahl for the permission to use Virtual Reality Centre in Universiti Malaya and a Young Lecturer Scheme (PLB) Grant from UPM in supporting this work.

8. References

Michael, S.P. (1992). *Real-time Systems Engineering And Application*, Kluwer Academic Publishers, 19-22

Philip, A. L. (2004). Real-time Systems Design And Analysis An Engineer's Handbook, *IEEE Press*, 17-18

Sang, H.S. (1995). *Advances in Real-Time Systems*, Prentice Hall, 4.

http://www.techweb.com/encyclopedia/defineterm. jhtml?term=virtualreality

Lai, J.M.; Faieza,A.A. & Sahari, B. (2008). A Study on Real-Time Auditory Feedback Technique in Manipulation Task, In: *Cognitive Informatics: Bridging Natural and Artificial Knowledge*, Halimah Badioze Zaman et al., (Eds.), 535-540,Kuala Lumpur.

Heeter, C. & Gomes, P. (1992). Sounds as Computer Feedback. *Journal of Educational Multimedia and Hypermedia*.

Rauterberg, M. (1999). *Different Effects of Auditory Feedback in Man-Machine Interfaces*, Eindhoven University of Technology,.

Winberg, F. & Bowers, J. (2004). *Assembling the senses: towards the design of cooperative interfaces for visually impaired users*.

Segal, M. & Akeley, K. *The Design of the OpenGL Graphics Interface*, Silicon Graphics Computer Systems 2011 N. Shoreline Blvd., Mountain View, CA 94039.

11

Crowdmags: Multi-Agent Geo-Simulation of the Interactions of a Crowd and Control Forces

Bernard Moulin and Benoit Larochelle
Department of Computer Sciences and Software Engineering, Laval University
Québec, Canada

1. Background and Motivation of the Research

Crowd simulation is a research and development field that has been greatly expanding during the past decade (Thalmann et al. 2007) and has found applications in various domains such as computer games, animation movies, the study of crowd behaviours for egress analysis and evacuation planning; the simulation of crowd situations and control interventions, to name a few. Crowd simulation is also applied to the domain of security planning and crowd management with the goal of helping civilian and/or military control forces to devise and assess intervention plans, and to train personnel in preparation of various kinds of crowd events: evacuation of densely populated areas in emergency situations, evaluation of contingency plans for emergency planning; initiatives to secure downtown infrastructures and populations (Levesque et al. 2008). Our current work takes place in this domain and aims at developing applications of crowd simulations to support decision makers when planning or monitoring crowd events. More particularly, we consider what we call '*purposive crowds*' in which people gather for a specific collective purpose such as demonstrating against measures or regulations enforced by civil or military authorities, celebrating specific events or persons, participating in rallies to promote particular causes, and so forth. It has been observed that in a 'purposive crowd', most people come in groups, often small groups of friends or colleagues, acquaintances or even families (McPhail 1991). It is clear that in such a crowd, different groups of people may try to put different messages to the fore, but they usually fall under the 'umbrella' of the global theme that was used to call for the gathering.

As it appeared in our literature review of a large number of crowd simulations developed during the past 15 years, it is clear that most research teams consider a crowd as an emergent phenomenon resulting from the interactions of a multitude of individuals that have temporarily assembled in a given location. In such a context, researchers proposed various ways to model individual agents, their behaviours and their interactions, so that the emerging collective behaviours resemble the patterns of collective behaviours observed in real crowd phenomena. However, we suggest that when modeling and simulating a purposive crowd, the most important element is not the individual, but the group!

Indeed, individuals reason and make decisions on an individual basis, but their references are groups: groups of demonstrators that they can recognize around them ('in-groups' as sociologists call them), but also 'out-groups' that they perceive as adversaries. A bystander may observe a crowd event as an uninvolved individual, but if she decides to join the demonstration, it is most likely that she will try to join a nearby group of demonstrators which attracts her and will offer her the opportunity to participate in collective actions. Hence, there is a need to simulate the attraction and repelling of agents by groups.

Very few approaches provide ways to model groups explicitly and, to our knowledge, none of them allows for the specification of group interactions. We suggest in this chapter that this is the reason why a true social dimension is still missing in current crowd simulations. Indeed, this social dimension is at the center of the crowd phenomena that have been studied by psychologists and sociologists during the past 30 years. Scholars have shown that such notions as social identity, self-categorization, emotions and inter-group relations, play an important role in understanding and analysing crowd behaviours (Section 2). Indeed, it seems more plausible to model a purposive crowd using an approach based on 'group dynamics' rather than on one which is solely based on individuals' interactions. We adopted such an approach in the *CrowdMAGS Project* in which we developed a simulation framework to simulate the behaviours and interactions of a crowd and of control forces in urban environments in order to assess different intervention strategies using non lethal weapons (fences, tear gas, plastic bullets).

These geo-simulations (Moulin et al. 2003) take place in a virtual geographic environment (VGE) generated from GIS data (GIS = 'Geographic Information System') that faithfully reproduce urban features (roads, buildings, pavements, etc.). Individuals, be they part of the crowd or of control forces, are modelled by autonomous agents which are able to: 1) perceive the environment's characteristics and content; 2) perceive the characteristics and behaviours of other agents and groups; 3) assess all these characteristics in order to choose their own behaviours; 4) carry out individual behaviours as well as collective ones in the group in which they participate; 5) interact with other agents and groups. Our behavioural models extend and adapt in an operational way, the main principles of the *Social Identity Theory* (Reicher 1982) which essentially states that an individual tends to self-identify with one or more social groups and, then aligns her behaviour with what she finds acceptable according to her values.

The proposed approach and the associated software put forward several innovations. This is the first approach of crowd simulations that explicitly models individuals, groups and their interactions, based on their social characteristics, as well as on the assessment of these characteristics by autonomous agents. This approach enables us to plausibly simulate the interactions of a crowd and control forces resulting from both individual and collective actions. Indeed, an agent perceives individuals and groups, assesses their behaviours and may decide to join a group (to participate in its collective actions) or to leave it (and again behave individually) according to its preferences ('social values'). Moreover, agents also react to simulated non-lethal weapons (NLW) that might be used by control forces. We developed the *CrowdMAGS System* which fully implements in multi-agent geo-simulations all the above-mentioned features of our models of crowds and control forces.

In section 2 we review some of the main existing crowd simulation approaches and show how they fail to integrate the notion of individual and collective actions based on sound social theories of crowds. We also briefly review some of the main social theories of crowds

that may be of interest for the simulation of crowds. In Section 3 we propose a number of extensions that might be introduced in crowd simulations to explicitly introduce social notions and mechanisms to explicitly manipulate groups and agents' and groups' interactions. Sections 4, 5 and 6 present the main characteristics of our Agent and Group models as well as their interactions that we have implemented in the *CrowdMAGS System*. Section 7 and 8 present the architecture and the main components of the *CrowdMAGS System*. They also provide illustrations of its practical use. Section 9 concludes the chapter and identifies several perspectives opened by this research work.

2. Crowd Simulation Approaches and Collective Actions

Since critical situations such as escape panic and unplanned evacuations may threaten the public safety, many research works have been carried out on the simulation of dense crowds and models based on particle and fluid dynamics have been proposed to explain people's behaviours in such constrained situations (Helbing et al. 2000, 2001). In these models individuals' behaviours are very simple and mainly consist of reactions to surrounding forces. These physics-based models try to reproduce the geometric characteristics of the observed patterns of 'group movements' in a crowd. However, with the exception of the HIDAC system (Pelechano et al. 2007), these approaches fail to explain why these patterns occur because they lack references to the psychological and sociological characteristics of crowd members. Such models are applicable to simulate certain situations such as pedestrian flows and high-density crowds (as in the case of evacuations), but they are not sufficient to plausibly simulate crowd behaviours in other situations in which people are not physically too much constrained.

Other approaches try to incorporate psychological factors in crowd simulations (Kenny et al. 2001) (Silverman et al. 2002). Most approaches offer models to specify the individual's characteristics (physiological, psychological and emotional) and the individual's behaviours. However, they do not provide sufficient constructs and mechanisms to specify and simulate the interactions between individuals and groups. When it comes to modeling police forces, we did not find any system that convincingly models agents and groups and their interactions with crowds. In the few simulations that introduce agents simulating policemen or soldiers, these agents have limited autonomous behaviours as in the *Crowd Federate System* (McKenzie et al. 2007).

Several systems are able to simulate some aspects of the dynamics of groups in a crowd. However, these systems essentially simulate the dynamics of groups in a kinematic way, taking advantage of the geometric properties (such as distance between group members, orientations, personal space) of agents moving in groups and of attraction/repulsion rules/forces that enable the system to maintain the group's geometrical coherence. Simulating groups in a kinematic way may be sufficient for animation purposes as in the *V-Crowd System* (Musse and Thalmann 2001). However, there is a need for more elaborated models integrating both the individual's characteristics (psychological, emotional) and social rules/behaviours in order to explain why agents may join or leave a group, why perceiving and interpreting the actions carried out by the members of a group (out-group) may induce an agent to change behaviour or even 'change of identity' as some sociologists call it (Reicher 1982).

In the large body of literature on the sociology of crowds and on 'collective actions', several theories have been proposed over the past hundred years such as the 'social contagion' (Le Bon 1895) (McPhail 1991), the 'social identity theory' (Reicher 1982) and the 'social comparison theory' (Festinger 1854). These theories may provide useful insights (i.e. Drury and Reicher 1999, Reicher et al. 2004) to researchers who want to explicitly introduce social interaction models in crowd simulations. However, very few have been used in current crowd simulations, and when they are used (Kaminka & Fridman 2006), the authors only tackle what we call the kinematics of groups: the dynamic geometrical properties of agents' positions in a group.
To conclude, most existing crowd simulations are based on the specification of individual agents' behaviours, and group behaviours are thought of as an emergent phenomenon. Our literature review showed that very few approaches provide ways to explicitly model groups, and none of them allows for the specification of group interactions. This is why we claim that a true social dimension is still missing in current crowd simulations.

3. Extending Crowd Models with Explicit Social and Group Notions

Our model is based on an adaptation of the *Social Identity Theory* (Reicher 1982). This theory states that an individual tends to self-identify with one or more social groups and then aligns her behaviour with those deemed acceptable by the members of that social group (what can be called 'the norms of the group'). Depending on the situation, an individual can shift from a personal identity to a social identity, or from one social identity to another one, and change her behaviours accordingly. We claim that current crowd simulation approaches need to be extended by explicitly introducing social concepts and mechanisms to enable agents to recognize, join or leave a group, and to react to groups' behaviour. Here are the main extensions that we propose:

- Social notions in the agent models such as the social identity and mechanisms to enable an agent to adopt a new identity under certain conditions; this change being triggered by its emotional and cognitive states and by the situation that the agent perceives and interprets;
- The notion of social group to which an agent may belong, and identify to (as for example a group of agitators, a family, etc.);
- The notion of what we call a 'spatial-temporal group' (STG), a group that is easily recognizable in space and time such as a line of policemen and a group of friends walking together;
- Mechanisms for an individual agent to recognize groups (through a perception and interpretation mechanism), to assess their characteristics (by their physical appearance, their actions) and compare them to his expectations, so that the agent may wish to join the group and participate in its collective actions (at least temporarily);
- Mechanisms that enable an agent to join a group or to leave a group.

We suggest that these mechanisms are necessary if we want to simulate and explain collective behaviours and attitude changes in crowd situations involving different kinds of agents and groups such as demonstrators, instigators /agitators and police squads.
As a proof of concept we developed an agent-based model based on the proposed extensions. In the following paragraphs we briefly present these notions.

The notion of social identity. An agent modeling a crowd member should be able to change its behaviour depending on the situation (what happens around the agent), and on the way he interprets this situation. The concept of social identity is used to factor objectives and behaviours of an agent so that it can change them during the simulation. For example, a bystander observing a demonstration may decide to join the demonstrators. A demonstrator may decide to join the instigators in the crowd and consequently may adopt behaviours that are typical of instigators.

The notion of projected image. In most current crowd simulation systems the agents' perception is usually simulated by a simple function that is able to identify the presence of other agents in the vicinity of the perceiving agent. In reality, different agents may observe the same situation and react to it in different ways. Hence, the way that an individual interprets the crowd situation (essentially the perceived behaviours of other individuals or groups in the crowd or in the control forces) significantly influences her decisions and behaviour changes. Hence, it is important to model this interpretation if we want to plausibly simulate phenomena such as social identity change or social contagion. To this end, we introduce the new notion of an agent's projected image. An agent *Ai*'s projected image is a data structure that contains the information made available to the other agents when they perceive *Ai*. An agent *Ai*'s projected image contains the attributes that can be perceived from the outside such as age category, clothing, equipment and attitudes. Moreover, we extend an agent's perception mechanisms with a function that is used to interpret the information contained in the projected images of the agents that he perceives. This function is used to change the agent's beliefs and possibly to trigger some goals or identity changes.

The notion of social group. The notion of group clearly plays an important role in crowd situations, but this notion is poorly modeled in currently existing crowd simulation tools. We propose to introduce the notion of *social group* which characterizes the common characteristics of a group of agents that do not change during the simulation. Crowd members and control forces are examples of global social groups with which agents can be associated. A family, a group of friends or a police squad, are other examples of social groups. In these groups, agents may play different roles, as for example the leader, the deputy-leader and the group members. Roles are associated with typical behaviours of these agents in their social groups. Hence, an agent can belong to one or several social groups and has a current social identity, chosen among a set of possible social identities. The agent's social identity may change depending on the circumstances as it was previously mentioned. Hence, a peaceful demonstrator agent may adopt a social identity of an instigator for a while. But, it can change it during the simulation and become again a peaceful demonstrator. Let us emphasize that social groups exist in the simulation, but they do not appear as spatial entities: they are merely part of the knowledge available to the agents. In contrast, we will call STGs, the groups that spatially appear in the simulation and that can be recognized by the agents and by external observers (users of the simulation).

The notions of a spatial-temporal group (STG) and of a formation. In most simulations, groups are not explicitly modelled; group behaviours are viewed as patterns emerging from the simulation such as the formation of pedestrian flows. In contrast, we propose to introduce mechanisms that will allow agents to purposively join groups during the simulation. For example, after changing its social identity, an agent may decide to join a group of instigators. Hence, we introduce the new notion of *Spatial-Temporal Group (STG)*

which models the groups that can be perceived by agents in the simulation. An STG is associated with mechanisms that allow individual agents to joint it, to dissociate from it and to recognize it. We also associate with an STG the notion of *Formation*, which characterizes the geometrical arrangement of members in the group. For example, a squad of policemen can adopt a line formation or a wedge formation. But, agents in the crowd can also move in formations as simple as 2 agents moving side by side (simplest line) or a group of instigators aligning behind a fence. Conceptually, STGs are also agents and hence have a projected image that agents can perceive.

The notions of interest point and interest area. People are attracted by various kinds of elements in the environment such as for example restaurants, tourist places, shops and monuments. Individuals may also be attracted by other people such as charismatic leaders or even by groups such as a group of demonstrators chanting songs. There is a need to model and simulate these attraction mechanisms. To this end, we introduce in the simulation environment interest points that are objects (which may not be visible to the user) that display different characteristics (thanks to a perceived image) and that the agents may selectively perceive. Hence, depending on his state, an agent may be attracted by some interest points. Interest points may be generalized in terms of interest areas, so that agents can detect areas of interest in the virtual environment.

Interest points/areas are not only associated with objects in the VGE, but also with STGs. In this way, we take advantage of the same mechanisms to simulate the agents' attraction to stationary points/areas and to moving points/and areas. For example, an instigator leader agent (with 'charismatic characteristics') may call for other agents to join. We emulate this potential group as an STG associated with the instigator leader agent. An interest point is attached to this STG and has the potential to attract members to the STG. The crowd member agents that favourably respond to the instigator's call are attracted by the STG's interest point: they move to join the STG and then participate in the associated formation.

The notion of resource. Several types of agent behaviours may need resources which are objects used to carry out these behaviours. For example, tear gas canisters are resources that control forces may launch over the crowd. Stones are resources that instigators and rioters may throw on control forces or on shop windows. A gas mask can also be considered as a resource that an agent may own or may give to another agent. Typically, resources are objects that are needed to carry out various activities. Some resources are limited and agents may compete to acquire them. There are very few crowd simulation systems that explicitly manage resources that agents may use.

4. The Agent Model

Considering Newell's pyramid (Newell 1990) which comprises the physiological, reactive, cognitive, rational and social levels of agent behaviours, we mainly concentrate on the social level in this section. According to the principles presented in Section 3, we suggest to introduce in the individual agent's model some minimal social notions in order to allow him to participate in collective actions: 1) Agent's projected image and interpretative process; 2) Social identity and mechanisms to enable an agent to adopt such an identity under certain conditions; 3) Social affiliation which characterizes social groups to which an agent may belong; 4) Mechanisms allowing an individual agent to recognize groups and assess their

characteristics in order to decide to join/leave a group; 5) Mechanisms that enable an agent to join a group or to leave a group. Here are these notions.

Agent's projected image. An agent *Ai*'s projected image is a data structure that contains the information made available to other agents when they perceive *Ai*. An agent *Ai*'s projected image contains the attributes that can be perceived from the outside such as age category, clothing and equipment. We may also include in *Ai*'s projected image a list of the last *n* activities carried out by *Ai*, so that an agent observing *Ai* may get this information and act accordingly. This enables us to simply and efficiently simulate a mechanism of memory of the perceivable activities carried out by agents.

Agent's interpretative perception. It is well known that different people may interpret in different ways a given piece of information that they perceive. This interpretation process is seldom accounted for in crowd simulations. We developed a mechanism of *interpretative perception* as a function that is added to other perception functions and enables the agent to interpret the information contained in the projected images of the agents that he perceives. This function can be used to change the agent's beliefs and possibly to trigger some goals or identity changes.

Agents and interest points. In Section 3 we introduced the notion of interest point /area. We need to model and simulate agents' attraction mechanisms toward interest points. To this end, an interest point/area is defined as an object (which may not be visible to the user) which displays different characteristics (thanks to a projected image) that the agents may selectively perceive. Hence, depending on the state of an agent, he may be attracted by some interest points/areas located in the VGE.

The notion of social affiliation. An agent may belong to various social groups (a family, a group of friends, a group of co-workers, a sports association, an agitators' association such as the *Black Block*, a police squad, etc.). We suggest to introduce the notion of *social group* which is defined as a set of agents sharing common social characteristics. An agent can be affiliated with several social groups and play various roles in them. For the simulation purposes, we characterize an agent's affiliations as a knowledge structure that identifies the social groups to which it belongs. This information does not change during the simulation and provides the agent with some background knowledge which can be used to recognize agents having common affiliations. Some agents may have distinctive characteristics that highlight their social affiliation as in the case of policemen who wear uniforms and carry equipment that identify them.

The notions of fundamental identity and of social identity. An agent modeling a crowd participant should be able to change its behaviour depending on the situation (what happens around the agent), and on the way it interprets this situation. In line with Cronin and Reicher's *Extended Social Identity Model* (Cronin et al. 2006)(Reicher 1996, Reicher et al. 2004), we suggest that an agent be associated with a *fundamental identity* (mainly composed of its personality traits) and that, in addition, it may temporarily adopt different *social identities* when participating in collective activities. The concept of social identity is used to aggregate an agent's objectives and behaviours, so that it can change them during the simulation. During the simulation an agent may change its social identity any time and depending on the circumstances. For example, a bystander observing a demonstration may decide to adopt a demonstrator's social identity and to join a group of demonstrators. A demonstrator may decide to join a group of instigators and consequently may adopt a social identity and the associated behaviours that are typical of instigators. In our system an agent

has access to a repertoire of social identities that it can use and which are consistent with its current social affiliations as well as with the groups (that we call STGs – Section 5) around him.

Knowledge, beliefs and memory. The agent has some knowledge about itself (attributes such as gender, age and profession), about its environment (location of certain buildings and places) and about other agents. Using perception mechanisms and exploiting information contained in the projected images of the entities (objects, agents, groups) that it perceives, an agent can also acquire new knowledge, while exploring the environment and participating in the crowd situation. This knowledge is application-dependent in the sense that a designer will integrate in an agent model the data structures that are appropriate to record the knowledge that is useful to this kind of agent during the simulation. These knowledge structures are part of the agent's memory: they are often called the 'agent's beliefs'. Several mechanisms can be used to manage the agent's memory (Perron and Moulin 2003).

Needs and resources. Some agents' characteristics may change during the simulation. We call them dynamic states (Moulin et al. 2003). For example, an agent's level of energy can change during the simulation. A dynamic state is represented by a variable associated with a function which is used to compute how this variable changes values during the simulation. The variable may be characterized by an initial value, a maximum value, an increase rate, a decrease rate, an upper threshold and a lower threshold which are used by the function. Using these parameters, the system can simulate the evolution of the agents' dynamic states and trigger the relevant behaviours by relating the dynamic states and the agent's goals (Moulin et al. 2003). More specifically in a crowd simulation, we can model the agent's needs and emotions using such dynamic states. As we mentioned earlier, it is important to take into account the resources that are available to the agent so that it can perform its behaviours. The knowledge of resources, available both internally and externally, and of the agent's dynamic states influence the agent's decision making process, and the selection of its goals.

Autonomy of agents and group belonging. Agents need some capabilities to recognize groups around them, to decide to join or to leave them, when needed. The interpretative perception process carried out by an individual is very important in this context. Indeed, an individual makes decisions about his own actions with respect to the actions of people who are around him. Hence, an agent will need at least to be able to assess the actions of the groups located around him: the group to which it belongs and the groups that it perceives, be they 'in-groups' (i.e. fellow demonstrators) or 'out-groups' (i.e. control forces). Our hypothesis is that an agent A_i continually monitors the actions of the groups located around it in order to determine if it is attracted or repelled by them. The attraction or repulsion occurs when the agent compares the actions of the group to its personal norms (its appreciation of what is a 'good or bad' action in given circumstances). In fact, this is not exactly the same idea as Festinger's *Social Comparison Theory* (1954), since here the objective is not to evaluate one's opinions and actions with respect to other's opinion and actions, but the opposite: an agent A_i compares the collective actions of a group (or of individuals in a group) with respect to its personal norms in relation to this type of actions. Depending on the level of 'acceptability' of the perceived actions, the agent will make a decision about its belonging (adhesion) to the group. It is clear that an agent may need some time before deciding to join or leave a group. Such a decision will result from a cumulative effect of

perceived actions that reinforce the agent's opinion that the group behaves in a way that suits, attracts or repels it.

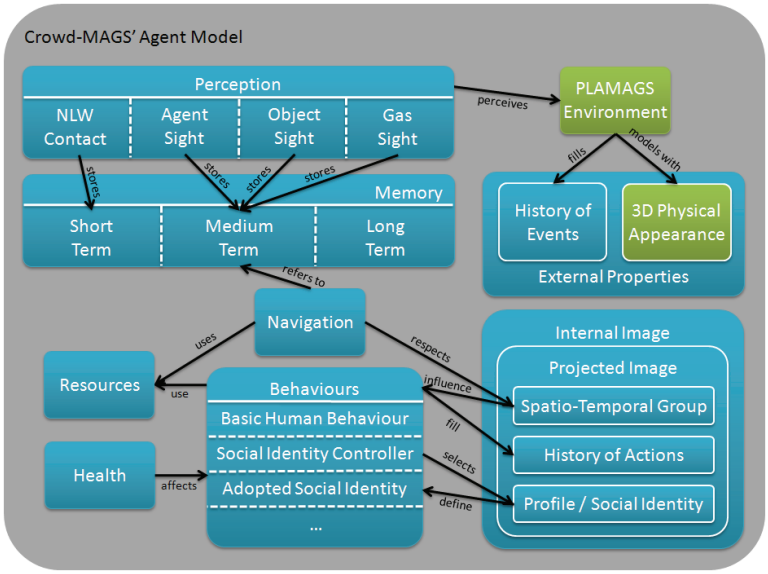


Fig. 1. The Agent Model of the *CrowdMAGS System*

The CrowdMAGS Agent Model. All these notions have been implemented in the CrowdMAGS Agent Model (Figure 1). Here, we briefly present the main elements of this model. Let us mention that the *CrowdMAGS System* has been developed on top of the *PLAMAGS* simulation platform (Garneau & Moulin 2008, 2009) which provides the fundamental functionalities of a simulation engine and an agent behaviour management engine (more details in Section 7).

Thanks to the PLAMAGS simulation engine, the agent can perceive objects, other agents, gas and the effects of Non-Lethal Weapons (NLW) such as gas and plastic bullets. The perceived data is recorded either in a short term memory for immediate processing or in a medium term memory that can be accessed during the simulation run. The agent possesses resources that can be used by its behaviours, including its navigation behaviour. Basic individual behaviours are based on the perception-decision-action loop representation (Lord and Levy, 1994) and correspond to: resource acquisition, navigation management (physical displacement, maintaining a position in an STG), perception mechanisms, memory management, appreciation of aggressiveness. Other behaviours are specifically related to the social identity that the agent currently holds (these behaviours and associated goals are defined in the agent's *Profile*). Eventually, the agent can carry out collective behaviours when it has joined an STG (see Section 5). The agent possesses an Internal Image and a *Projected Image*. In order to be able to collect simulation data for analysis purposes, we also developed mechanisms to record the *History of Events* that may occur during the simulation and that are fed by the *PLAMAGS Engine*, as well as a *History of the Actions* that each agent carries out.

In addition, the Agent Model manages the agent's health level in a simplified way in order to take into account the effects of events or of actions that influence the agent's physical health and its mobility (this needs to be considered in order to take into account the effects of NLW).

5. Spatial-temporal groups and their dynamics

The notion of group clearly plays an important role in crowd situations. However, as it has been shown in Section 2, this notion is usually poorly modeled and used in existing crowd simulations. In most simulations, group behaviours are viewed as patterns emerging from the simulation such as the formation of pedestrian flows. In contrast, we propose to explicitly introduce mechanisms to manage the dynamics and interactions of groups and of indivicual agents during the simulation. In this section we define the notion of Spatial-*Temporal Group* (STG) as well as the associated simulation mechanisms.

The creation of a spatial-temporal group (STG). We first need to examine the mechanisms that are used to create and dissolve STGs. We use a holonic approach to model STGs and to justify the structures that we suggest to include in them. The holonic approach has been used in several domains such as ecology, biology and the design of manufacturing systems. Holonic multi-agent systems have been developed in recent years (Fisher et al. 2003, Rodriguez et al. 2006). The term "*holon*" was originally coined by Koestler (1967) and defined as a self-similar structure that is stable, coherent and composed of several holons as sub-structures. A holon is itself a part of a greater whole, which is also called a holon. Holons are systems that have self-organizing properties and can be used to implement decentralized control. They are also dynamic systems in the sense that they can dynamically aggregate new members, while some members can leave the holon at any time. Indeed, the rules which govern the self-organization of agents (holons) into groups need to be carefully defined. Recently, the holonic approach has been applied to pedestrian simulation (Gaud et al. 2007), but in the context of a physics-based model that does not emphasize the social characteristics of groups.

Coming back to the simulation of purposive crowds and to our concept of STG, we consider that an STG emerges around what we call a '*seed*': an agent that is the origin of the STG. A leader agent may call for the creation of an STG by broadcasting messages around it in order to attract other agents. But, an STG may also be automatically created around its seed as for example police squads are created around their leaders when they appear in the scene. Using a 'holonic vocabulary', we will say that a leader agent is the *head* of its STG. A leader agent provides its STG with directives (objectives) that are used by the STG to coordinate the collective actions carried out by its members. The leader agent also provides the STG with some characteristics that will be recognizable by agents observing it. **A**n example of such characteristics is the STG type. For our crowd simulation we distinguish three types of STGs: demonstrator STG, Instigator STG and Squad STG. Another example of such a STG characteristic is the maximal number of members that the STG can accept. In cases where group membership needs to be limited, the STG is associated with a mechanism that enables it to accept or refuse new members. This is a standard function of a holon's head. Consequently, an agent that wants to join an STG must request the STG's acceptance. If the agent is accepted, it becomes a member of the STG and must behave accordingly. If the agent is not accepted by the STG, it must find another STG that will accept it. If an STG does

not have any member left, its leader may decide to dissolve it. Indeed, the dissolution conditions depend on the STG's properties and on the application domain. An STG is also associated with mechanisms that allow individual agents to recognize it, to join it and to leave it. We discuss these mechanisms in Section 6.

STG's projected image. As other entities that can be perceived by agents in the VGE, an STG is associated with a projected image (Section 3) which contains the information that agents may get about the STG when perceiving it. Examples of such information are: the characteristics of the STG (such as the STG's type and number of member agents), information about the collective actions carried out by the STG's members (during a parameterized duration), and possibly the global emotions that result from these actions. These actions can be computed using an algorithm based on the results of socio-psychological studies of collective actions performed by groups in crowds. As for the general mechanisms related to projected images (Section 3), all agents perceive STGs' information, but each agent can interpret it in its own way.

The hot-spot and the STG's attraction mechanism. Using another general notion introduced in Section 3, we use the notion of interest point or area (that we call hot-spot when it is associated with an STG) in order to allow agents to locate the STG in the VGE and to get information from its perceived image. We associate each STG with a hot-spot (interest point/area) which is attached to the leader agent that controls the STG. Hence, an STG's hot-spot moves with the leader agent. This simple mechanism enables us to efficiently simulate the agents' perception of STGs. Consequently, an individual agent can easily identify the STGs located around it, in order to eventually decide to join one of them. Suppose for example that an instigator leader agent calls for agents to join it. We simulate this potential group as an STG Some crowd member agents perceive the STG's hot-spot and projected image content and may decide to favourably respond to the instigator's call. Hence, they move toward the STG's hot-spot in order to join the group and to participate in its collective actions. When joining the STG, the individual agents take a position in the STG's formation.

STG's choreography and the notion of a formation. Depending on its specific characteristics, an STG is associated with a particular geometric configuration and rules that govern the movements of its member agents, when they participate in collective actions requiring particular coordinated movements. We call this aspect: the *STG's choreography*. Typical examples are the different geometric configurations (also called formations) of police squads. In the police procedures (also called 'doctrine') there are standard geometric configurations such as the line and wedge formations, that a squad may adopt when facing demonstrators. In order to deal with this aspect, we introduce the notion of *Formation* which characterizes the geometrical arrangement of members in the group. An STG is associated with a number of formation types that can be used when carrying out certain collective activities. In the simulation, when an agent is accepted by the STG's leader, it is assigned a position in each formation associated with the STG. The position corresponds to parameters that refer to the relative movements that the agent will have to perform in the formation. The formation management mechanisms that we developed are fairly generic and allow for a variable number of participants and various geometric configurations and agent movements in these formations (wedge and line formations in relation to the leader agent's position, two agents side by side, unorganized formation of agents around a hot-spot in a given area, etc.). Since an STG formation is represented in the VGE, the associated area can be perceived by the agents: this area is used as a 'hot-spot' by attraction mechanisms. We

can see such areas in Figure 3B: these areas are displayed so that the user can easily identify the STGs which have been formed during the simulation.

STG's attraction / repulsion. In our approach, we consider that an STG assigns behaviours to its member agents according to their roles and to the 'choreography' of the collective action performed by the members of the STG. Consequently, whenever an agent has decided to join an STG, it agrees to carry out the actions imposed by the STG in the context of the collective behaviour. Empirical observations of structured groups in crowd situations show that the members of these groups usually behave as if they were performing actions imposed by the role they play in such a group. For example, let us mention military and police groups as well as sports teams that are trained for such coordinated behaviours (movements), but also agitator and demonstrator groups when they are well supervised and trained. It is clear that in reality, leaders of less structured groups find it more difficult to impose uniform behaviours to their members: hence we allow for loose formations of agents around a leader agent. Moreover, an agent can always leave an STG if it does not agree with the individual actions that are collectively imposed by the STG.

Acting collectively or individually. Since we are particularly interested in collective actions, we distinguish the individual and collective actions that agents may perform during the simulation. All the agents that do not participate in collective activities as STGs' members, act on their own and hence carry out individual actions. Members of STGs participate in collective actions. Drawing such a distinction helps us to simulate collective actions in a more efficient way since we do not have to trigger complex reasoning mechanisms while agents act as members of STGs. When it comes to individual behaviours, we distinguish two categories of agents: the leaders of STGs and the other agents acting on their own. In agreement with our assumption that STGs' members carry out collective actions imposed by the STG to which they belong, we propose that a leader agent makes decisions on behalf of the STG that he leads. Examples of such leader agents' activities are: managing collective goals, collective needs and resources in relation to the current situation. Using such an approach, the more complex decision making activities are carried out by a limited number of agents, the STGs' leaders, which again enhances the performance of the simulation. Agents who are not leaders or members of STGs, also carry out individual actions: we call them 'uninvolved agents'. However, in the simulation of purposive crowds, the actions of uninvolved agents are usually less complex because most of these agents are observers (bystanders) that may eventually decide to leave the scene or to join demonstrators (and consequently participate in collective actions).

The role of the STG's leader agent. We mentioned that for simplification purposes an STG is created around a leader agent using one of three possibilities: 1) there exists a social group that appears in the scene with the leader (i.e. police squads); 2) a leader agent moves around in the VGE and is able to attract other agents that adhere to its STG; 3) a pre-existing social group appears with its leader agent in the VGE and creates an STG that is able to attract new members during the simulation. We also assume that the leader agent makes decisions for its STG and that member agents perform collective actions under the command of the leader agent. Certain categories of agents may be 'programmed' to follow the orders of the STG leader without questioning them, as in the cases of control forces or of groups of trained instigators. In the case of other agent categories, an agent may decide to leave a group if the imposed collective actions do not agree with its personal values/standards

STG's resources. Makie and her colleagues (2000) emphasized the crucial role that the availability of resources plays when one group actively aggresses against another one and that the appraisal of the in-group's strength produces emotions (anger or fear) towards the opponent group ('out-group'). Hence, we can reasonably posit that the power of a group can be evaluated by considering certain characteristics such as the number of group members, the equipment that they possess, the material or 'moral' possibility for group members to use this equipment, and the training that group members have acquired in carrying out collective actions with or without equipment. In our approach, a leader agent manages the tactics of its STG with respect to its goals (which are assumed to coincide with the goals that the agent pursues on behalf of its STG) and to the resources that it manages. We consider a number of variables which characterize an STG's resources such as the number of STG members, their skills to perform collective actions, the 'compound power' of the individual resources.

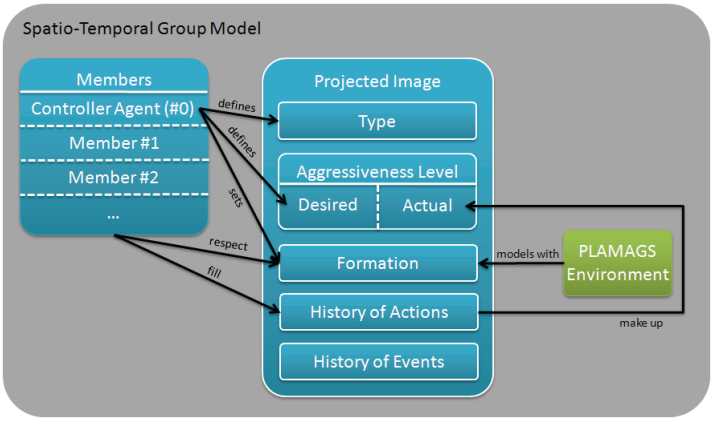


Fig. 2. The CrowdMAGS' STG Model for Spatio-Temporal Groups

The formations are managed by the PLAMAGS engine and are used to coordinate the group members displacements in different ways (Figure 3).

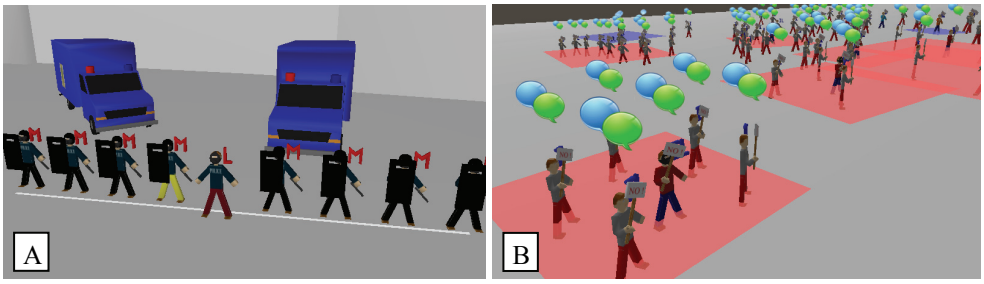


Fig. 3. A: Formation of squad members (M) around the squad leader (L). B: Chanting demonstrators gathered in different groups (STGs identified by their interest areas)

The CrowdMAGS' STG Model. All the above mentioned notions have been implemented in the *CrowdMAGS'* STG Model (Figure 2) which provides the data structures and behaviours

required to manage the STGs. An STG possesses a Projected Image in which are accessible the STG type, its History of Actions and History of the Events in which it has been involved during the simulation. It also contains information about the formation(s) that the STG members should adopt when participating in the STG's collective actions. At any given simulation step, the formation is chosen by the STG leader. An STG also contains a list of its agent members. The STG leader is distinguished and provides the decisions (goals) that influence the STG's collective actions carried out by its agent members. The STG possesses all the basic behaviours that enable it to manage its membership, as well as the procedures that are needed to compute the STG overall aggressiveness.

6. Dynamics of Individual Agents and STGs

An individual can be attracted by a group because it feels some commonalities with this group; either because it wants to participate in the collective actions performed by the group's members, or because the emotions displayed by the group fit the individual's mood. From a psychological point of view, we can say that the group offers the individual the opportunity to express her feelings and emotions through the participation to certain collective actions (Smith 1993, Mackie et al. 2000). In our approach, we emphasize that for an agent A_i an important individual process consists in constantly choosing if it will join a new STG, or continue to stick to the STG it belongs to, or dissociate from it; while considering the STG's actions (resulting from the collective actions of the agents, members of the STG) as they are perceived/interpreted by agent A_i. The decision to join or leave an STGs is based on the individual agent's appreciation of aggressiveness.

Enthusiasm and Appreciation of Agressiveness. By observing the collective activities that take place around him, an agent may become excited and feel an urge to participate in the collective actions. Conversely, an agent may be disapproving the collective activities that it observes and become reluctant to participate in them. To this end, we introduce the notion of enthusiasm which basically represents the overall appreciation of the collective actions being carried out around an agent. Thus, enthusiasm takes its values in [-1, +1] with +1 expressing an extreme enthusiasm (or excitation), 0 being neutral and -1 expressing a complete reluctance to participate in collective actions.

In order for agents to plausibly make decisions based on the collective actions carried out in their surroundings, it is hypothesized that agents must express a certain appreciation for different levels of aggressiveness. For example, a bystander might highly appreciate chanting and probably does not appreciate instigators throwing projectiles. On the other hand, instigators might appreciate throwing projectiles and may find chanting too 'passive'. Due to the complexity of simulating these 'feelings', each agent cannot interpret in its own manner the actions carried out around it. Thus, we propose that agents be characterized by appreciation profiles, which can be adopted by multiple agents. For normalization purposes, aggressiveness takes its values in [0, +1] with +1 expressing an extreme aggressiveness and 0 expressing a complete absence of aggressiveness.

In crowd control events, violence is a major concern: peaceful demonstrators want to avoid violence, instigators may seek opportunities for violent actions and control forces want to limit violence and disruption of public order. Consequently, collective actions are often qualified on an aggressiveness scale. We propose that this scale range from -1 (very peaceful) to +1 (very aggressive). Estimating a ranking of collective actions on such a scale is

feasible. For example, McKenzie and his team (McKenzie et al., 2005) qualified the actions of individuals and groups in a crowd using an aggressiveness scale. Once an acceptable set of collective actions has been defined, the scale of the agents' appreciation of aggressiveness can also be defined. In CrowdMAGS System, different scales can be defined depending on the agent's profile.

Adhesion to Spatio-Temporal Groups.

It has been already mentioned that individual agents seek to participate in STGs based on the collective activities being carried out. In addition, if an agent is already participating in the collective activities of an STG, it needs to decide if it will stay or leave the STG. Thus, these decision rules must be modelled taking into consideration the collective actions. Because all STG members are autonomous, the collective actions being performed might change quite rapidly. Current members might not appreciate the actions of some new members, but they may not necessarily want to leave the STG immediately. In consequence, we introduce the notion of *support* which is defined as an agent's long-term appreciation of the STG's collective actions. Because the appreciation of aggressiveness takes its values in [-1, +1], support is constrained to the same range. Coming back to the previous example, the new members' actions might bring down the other members' support values towards the STG, but the recent actions will bear only a certain weight with respect to the history of actions that have been performed before. Thus, some members might eventually leave, when their support towards the STG drops below a certain level. Agents who want to join a STG might also wait before their support goes above a certain threshold.

All these notions have been implemented in the *CrowdMAGS System*.

7. The CrowdMAGS System

In order to develop micro-simulations of crowd situations, we needed a software platform that provides agents with basic capabilities such as reactive navigation in a virtual geographic environment (VGE), perception of agents and of features/objects of the VGE, decision making capabilities including the manipulation of hierarchies of goals. Several platforms allow such agent micro-simulations such as (Moulin et al. 2003, Lamarche et al. 2004, Paris et al. 2007, Pelechano et al. 2007, Garneau and Moulin 2008, 2009), to name a few. We chose the PLAMAGS Environment (Garneau et al. 2008) that provides us with a 3D engine to manage the 3D Virtual Geographic Environment and the agents immersed in it (physical appearance, navigation and collision avoidance, etc.), as well as a powerful behavioural engine that manages the agent's behaviours in relation to their goals and available resources.

Figure 4 presents CrowdMAGS' main architecture (the main components and mechanisms) and its relationship with the PLAMAGS Platform.

At the lowest level of all components sits the PLAMAGS simulation engine, which can be viewed as the "master" component. The engine creates the PLAMAGS environment (i.e. the VGE) to handle physical interactions of agents and it manages on its own the behavioural aspects of the agents. In fact, it is the simulation engine that starts counting iterations and increases the iteration number after all components in the iteration have been executed. The Crowd-MAGS' architecture simplified this process by ensuring that only two types of components would be executed by the simulation engine at every iteration. First of all, the behavioural graphs are executed for each agent (Garneau and Moulin, 2008, 2009). Then, the

simulation engine notifies the *Scheduler*, which executes the components that need to be executed at this particular iteration (since not all components are executed at every iteration).

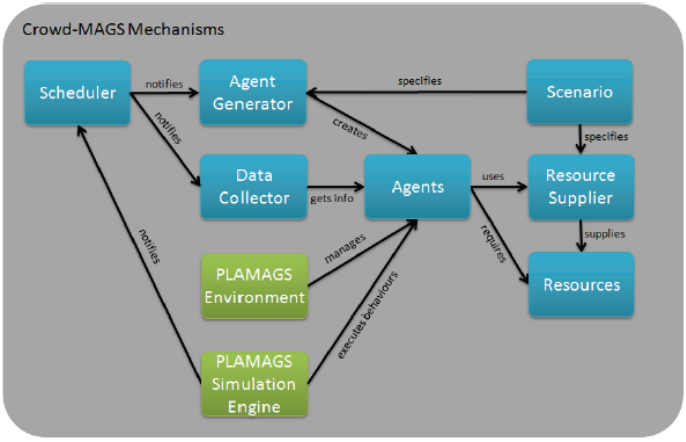


Fig. 4. Overview of the CrowdMAGS' Architecture

The *Scheduler* is one of the most generic and important mechanisms in the *Crowd-MAGS System*. Not all components need to execute code at every iteration. For example, the navigation behaviour needs to be called ten times per second, but the behaviour that regulates social identity changes is called only once per minute. The Scheduler has been created to manage these different requirements. The simulation engine notifies the Scheduler at every iteration, which in turn notifies only the components that need to be activated at that iteration. The Scheduler can also allow a single component to be called at different frequencies for different purposes. The Scheduler can also be used to register punctual events rather than periodical events. For example, gas cans last for only a short moment once they are set off. Thus, when thrown, the gas can objects register with the Scheduler to be called once to start the emission of gas, and once again to stop it.

The resource supplier is a fairly complex component that can provide agents with one or more specific resources. The resources provided by a resource supplier may be anything from a physical object (e.g. banner, tear gas can) to less concrete concepts (e.g. "relaxation" and "healing power" to aid agents who have been hurt). Thus, a resource supplier can represent any concept from a pile of banners left on the ground to a first aid tent. In fact, a resource supplier does not even need to be embodied in the VGE (e.g. a zone such as a park could provide "relaxation" to bystanders). When considering a simulation scenario (Section 8) a designer must indicate which resources can be provided by each resource supplier of the simulation. Three parameters are required as inputs: the name of the resource to be provided, the quantity available (0 to infinity), and the time necessary to obtain the resource. The interactions that agents can have with resource suppliers consist of requesting and cancelling resources. Agents may request multiple resources at one time from the same supplier. For example, if an agent requests three resources that each have a delivery time of 15 seconds, then he can get its three resources after 15 seconds, not 45 sec. For example, it

may take 15 seconds for an officer to get his helmet from the back of a police truck, and it would take approximately the same time to get a helmet and a baton. After an agent has made a request, it may cancel it, or cancel all of its pending requests. Resource suppliers can answer queries with respect to which resources they provide and to how many are left. Suppliers can also be refilled with more units of certain resources, whenever necessary.

The agent generator. This component allows the scenario designer to specify the types of agents to be generated at certain locations and at certain moments during the scenario. An agent generator is a component in the VGE that is not embodied, but that has a conceptual location. Using a generator is very simple: first it must be created and then generation requests must be assigned to it. The first step to add an agent to the simulation is to create an agent object. Then, *PLAMAGS* creates the component that will represent this agent in the environment. This PLAMAGS component mostly contains the physical attributes of the agent such as its mass, its perception capabilities and its visual representation. Next, the *Data Collector* is notified that a new agent was created, so that the collector can keep track of this agent. Then, agent configurators perform initialization and configuration tasks. Each agent possesses an initialization method that it uses to adopt its fundamental social identity and profile, effectively giving it the appropriate behavioural graphs, icons, and color of clothes. The initialization method also sets the necessary variables such as destination, speed, and projected image. Finally, the initialization method registers the agent with the *Scheduler* for all generic mechanisms and behaviours such as perception and the evaluation of enthusiasm. Once the initialization is complete, an event is added to the agent's *History of Events* recording that it has been added to the simulation.

Generic Data Collection. There is no central algorithm to collect simulation data, but rather a collection of small algorithms distributed throughout various components of the system. However, there is a central repository (called the *Data Collector*) that stores all data collected system wide. The reason for this decentralization is that a multi-agent simulation is so dynamic that monitoring all components entails a large overhead. For example, a central "monitor" would have to maintain lists of components in the simulation and run through every element periodically to gather whatever information is necessary. Instead, all components in the *Crowd-MAGS System* are free to register themselves with the *Data Collector* and to provide any information that they want to make available publicly. Overall, most of the data collection work is done when actions and events happen to individual components, such as an agent being dragged or receiving a plastic bullet.

One of the only active tasks of the Data Collector is to write a text file about each agent when it is removed from the simulation. This file can contain basic information, the histories of actions and events, and any other information that may be useful for the analysis of simulation results. Similar files are created for STGs as well.

The Scenario Manager. This component manages the scenarios that the user has created for the simulation. The user can create scenarios, edit them and record them. More details in Section 8.

System Main Interface. The *CrowdMAGS System* offers a sophisticated interface (Figure 5) that enables the user: 1) to create a virtual geographic environment (VGE) and agents (both for the crowd and for control forces); 2) to specify various scenarios for the simulation; 3) to play the role of a commander of control forces who chooses intervention strategies (mobilization level and choice of NLW) that agents composing the control forces will carry out autonomously. Using the control bar on top of the main window (see letter A in Figure

5) the user can control the simulation (accelerate the simulation step, pause, search for a particular agent. The main window displays the simulation in the VGE: the user can navigate in it (controlling the camera) to observe the simulation from different angles. When the simulation is paused, the user can modify the content of the VGE: he can add or remove fences and agents and objects in the VGE. The user can also modify the control forces' strategy using the panel located under the main window (see letter C in Figure 5). The user does not dictate the behaviours of the control forces, but rather chooses a degree of involvement (which conforms to control forces' doctrine) that he communicates to the squad leaders by clicking on the corresponding button in the panel. The user, playing the commander's role can also allow the use of tear gas and/or plastic bullets. We can see in Figure 5 that the user has allowed the use of tear gas.
The window on the left hand side (see letter B in Figure 5) enables the user to inspect all the agents and the STGs using different tabs: basic information, resources, memory content, spatio-temporal group information, history of actions and history of events.

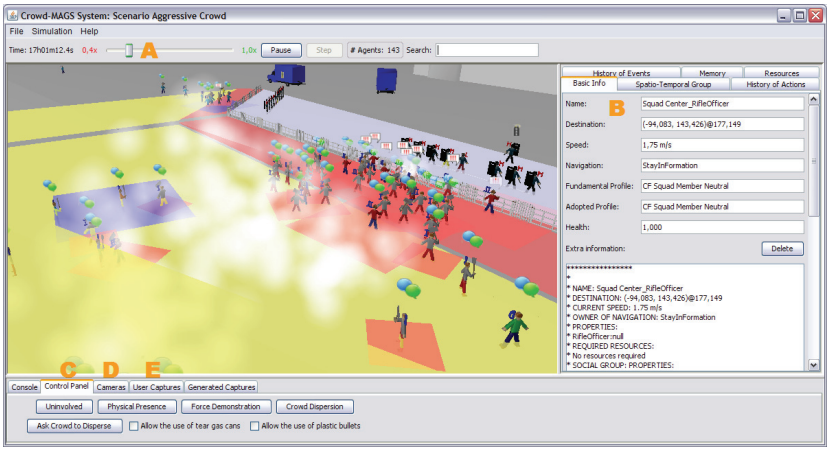


Fig. 5. The Main Interface of the *CrowdMAGS System*

When clicking in the *Cameras* tab (see letter D at the bottom of Figure 5), the user can activate new cameras for inspecting the scene from different view-points. The camera is set at the position from which the user is currently observing the scene. Another window is created and opened and will display the scene from the chosen point of view as long as it is opened. The user can create several external camera windows (Figure 6). The user can also take still pictures of the scene using the tab *User Captures* (see letter E at the bottom of Figure 5) and record them in files for future use.

8. Specifying Scenarios and Running Simulations using the CrowdMAGS System

Scenarios. The user can specify and play different scenarios. Each scenario is recorded as an XML file for inspection and subsequent use, either for replay purposes or to be used as a basis for the creation of variations of a given scenario.

Figure 7 shows the main window used for editing scenarios. The middle part shows the simulation window, just like in the regular interface (Figure 5). The only difference is that all elements are motionless, since the simulation is not running. On the left of Figure 7 is the palette, which allows adding simulation components to the scenario with a single click. The user must first select what type of component he wants to add to the scenario, and then click in the main window wherever he wants one instance of the component to appear. The available types are: agent, fence, interest point, police truck, agent generator, media truck, journalist, and tear gas can. An information panel to edit components can be seen on the right of Figure 7.

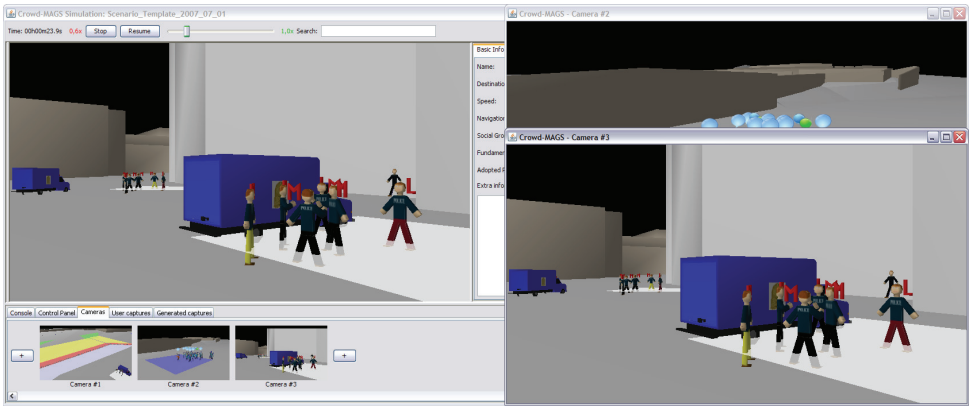


Fig. 6. External Camera Windows

Options are available to select the time of appearance, the position, the fundamental social identity for agents, the available resources, and other details. Finally, scenario parameters, such as the start time, the crowd distribution, and non-lethal weapons impacts, can be edited with the configuration panels, shown in Figure 8. This panel is accessible from the menus in Figure 7. Nearly all parameters in the XML scenario File can be edited with these configuration panels.

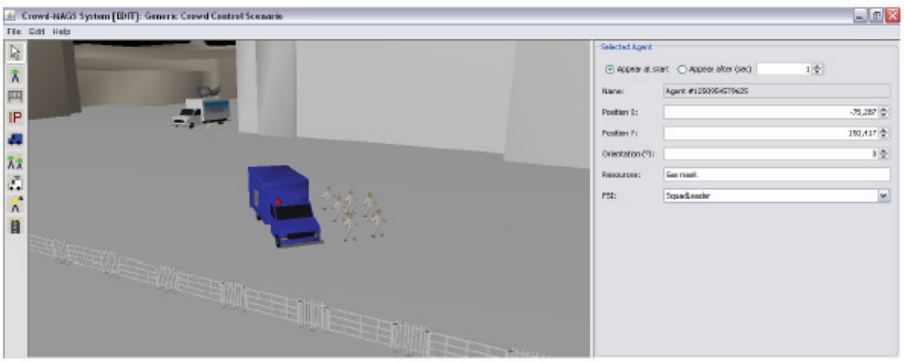


Fig. 7. The Scenario Specification Main Screen

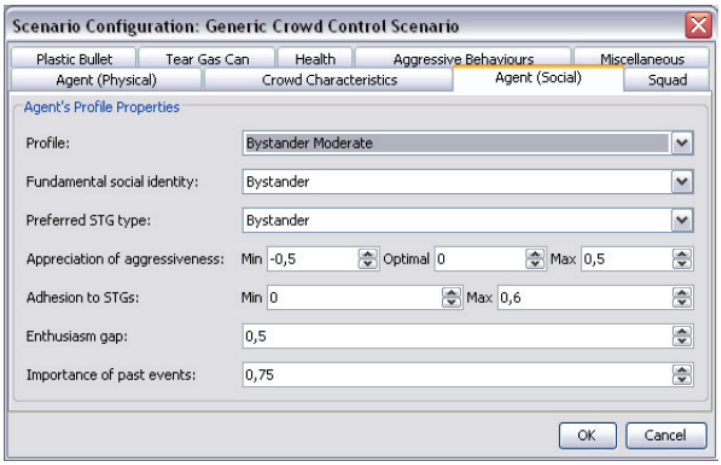


Fig. 8. The Scenario Configuration Panel

In a scenario the user can choose how many agents of each category he wants to be involved in the simulation. In Figure 5 one can notice that crowd members wear outfits of different colours that show their fundamental social identity (combinations of yellow, blue and red pants). They have also an adopted social identity which is shown by a letter on top of their shoulder. Agents can also carry out accessories such as placards, helmets, guns, and shields. When the agents are shouting or chanting, this is shown as the dialogue bubbles on top of them (see the blue and green bubbles for example in Figure 5). These icons are easy to understand and enable the user to have a global view of the crowd simulation. In a scenario the user chooses the locations in which each agent or set of agents (agents are created in a set at the same time by an agent generator) must appear and when during the simulation. If required, the user can also assign particular itineraries to specific groups of agents (as for example the succession of interest points where a march should move through).

When the simulation runs, the agents are autonomous and behave according to their behaviours, profile, decision making process (taking into account the agent's objectives, behaviours, appreciation of aggressiveness, etc.). The user can pause the simulation to modify the VGE (i.e. add fences), add or remove agents. He can also give orders to the control forces as we mentioned earlier. Moreover, the *CrowdMAGS System* contains another model, that we call the Information Model, managed by the *Data Collector* (Section 7), and which is used to record the history of events and actions of the agents and STGs. Taking advantage of this model, the user can specify which variables he wants to be recorded during the simulation and in which format. After the simulation, he can use the generated files for thorough analyses using statistical packages. More details in (Moulin and Larochelle 2009).

Calibration and Experimentations. The Crowd-MAGS system allows for the customization of a large number of parameters, some of which can be calibrated based on isolated tests and on scientific studies found in the literature. Others, mostly related to crowd characteristics, must be calibrated in plausible test scenarios before being considered acceptable for more complex scenarios. As an illustration, let us comment upon the calibration approach that we followed to develop our prototype of crowd simulation. Many parameters were initially

calibrated while the system was being developed. The values were calibrated on the basis of available data in the literature, and relying on the qualitative assessment of the realism of the effects of each parameter during isolated tests. To further calibrate the models, we chose a fundamental scenario taking place in front of the Quebec Parliament involving a crowd and a fixed number of control forces' squads. We created 3 kinds of crowds (passive, moderate and aggressive) with different proportions of bystanders, demonstrators and instigators (the proportion of leaders of each category was also adjusted).

Control Forces' Strategy	**Crowd**		
	Passive	**Moderate**	**Aggressive**
Uninvolved	1a	1b	1c
Force Demonstration + Communication	2a	2b	2c
Physical Presence + Tear Gas		3b	3c
Force Demonstration + Tear Gas		4b	4c
Physical Presence + Plastic Bullets		5b	5c
Force Demonstration + Plastic Bullets		6b	6c
Physical Presence + Tear Gas + Plastic Bullets		7b	7c
Force Demonstration + Tear Gas + Plastic Bullets		8b	8c
Crowd Dispersion		9b	9c
Crowd Dispersion + Tear Gas		10b	10c
Crowd Dispersion + Plastic Bullets		11b	11c
Crowd Dispersion + Tear Gas + Plastic Bullets		12b	12c

Table 1. An overview of scenarios involving different control forces' strategies for different types of crowds.

Then, we created different scenarios in which the commander chooses different degrees of involvement for the control forces, eventually using NLWs (tear gas and/or plastic bullets) (See Table 1). One of the purposes of this experiment was to assess how Control Forces adopting different levels of involvement (and eventually using different types of NLWs) would influence the crowd behaviour. Using Keeney's top-down approach (Keeney and Gregory, 2005) we identified fundamental attributes, such as '*Ensuring public safety*' and '*Minimizing costs*', that were relevant for analysis purposes. Then, we refined these general attributes in terms of variables that would be either measured by the system or computed from variables measured during the simulation. Here is a partial list of these variables: *Control forces' intervention level*, *Number of people harmed*, *Amount of resources used*, *Crowd size and ratio with respect to initial size* (every 15 seconds), *Crowd aggressiveness* (every 15 seconds). Several parameters needed to be set in the simulation. For instance, we chose initial settings to assess the aggressiveness associated with typical collective actions such as: *Chanting* (0.15), *Showing banner* (0,15), *Yelling* (0,2), *Insulting* (0,3), *Fighting* (1). Let us emphasize that it was chosen to assess the overall behaviour of the crowd by computing the crowd's aggressiveness every 15 seconds during a simulation run (5 minutes). The tests carried out with these initial settings were surprising and showed relatively small differences in the resulting aggressiveness of our 3 crowds. Several elements could be changed in addition to these settings, in particular the functions that set the tolerance to aggressiveness for the different categories of crowd members. We did a series of trials for the different scenarios and for different settings of these parameters. For illustration purposes, Table 2 shows a

comparison of the aggressiveness of the 3 crowds for scenario 2 (see Table 1) as obtained after the 6th trial. Obviously, we cannot detail further these experiments in this chapter. See for more details (Larochelle 2009).

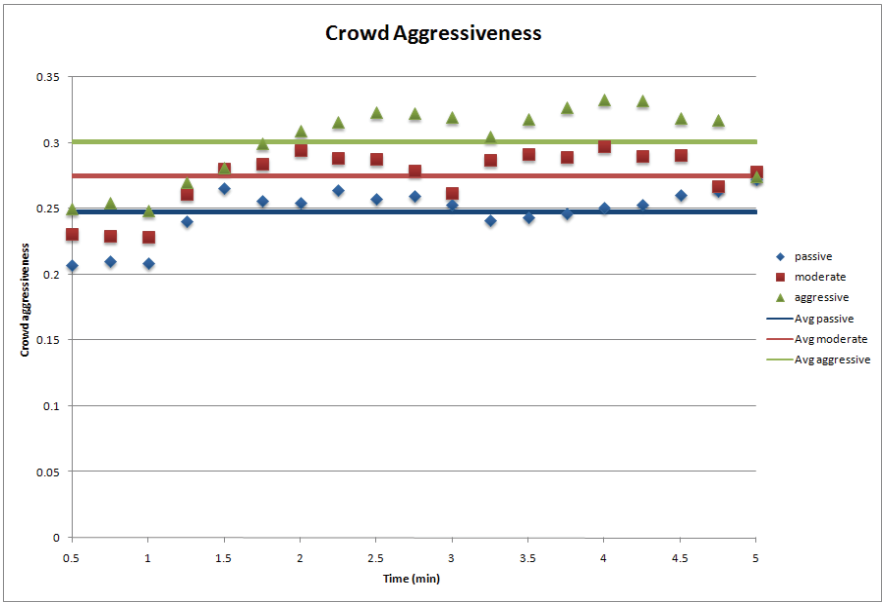


Table 2. Comparison of the aggressiveness levels of the three crowds for Scenario 2

9. Conclusion

In this chapter we proposed new agent and group models that explicitly take into account the social dimension that is used for the management of collective actions in groups of agents that we call spatial-temporal groups (STG). Our models apply to the simulation of both crowd members and control forces' officers, as well as to their collective behaviours in groups and their interactions with groups. These new models push further currently existing approaches for crowd simulation, while explicitly introducing a social dimension in relation to the management of groups of agents. These generic models have been adjusted in the context of the CrowdMAGS Project while using the PLAMAGS platform.

We used PLAMAGS as a development environment and a language to create multi-agent geo-simulations and we extended its capabilities in order to create the proposed models. Hence, we discussed in details the architecture of our CrowdMAGS system and presented details of the system's practical use (scenario-based development, user interface, data collection and analysis).

We developed an Information Collection Model which is composed of the various structures that are used to collect and organize data obtained during the simulation. This data can be used for analysis purposes.

In conclusion, we must mention that this project has been fairly effective in opening new grounds for the development of crowd simulations with agent models in which the social

dimension is explicitly taken into account not only at the individual level, but also at the group level. Still numerous enhancements might be considered as a continuation of this project. The current simulations allow the introduction of a maximum of 800 agents with reasonable execution time. These performance limits are mainly related to the PLAMAGS behaviour engine which, it must be emphasized, provides a sophisticated management of agents' behaviours, states, goals, pre- and post-conditions of actions, concurrent resource management, as well as the management of concurrent goals. We need to examine how PLAMAGS' behaviour engine can be improved.

For our demonstration purposes we developed a set of profiles, social identities and behaviours for the different types of agents involved in the simulation of crowd members as well as control forces. These models could be greatly improved as a result of careful socio-psychological analyses of the typical behaviours of people that can be observed in various crowd situations. Such analyses should be carried out by multi-disciplinary teams that might take advantage of the *CrowdMAGS Platform* to test and compare them.

It would also be very fruitful to develop a variety of simulation scenarios in different urban environments, with different kinds of crowds (more or less aggressive) gathering agents of different categories (and variable numbers) and to further calibrate the system and models.

Acknowledgements

The CrowdMAGS Project has been mainly financed by the Canadian Defence (RDDC Valcartier). It has also benefited from the support of Geoide, the Canadian Network of Centers of Excellence in Geomatics, and NSERC, the Natural Sciences and Engineering Research Council of Canada.

10. References

Cronin, P. & Reicher, S. (2006). Study of factors that influence how senior officers police crowd events: On SIDE outside the laboratory. *British Journal of Social Psychology*, 45, pp. 175-196.

Drury, J. & Reicher, S. (1999). The Inter-Group Dynamics of Collective Empowerment: Substantiating the Social Identity Model of Crowd Behaviour. *Group Processes and Intergroup Relations*, vol. 2(4), pp. 381-402.

Festinger, L. (1954) A theory of social comparison processes. *Human Relations*, 7(2), pp. 117-140.

Fischer, K., Schillo, M. & Siekmann, J. (2003). Holonic Multiagent Systems: A Foundation for the Organisation of Multiagent Systems. In: Holonic and Multi-Agent Systems for Manufacturing, pp. 71-80.

Garneau, T. & Moulin, B. (2008). PLAMAGS: A language and environment to specify intelligent agents in virtual geo-referenced worlds. *Proceedings of the 19th IASTED International Conference on Modeling and Simulation*, Quebec city, May 2008.

Garneau, T., Delisle S. & Moulin, B. (2009). Effective agent-based geosimulation development using PLAMAGS, Chapter 30, In A. Lazinica (Edt.), *Modelling, Simulation and Optimization*, ISBN 978-953-7619-36-7, In-Tech, Rijeka, Croatia.

Gaud, N., Gechter, F., Galland, S., Koukam, A. (2007). Holonic multiagent multilevel simulation : Application to real-time pedestrians simulation in urban environment, *Proceedings of the 30th International Joint Conference on Artificial Intelligence (IJCAI'07)*, Hyderabad India.

Helbing, D., Farkas, I.J. & Vicsek, T. (2000). Simulating Dynamic Features of Escape Panic. *Nature*, n. 407, pp. 487-490.

Helbing, D., Molnar, P., Farkas, I. J. & Bolay, K. (2001). Self-Organizing Pedestrian Movement. *Environment and Planning B: Planning and Design*, vol. 28, pp. 361-383.

Kaminka, G.A. & Fridman, N. (2006). A Cognitive Model of Crowd behaviour Based on Social Comparison Theory. In: *Cognitive Modeling and Agent-Based Social Simulation*, M. Afzal Upal & R. Sun (Eds.), Papers from the 2006 AAAI Workshop, American Association for Artificial Intelligence, pp. 25-34.

Keeney, R. L. & Gregory, R. S. (2005). Selecting attributes to measure the achievement of objectives. *Operational Research*, 53(1), pp. 1–11.

Kenny, J. M., McPhail, C., Farrer, D. N., Odenthal, D., Heal, S., Taylor, J., James, S., & Waddington, P. (2001). *Crowd Behaviour, Crowd Control, and the Use of Non-Lethal Weapons*, Technical Report, Penn State Applied Research Laboratory.

Koestler, A. (1967). *The Ghost in the Machine*. (reprint Pinguin Group 1990).

Lamarche, F. & Donikian, S. (2004). Crowds of virtual humans: a new approach for real time navigation in complex and structured environments. *Proceedings of the Computer Graphics Forum, Eurographics'04*.

Larochelle, Benoit (2009). *Multi-Agent Geo-Simulation of Crowds and Control Forces in Conflict Situations: Models, Application, and Analysis*. MSc Thesis, Université Laval, Département d'informatique et de génie logiciel, August 2009.

Le Bon, G. (1895). *La psychologie des foules*. Paris : Édition Félix Alcan.

Levesque, J., Perron, J., Hogan, J., Garneau, T. , & Moulin B. (2008). CAMiCS : Civilian activity modelling in military constructive simulation. In *Proceedings of the SCS Spring Simulation Multi-Conference*, Ottawa, Canada, April 2008.

Lord, R. G. & Levy, P. E. (1994). Moving from cognition to action: a control theory perspective. *Applied Psychology*, 43(3), pp. 335–398.

Mackie, D. M., Devos,T. & Smith, E. R. (2000). Intergroup emotions: Explaining offensive action tendencies in an intergroup context. *Journal of Personality and Social Psychology*, vol. 79 n. 4, pp. 602-616.

McPhail, C. (1991). *The Myth of the Madding Crowd*. New York: Aldine de Gruyter.

McKenzie, F. D., Xu, Q., Nguyen, Q.-A. H., & Petty, M. D. (2005). Designing physical layer components in a reconfigurable crowd federate. *In Proceedings of the Spring 2005 Simulation Interoperability Workshop*.

McKenzie, F.D., Petty, M.D., Kruszewski, P.A., Gaskins, R.C., Nguyen, Q-A. H., Seevink, J. & Weisel, E.W. (2007). Integrating Crowd Behaviour Modeling into Military Simulation Using Game Technology. *In Proceedings of Simulation and Gaming Online First*.

Moulin, B., Chaker, W., Perron, P., Pelletier, P., Hogan, J. & Gbei, E. (2003). MAGS Project: Multi-agent geosimulation and crowd simulation. In: *Spatial Information Theory*. Kuhn, Worboys & Timpf (Eds.), Springer Verlag, LNCS 2825, 151-168.

Moulin, B. & Larochelle, B. (2009). *The CrowdMAGS System on the PLAMAGS Platform: A Scientific and Technical View*. Contract Report, Defence RD Canada Valcartier, March 2009.

Musse, S.R. & Thalmann, D. (2001). A Hierarchical Model for Real-Time Simulation of Virtual Human Crowds. *IEEE Transactions on Visualization and Computer Graphics*, vol. 7(2), pp. 152-164.

Newel,l A. (1990). *Unified Theories of Cognition*. Harvard University Press, Cambridge.

Paris, S., J. Pettré, J. & Donikian, S. (2007). Pedestrian reactive navigation for crowd simulation: a predictive approach. Computer Graphics Forum. *Eurographics'07*.

Pelechano, N., Allbeck, J. & Badler, N. (2007). Controlling Individual Agents in High-Density Crowd Simulation. In *ACM SIGGRAPH / Eurographics Symposium on Computer Animation (SCA)*.

Perron, J. & Moulin, B. (2004). Un modèle de mémoire dans un système multi-agent de géo-simulation. *Revue d'Intelligence Artificielle*, vol 18 – n.5-6, Hermes, pp. 647-678.

Reicher, S. D. (1982). The determination of collective behaviour, In: *Social Psychology and Intergroup Relations*. H. Tajfel (Edt.). Cambridge University Press.

Reicher, S., Stott, C., Cronin, P. & Adang, O. (2004). An Integrated Approach to Crowd Psychology and Public Order Policing. *Policing: An International Journal of Police Strategies and Management*, vol. 27(4), pp. 558-572.

Rodriguez, S., Gaud, N., Hilaire, V. & Koukam, A. (2006). Holonic Modeling of Environments for Situated Multi-agent Systems, *Environments for Multi-Agent Systems II*, pp. 18-31.

Silverman, B. G., Johns, M., O'Brien, K., Weaver, R. & Cornwell, J. B. (2002). Constructing Virtual Asymmetric Opponents from Data and Models in the Literature: Case of Crowd Rioting. *Proceedings of the Eleventh Conference on Computer-Generated Forces and Behaviour Representation*, pp. 97-106.

Smith, E. R. (1993). Social identity and social emotions: Toward new conceptualizations of prejudice. In: *Affect, cognition, and stereotyping: Interactive processes in group perception*. D. M. Mackie & D. L. Hamilton (Eds.), San Diego, CA: Academic Press. 297-315.

Thalmann, D. & Musse, S. R. (2007). *Crowd Simulation*, Springer Verlag.

12

PLAMAGS: A Unified Framework and Language for Efficient Multi-Agent Geo-Simulation Development

Tony Garneau and Bernard Moulin
Département d'informatique et de génie logiciel, Université Laval
Québec, Canada
Sylvain Delisle
Département de mathématiques et d'informatique, Université du Québec à Trois-Rivières
Québec, Canada

1. Introduction

The micro-simulation of social and urban phenomena using software agents in geo-referenced virtual environments is a field of research whose popularity has strongly grown recently. Geo-simulation (Benenson and Torrens 2004) is an approach which became popular in geography and social sciences in recent years. It is a useful tool to integrate the spatial dimension in models of interactions of different types (economical (Fagiolo et al. 2007), political, industrial (Gnansounou et al. 2007), medical, social, etc.) and it is thus used to study various complex phenomena, especially in the domain of urban dynamics (Foudil and Noureddine 2007) and land cover planning.

Since these phenomena usually involve large populations in which individuals behave autonomously, several researchers thought to take advantage of multi-agent simulation techniques (d'Aquino et al. 2003; Guyot and Honiden 2006; Gnansounou et al. 2007; Papazoglou et al. 2008), which resulted in the creation of the new field of Multi-Agent Geo-Simulation (Koch 2001; Moulin et al. 2003). However, most geosimulation applications deal with very simple agent models, mainly expressed in terms of simple behavior and decision rules, either attached to spatial portions (i.e. cells in cellular automata) or to simple agents moving around in a virtual geo-referenced space (Benenson and Kharbash 2005; Müller et al. 2005). Indeed, the degree of sophistication of agent models depends on the scale of the simulation. For example in the traffic simulation domain, different kinds of simulations are developed at macro-, meso- and micro-scales in order to respectively study traffic flows in regions of different extent (macro- or meso-level) or to create micro-models based on individual vehicles' behaviors (Helbing et al. 2002; Bourrel and Henn 2003). Nevertheless, most models that drive such simulations of agents' movements in geographic space are either based on mathematical models (usually systems of differential equations) or on simple rules (Torrens and Benenson 2005; Levesque et al. 2008).

However, whatever the sophistication of the models, specifying agent behavior models is a difficult task and designers need efficient and user-friendly tools to support them. Some

existing tools for agent-based simulations, such as HPTS (Donikian 2001), AI.Implant (AI.implant 2009) and PathEngine (PATHEngine 2009), deal with the spatial aspects of agent behaviors by providing good navigation mechanisms for the characters. Unfortunately, they tend to neglect the proactive aspects of agents and their interactions with the environment. Other tools such as SimBionic (Fu et al. 2002) and SPIR.OPS (SPR.OPS 2009) offer sophisticated mechanisms to specify objects/agents behaviors based on models inspired by finite state machines. But, the use of finite state machines leads to complex graphs to represent relatively simple reactive behaviors. Behaviors developed using these tools lead to reactive agents or "navigation driven" agents (Cutumisu et al. 2006). Hence, they are not sufficient for the development of geo-simulations of social phenomena in which agents need to implement knowledge-based capabilities in relation to the space in which they evolve. In both cases, the resulting agents do not have decision-making capacities. Moreover, since most of these tools do not provide perception mechanisms, agents cannot apprehend the virtual environment (act in the environment and interact with the objects/agents contained in it).

To help solve these problems, we claim that software agents with space-related capabilities should be introduced in the virtual spatial environments associated with geo-simulations. These agents, that we call "spatialized agents", are characterized by the following properties:

- Autonomous and individual perception mechanism
- Decision-making in relation to a geo-referenced virtual environment
- Proactive and autonomous behaviors taking into account their knowledge about the world (the virtual environment).

The specification of this type of agents is a difficult task and, to our knowledge, no existing simulation environment enables designers to fully specify spatialized agents. In this chapter we present the PLAMAGS Project in which we developed an agent-oriented language, a development environment and a 3D visualization engine completely dedicated to the development and the execution of multi-agent geo-simulations (MAGS), involving spatialized agents.

Section 2 presents the PLAMAGS methodology that we propose to develop MAGS. Section 3 introduces the architecture and the main concepts on which the PLAMAGS language is based. Section 4 presents the main elements that are composing a PLAMAGS simulation. Section 5 discusses the characteristics of the language and its IDE and Section 6 concludes the paper with a discussion and some future work.

2. PLAMAGS method

Contrary to the majority of existing MAGS and MABS methods which generally put the emphasis on modeling and design, our PLAMAGS method is tightly linked to the specification language that was designed to support each step of the development cycle, using a syntax that is both declarative and procedural. The user can thus easily carry out the modeling, the implementation, the execution and the validation of the simulation within a single framework. Our method is composed of 12 steps shown in the Figure 1.

The PLAMAGS method proposes a generic and progressive approach which aims at supporting a designer when creating a MAGS, which is specified and tested in an

incremental way, thanks to the PLAMAGS Development Environment and the associated language.

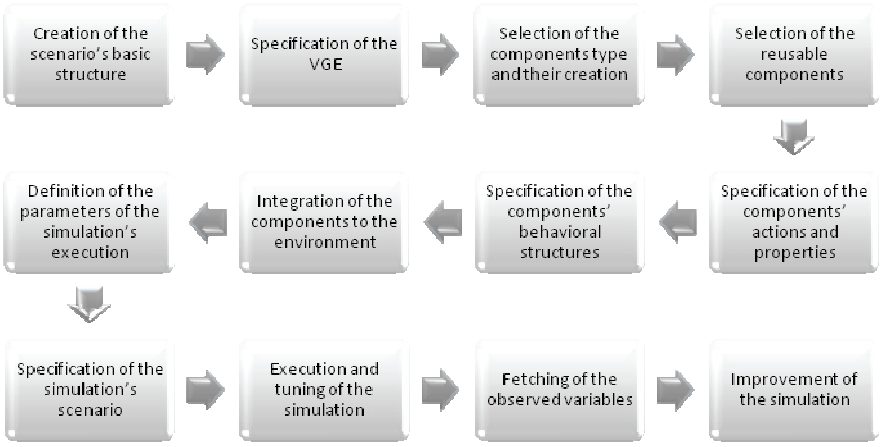


Fig. 1. Overview of the PLAMAGS' method.

As soon as the designer has specified the elements required at a given step of the methodology, he can implement these elements using the PLAMAGS language. Then, the PLAMAGS environment allows him to execute this partial simulation, so that the designer can observe the results and eventually detect any anomaly/error resulting from the specification. Indeed, he can make the required corrections and run the partial simulation again. In that way, the designer can readily get partial results of the simulation under construction, and make all the adjustments that are required to make sure that the specification is correct and yields the expected results. Each step of the method is supported by a set of statements of the PLAMAGS language and by specific components of the Development Environment. The method supports a 2-level iterative design process during which it is possible to come back to the previous steps (or sub-steps) when needed.
Each of the method steps leads the user to carry out an action sequence which aims at getting a partial simulation result, when completed. Figure 2 presents the action sequence (of sub-steps) that applies to most of the steps of Figure 1. These sub-steps can also be carried out in an iterative way and can be refined by the user if needed.

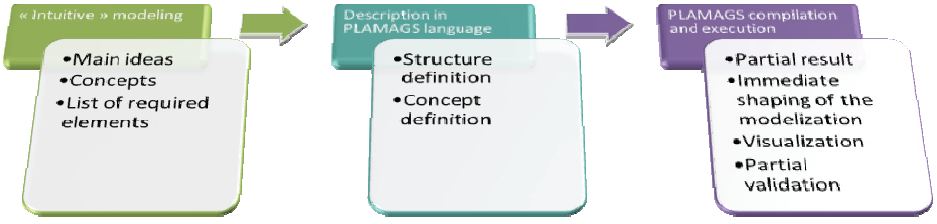


Fig. 2. Sub-steps that apply to each step of the method.

For example, in a first iteration of the development process, at the VGE specification step (second step in Figure 1), the user might carry out the following actions.

- "Intuitive" modeling: the user chooses the place where the simulation will occur, the dimensions of the virtual geographic environment (VGE), etc.

- Description in PLAMAGS language: the designer uses the language to create a scenario and specifies the 3DS model of the ground, the textures, the system of coordinates, the dimensions, the model orientation on the screen, etc.

- PLAMAGS compilation and execution: the user compiles and executes this initial scenario and visually validates into the simulator that the VGE model is suitable to the simulation, that it has been positioned correctly, that the proportions are realistic, etc.

2.2 Advantages of the PLAMAGS approach

The PLAMAGS approach (Method + Language + Development Environment) allows the designer to get a partial result that can be executed at every step of the development process, which brings many advantages compared to other theoretical or conceptual methods. Here are some of them:

- Early detection of modeling or design mistakes
- Validation of the simulation result at every step
- Transparency between the conceptual and the implemented models
- Easy approach and quick materialization
- Progressive development and refinement of the simulation
- Flexibility in the definition

Developers are often forced to use simulation development tools that do not directly support the concepts that are defined in the conceptual models, which inevitably leads them to modify the models in order to implement them. Indeed, such modifications lead to an additional work-load and increase the risks of introducing errors and discrepancies with the conceptual models. Moreover, even when the tools allow for a direct translation of the specification to the simulation code, the fact that the code is written in a general programming language that is not dedicated to the simulation, the implementation task is slow, complex and propitious to mistakes of various kinds. The great advantage of the PLAMAGS Approach is that it provides a unique and complete language for specification, definition, implantation and execution of the models, and consequently eliminates the translation process between the theoretical/conceptual models and their implementations.
Compared to other methods, the PLAMAGS two-level process allows the user to make sure that the partial model is implemented, tested and validated at the end of each step. This allows him to progressively carry out the model validation during its creation. This kind of validation is usually impossible when using other simulation design methods.

3. PLAMAGS' Architecture

This section presents a synthesis of the relations that exist between the different elements that we use in the development cycle of MAGS. We will present the general principles of the PLAMAGS framework and show how they are related.

A fact that greatly contributes to make a MAGS development a challenge is the necessity to work with two very distinct sets of concepts expressed either using a Geographic Information System (GIS) or a Multi-Agent Based Simulation, which do not address the same modeling problems. Thus, we need to conciliate these differences in order to simplify the work of designers and developers. To this end, the PLAMAGS Approach proposes a way to easily model, specify and implement the Virtual Geo-referenced Environment (VGE) as well as the links and interactions between the agents and the VGE. Indeed, the PLAMAGS' architecture deals with the following issues: i) the characterization of the relations between the Geographic Information (GIS) and the multi-agent based simulation (MABS); ii) the description of the VGE within the simulation; iii) the interaction between the VGE and the simulation's components (objects and agents); iv) spatial and physical consistency; v) sensorial capacities of the agents and objects; vi) the VGE's influence on the agents and objects during the simulation.
As a matter of fact, Figure 3 illustrates the principles on which the PLAMAGS Approach is based. As illustrated, the main components of a PLAMAGS simulation's model are: 1) a VGE that renders the simulation environment; 2) the agents and objects located in this VGE (which are characterized by specific behaviors); 3) the simulation scenario and, finally 4) the results (or outputs) of the simulation. These four elements are in constant interaction and constitute the core of the system.

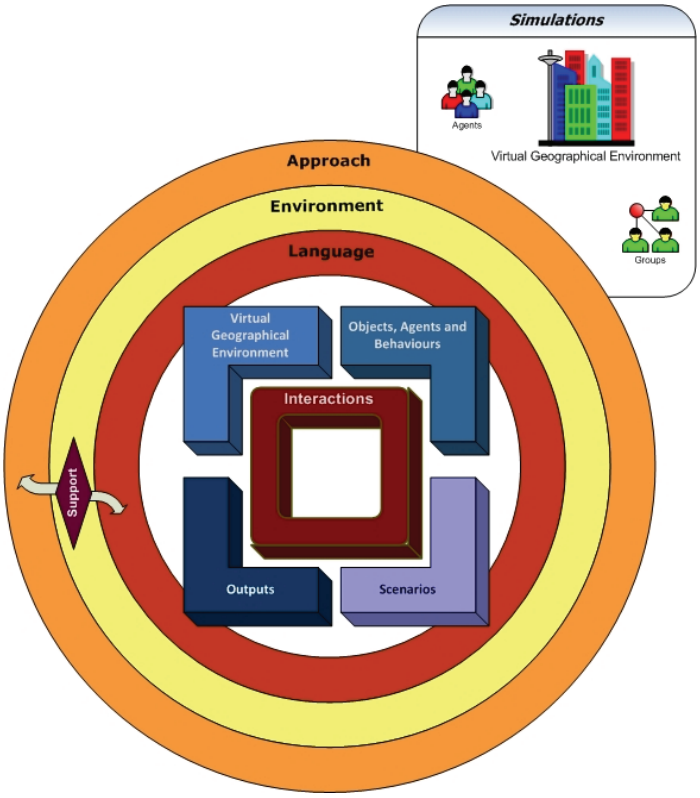


Fig. 3. PLAMAGS' architecture.

However, at a more global level, the PLAMAGS' architecture can be thought of as being composed of two distinct parts: 1) the programming language (Garneau et al. 2008) fully dedicated to the specification, definition, execution and the deployment of MAGS and all its elements; and 2) a set of tools to facilitate its use. These tools are bundled in an IDE (Integrated Development Environment) that simplifies as much as possible the language's usage. The IDE also provides the user with a development framework.
PLAMAGS also offers a powerful behavioral model that is modular and extensible and is used to specify the agents' behaviors (Garneau et al. 2010). Similarly to tools which have "navigation or spatial driven" behaviors, PLAMAGS offers a lot of predefined navigation actions. In addition to basic navigation, PLAMAGS introduces behaviors using multi-layered directed graphs that, contrary to other models inspired by finite state machines, manage the concurrent execution and multiple concurrent states (using "on the fly" context addition and withdrawal), infinite decomposition (sub-behavior layers), expressive and powerful transitions between nodes proceeding into three phases: activation, execution and completion (using rule lists, resources and priorities). Section 4.2 presents the main characteristics of the behavioral graphs.

3.1 Materialization of the geo-spatial aspect

The model presented in Figure 3 features a close bond between the VGE and the agents (as well as their behaviors). One of the main challenges at this level is to define in a simple and legible way for users who are not expert at manipulating GIS, the structures appropriate to describe the characteristics and properties of the VGE, while offering an acceptable level of details to create a complete VGE.
The PLAMAGS Approach simplifies the specification of the VGE thanks to dedicated constructs of the language and functions of the IDE. Again such an approach provides several advantages:

- To make available the structural and spatial composition of the VGE to the agents' decisional process
- To integrate the definition of characteristics of the VGE at the scenario level
- It allows agents to interact with the VGE during the simulation development
- To extract comprehensive spatialized information ("outputs", even for beginners in GIS).

In order to be able to integrate and use the spatial and physical characteristics of the VGE in a simulation, it is necessary that the properties of the VGE be defined in an intuitive way (in formats that are legible by everyone) and that these properties be directly usable during the execution of the simulation without needing that a developer carry out any transformation, conversion or complex calculations.
Practically, a geographical environment possesses an infinite number of characteristics and properties. It is thus unthinkable to exactly model it in the VGE at the simulation level. One must rather facilitate the specification of the VGE's characteristics by simplifying and selecting the appropriate features of the environment.
The integration and management of the geographical environment in the PLAMAGS simulation model are done at different levels. First, the visual, spatial and physical characteristics of the geographic environment as well as the objects/agents must be defined (section 3.1.1). Then, it is necessary to characterize the interactions and the relations between the VGE and the objects/agents of the simulation (section 3.1.2). Finally, when considering

the agents' behaviors, the interactions and the feedback coming from the environment must be taken in account (section 3.1.3).

3.1.1 Definition of visual, spatial and physical properties

The first step of the VGE's specification consists in defining the properties and characteristics of the geographical environment and of the objects/agents that will be used by the simulation. We can distinguish three categories: the visual properties, the spatial properties and the physical properties.

Visual properties

The visual properties allow the designer to configure the visual rendering of the simulation (simulation display). Several of these properties define pointers either to files containing the 3D structures (using the 3DS format) of the various components, or to images, as well as to their textures and colors ("png", "jpg" or other formats).

Although the visual properties have a limited influence on the development of the simulation, they are essential to display the simulation results at the execution time. Visualizing the simulation results is the first feedback that a user gets from its specification and that enables him to inspect the simulation output. The visual rendering also allows a developer to carry out a series of quick validations (checking the components present in the simulation, correcting positions, checking specific actions occurring when a specific object is perceived by an agent, etc.).

Spatial properties

Moreover, the spatial properties allow the designer to define the logical basis of the VGE structure and of the simulated objects/agents. Among these properties, let us mention the measuring units, the dimensions, the volumes, and the objects'/agents' positions in the VGE. These spatial properties relate the geographical environment to the simulated objects/agents and are used by their decision making processes (at the behavioural level).

Physical properties

The physical properties (gravity, components' weight, friction, etc.) allow for keeping a certain level of coherence and of reality in the management of the spatial interactions between the objects/agents and the VGE. For example, the physical properties can be used to determine the effects of a collision between two components, the effect of a movement of the object/agent on a steep ground, etc. They are also used to simulate the propagation of gases. The management of the physics is an extremely complex mechanism which is demanding in terms of calculations. This mechanism is managed in the PLAMAGS environment with the help of a very powerful external library called PhysX (PhysX 2010) which is used in many video games.

Using these visual, spatial and physical characteristics allow the designer to create a simplified definition, yet representative, of the geo-referenced environment, of the objects and agents as well as of their relationships.

3.1.2 Definition of the agents' sensory capacities

Once the visual, spatial and physical properties have been defined for the objects, the agents and the VGE, the designer only needs to define the properties corresponding to the sensory capacities of the objects and agents in order to be able to get back the perception information (these properties correspond to a subset of the visual, spatial and physical properties described in the previous section). Figure 4 presents some of the main properties allowing the designer to define the sensory capacities of an active object or an agent.

The "perceivable" property specifies that the agents that possess it can be perceived visually by other agents. The "perceiveSelfMouvements" and "perceiveSelfAltitude" properties offer to the agent the capacity of getting upon request the information relative to its location and its "comfort zone" described in the middle section of figure 5. The five last properties define the capacities of perception of the agent (described in block 1 of figure 5). For example, "fieldOfView 200,180,10" means that the agent perceives elements located in a circle of 200 distance units (with respect to the coordinates defined in the VGE) and that its sector of perception is of 180 degrees in front of it.

```
19    map perceivable : true
20    map perceiveSelfMovements : true
21    map perceiveSelfAltitude : true
22
23    map fieldOfView : "200, 180, 10"
24    map perceiveGroups : true
25    map gasFieldOfView : "100,180"
26    map perceiveGas : true
27    map personalSpacePerceptionDistance : 0.5
```

Fig. 4. Properties allowing to define some sensory capacities.

3.1.3 Application of geo-spatial and physical concepts in the simulation

Once the sensory capacities have been defined, it is possible to manipulate them directly and to integrate them in the agents' decision making process.

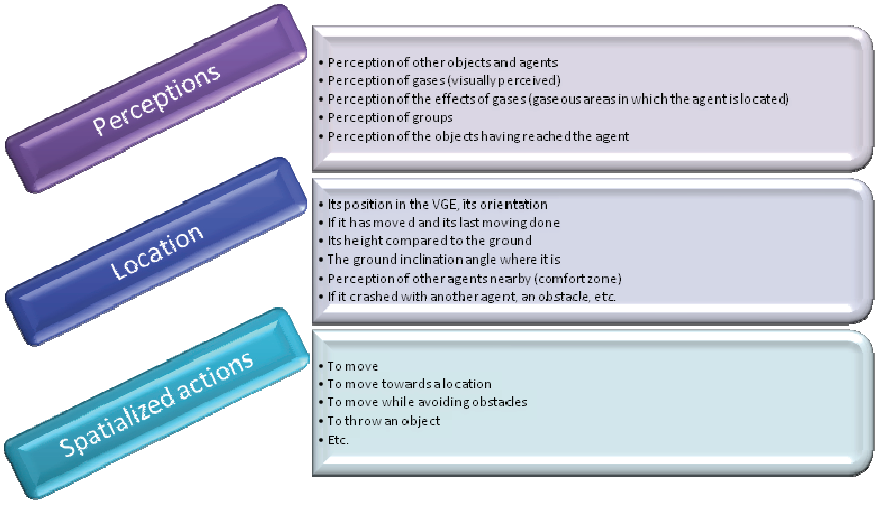


Fig. 5. Summary of the main interaction mechanisms between the agents and the VGE

To facilitate these manipulations and support properly the integration of the geo-spatial aspect into the simulations, the PLAMAGS allows the explicit definition of the spatialized interactions between the agents and the VGE. To this end, the different interaction mechanisms are directly integrated in the language and available through the use of 3 keywords*: percepts, references and perform.* Indeed, the interactions are sorted by categories, each of them grouping a collection of functionalities/capacities semantically related. Figure 5 sums up the spatialized interaction mechanisms defined in the language.

Perceptions
An agent can get its perception information at any time during the simulation run (this partial knowledge of the environment can be used by the decision process). The perceptions are sorted in five categories that can each be treated in an independent way. Figure 6 shows the way to get back the objects/agents perceived by an agent. The system only needs to access the list of perceived objects/agents (using *references.Sight*). Each of the components is accessible afterward.

```
public void perceiveAgents()

    local objects : references = [references.Sight]
    local objectList : array<object> = [objects.toArray()]
    local obj : object

    for [i : int = 0;i < objectList.size();i = i + 1]
        set obj = [objectList.get(i)]
        ...
    end for

end method
```

Fig. 6. Retrieval and manipulation of components perceived by an agent.

Location
The language also allows to obtain an agent's geographic location. (Figure 7)

```
public void printSelfInfo()

    call println("xMovement: " + percepts.xMovement)
    call println("zRotation: " + percepts.zRotation)
    call println("elevationAngle: " + percepts.elevationAngle)
    call println("orientationAngle: " + percepts.orientationAngle)
    ...

end method
```

Fig. 7. Retrieval of information about the spatial situation of the agent.

Given its perception list and its own spatial situation, an agent can determine which agents are located around it, as well as the objects located in the environment in which it evolves. This knowledge is called "location knowledge".
The location knowledge also allows for the definition of a minimal zone (comfort zone) that is necessary for an object/agent to exist in the VGE. If another object/agent gets into this zone, then some forces are automatically applied to attempt pushing back the intruder

(using the PhysX functions). The applied forces and their effects are configured by the parameters defined in the location knowledge of the objects and agents implied in the spatial conflict. Such a capability allows for the automation of the proximity management of objects/agents (figure 8), in addition generating more realistic spatial micro-behaviors (for example, the movements of an agent making his way through a crowd).

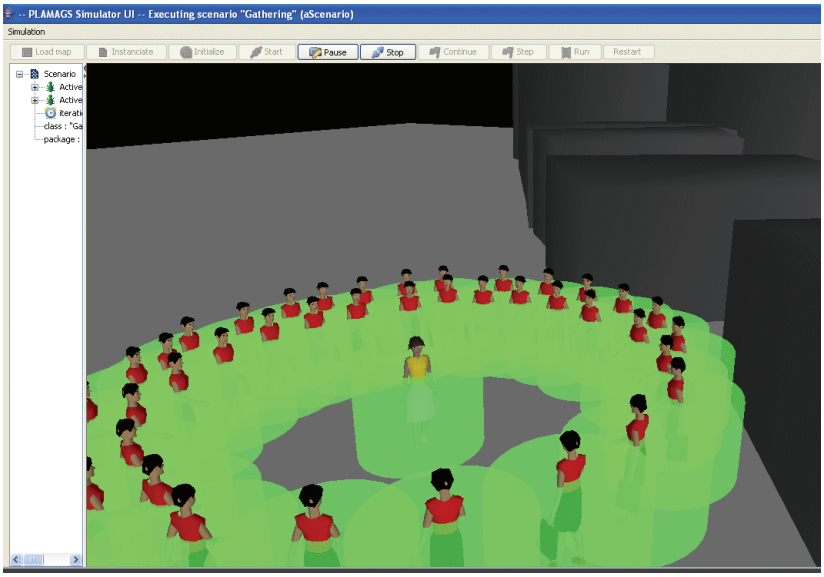


Fig. 8. Use of the confort zone (debug view)

Spatialized actions

The spatialized actions allow agents to spatially interact with the VGE and with other agents. Figure 9 shows how an agent can perform a 5-degree rotation and do a movement toward position 134, 158 at the speed of 3.6 units per iteration.

```
public void move()

    perform relativeRotation(5)
    perform moveToward(134,158, 3.6,true)

end method
```

Fig. 9. Use of spatialized actions

Availability and usability of the interactions model

The PLAMAGS interaction model is clearly defined and completely integrated with the language syntax: it is not necessary to set specific configurations, neither to use a complex syntax, nor to make use of an external module. The direct invocation of these mechanisms through the language allows a user to easily integrate the geo-spatial information in the agents' decisional processes. As an example, when a rioter in a demonstration sees a policeman approaching, he must decide either to continue to throw objects or to run away.

3.2 Geo-spatial and physical coherence

The mechanism responsible for managing the spatial and physical coherence of the simulation consists in controlling, in a transparent and automatic way, the spatial and physical state of each object/agent. With the help of the PhysX Physical engine, the PLAMAGS execution engine ensures that the spatialized actions carried out by each of the objects/agents are available and valid. For example, if an object/agent attempts to move in a given direction, and if there is a risk of collision with an obstacle (a building wall, another object/agent, etc.), then the execution engine must calculate the present forces (friction, weight, etc.) and correct the "desired" movement of the component and transfer it to another acceptable position which respects the spatial and physical restrictions.

The execution engine is also responsible for producing the edge effects associated with the physical forces which affect the objects/agents. For example, if a collision occurs with an object/agent at a location where the ground is characterized by a certain slope, then dependently of the physical properties of the ground and of the object/agent (gravity, friction, weight, etc.), it is possible that the colliding object/agent will "start to slip". In this case, the execution engine (once again with the help of the Physical engine) is responsible for managing the unwanted movement of the object/agent and to notify it of its position change, at each iteration.

4. Composition of a simulation

This section presents an overview of the elements composing a PLAMAGS simulation (see Figure 3). Generally speaking, a PLAMAGS simulation is composed of a VGE (mainly the 3D model, the coordinates system, the structures containing the elevation maps, physical forces such as gravity), the objects and agents that evolve in the VGE, each of them having its particular capabilities, either visual, spatial or physical, that allow it to carry out its specific behaviors and to evolve in the VGE. In addition, a simulation includes a scenario which defines the initial state of the simulation as well as how it will unfold, taking into account particular conditions that may dynamically change the VGE during the simulation. PLAMAGS also natively supports agent groups' (behaviors and interactions) and the simulation of any kind of particle systems (such as tear gas).

Of all the elements that are composing a simulation, the main actors of a PLAMAGS simulation are the objects and simulation agents, which are in constant interactions with each other and with the VGE. The various PLAMAGS components are formally defined in the language. Here are the definitions of the main categories of components.

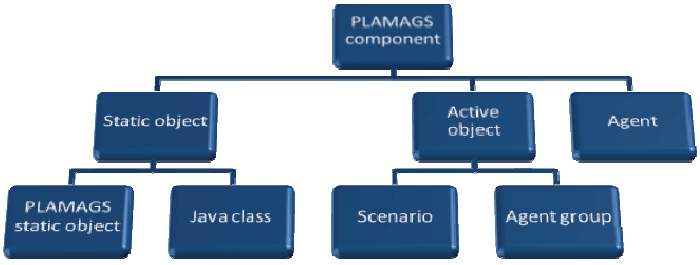


Fig. 10. Main categories of PLAMAGS components.

Components categories
PLAMAGS defines the three main categories of components: ***static object, active agent*** and ***agent*** which are distinguished by their different behavioural capabilities. Moreover, two sub-categories of the active objects are defined, whether the ***agent groups*** and the ***scenarios***. Also, two sub-categories of the static objects are defined: PLAMAGS static object and java class. The following sub-sections characterize each of the categories presented in figure 10.

4.1 Objects and agents

Whatever its nature (static object, active object, agent), the definition of a component type is minimally composed of a set of visual, spatial and physical characteristics, as described in section 3.1.1 and of a set of internal properties that can be either constant or dynamic. In addition, the designer can associate a set of actions to a component type.
The main difference between static objects, active objects and agents corresponds to the definition of the associated behaviors (Garneau et al. 2010). Static objects do not possess any behavior and therefore, are totally passive (they are used to represent objects such as trees, fences and street lamps). Active objects can be considered as reactive agents, their behaviors being defined using lists of powerful rules (Levesque et al. 2008). Moreover, agents possess the most expressive of PLAMAGS' behavioural structures as discussed in the next section.

4.2 Agent behaviors

Behavioral graphs are used to define complex agent's behaviors of simulations' agents. The power of these graphs lies in their high flexibility, customizability and their ability to represent behaviors in different ways.

Fig. 11. Composition of an objective

A PLAMAGS behavioral graph is a multi-layered graph in which nodes represent objectives which can be either atomic or composed of sub-behaviors. The objectives are organized in hierarchies such that elementary objectives (called *simple objectives*) are associated with actions that the agent can execute. Each agent owns a set of objectives corresponding to its needs (Moulin et al. 2003). An objective is associated with rules containing constraints

characterizing the activation, execution or completion of the objective: we call them *activation rules*, *execution rules* and *completion rules* (Moulin et al. 2003; Garneau et al. 2008). Constraints are dependent on time, on the agent's state, and on the environment's state. An objective is also related to resources that it either needs to acquire or already owns, these resources being required for the objective's execution. Figure 11 presents the different elements constituting an objective in a schematic way.
The selection of the current agent's objectives relies on the graph structure, on previously executed objectives and is influenced by required resources, the objectives' priorities, as well as by the associated activation/execution/completion rules (Garneau et al. 2010). The execution of an objective is always conditional to its execution rule being triggered and to the availability of the required resources. An objective's priority is primarily a discriminating function or expression which is used to choose the active objective among a set of potential objectives. It is also subject to modifications brought about by the opportunities that the agent perceives in the environment and by the temporal constraints applying to the objective. Resources are agents' assets that can be assigned exclusively to an objective's execution. The allocation of resources between objectives at a given iteration is subject to the objectives' priorities.

Control of each objective execution in many phases
Compared to the majority of graph-based models where the execution of a node (corresponding whether to a state, an objective or a goal.) is monolithic, the execution of a PLAMAGS objective (be it simple, composed or aggregated) is a flexible process associated with several specification steps that can be easily controlled (figure 12).

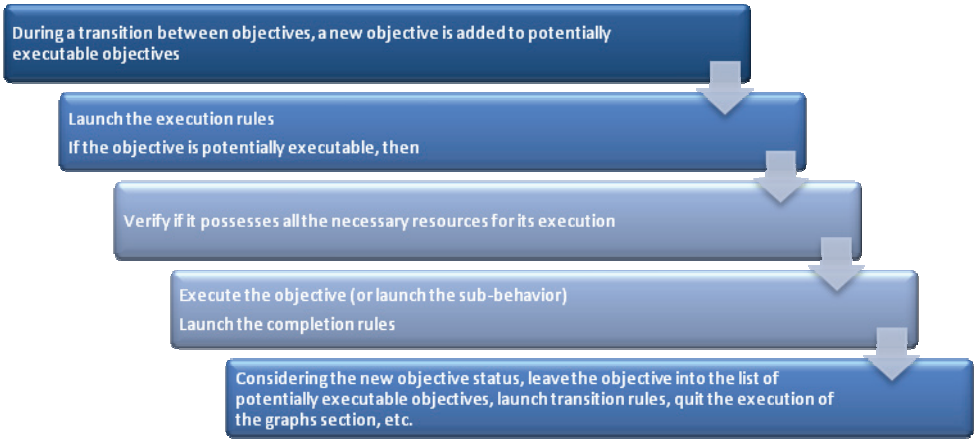


Fig. 12. Summary of an objective's execution dynamics.

The structure of the multi-layered directed graph allows us to define a behavior at different levels of abstraction and to divide behaviors into sub-behaviors. An abstraction level can be added by inserting a compound or an aggregate objective in the behavior.
A *compound objective* can be thought of as a decomposable structure representing a sub-behavior (its structure is similar to a behavior structure). *Aggregate objectives* are also decomposable structures: they are composed of a set of objectives which may not be related.

These objectives allow the designer to represent agents' objectives in which neither a hierarchical structure nor a predefined sequence of objectives is needed. Compound and aggregate objectives are well suited to regroup an agent's objectives as set of goals. Since "non-simple" objectives are composed of other objectives, any number of abstraction levels can be specified. The decomposition stops when an objective is composed of elementary actions which correspond to simple objectives: this corresponds to the "execution level" of the behavior.
Since agents often need to simultaneously achieve more than one objective, we provide an execution mode allowing to concurrently activate several objectives. The "mode" of declaration is specified for each objective because concurrent activation is not desirable everywhere in a behavior graph. This allows the designer to locally control the activation of parts of a behavior graph.
This graph structure and the PLAMAGS language associated offer several advantages.

- Intuitive modeling approach
- Executable specifications
- Abstraction and iterative refinement (based on the specification of sub-behaviors)
- Elimination of the translation step between a model definition and its implementation
- Concurrent execution of multi-layered objectives
- Automatic management of objectives' concurrency based on resources and priorities

4.2 Scenarios

Defining all the elements of a simulation is an error-prone and complex process. Since one of our aims in creating PLAMAGS is to guide the user when creating a simulation, we introduce a specific structure called "scenario" to specify in a structured way the various elements composing a simulation. Figure 13 presents the steps of the specification of a scenario.

Definition of VGE and execution properties
The scenario allows for the specification of the parameters of the VGE. Figure 13 presents the skeleton of a scenario that we discuss in the paragraph. In order to offer guidelines to the user, each of the VGE properties is defined by a specific key-word and associated value. Several other parameters need to be specified to configure the execution, the performance, the personalization, the interaction with external tools, the debugging as well as the optimization of the simulation. All these parameters are directly available in the scenario (within a specific block) and each of them possesses default values, minimally necessary to run the simulation. In this way, the user can concentrate on defining the properties he wishes to specify and can accept the default values for the properties that he is not interested in (see for example lines 7 and 8 in Figure 14).

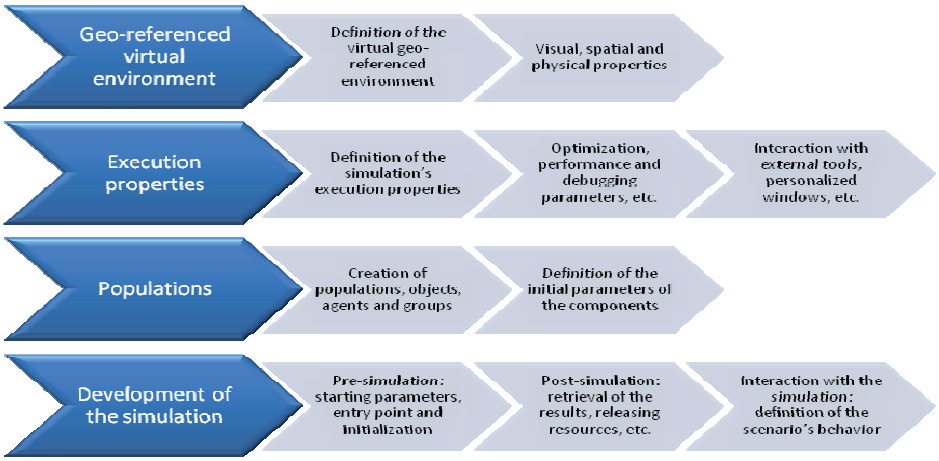


Fig. 13. Defined elements in a PLAMAGS scenario.

```
5   public scenario Scenario1
6
7       map coordinates : "0,500,0,900,0,100"
8       map mapModel ...
9       ...
10      component :
11          ...
12
13      rules Scenario1RB trigger every [50]
14
15      initialize call main()
16      initialize call println("Starting simulation...")
17
18      exit when [shouldStop()]
19
20      terminate call println("Ending simulation...")
21
22      logger debug = {
23              behaviour.algorithm,
24              behaviour.warning
25      }
26
27      method :
28
29          private void main(array<string> args)
30              for [i : int; i < args.size(); i = i + 1]
31                  ...
32              end for
33          end method
34
35          private boolean shouldStop()
36              ...
```

Fig. 14. Basic structure of a scenario.

Creation of populations and initial conditions

A block within the scenario (lines 10 and 11 in Figure 14) allow to instantiate (in a similar way as using calls of constructors in an object-oriented language) the objects and agents

populations that will participate in a simulation (static object, active object, gas, group, agent).

Preparation of the simulation (Pre-simulation)
The scenario offers the possibility to define the actions that will be executed only once before the execution of the simulation's main loop (main event loop), lines 15, 16 and 29 to 33 of figure 14, which is often necessary to carry out different initializations of a simulation.

Main execution loop
Once the pre-simulation is completed, the execution of the main loop is automatic and launches in an iterative way the behaviors of each active component of the simulation.

Interactions with the simulation
Since the scenario is itself a component, it can interact with the simulation. Its behavior is defined by a list of rules (line 13 of figure 14).

Ending of simulation
A scenario also allows for the definition of the actions that will be executed only when the simulation is over (line 20 in Figure 14). In this way, the designer can explicitly define and schedule the actions to be executed once the execution of the main loop simulation is over.

4.4 Gas

The PLAMAGS model allows to integrate components representing gases in a simulation using a type of active object directly supplied in PLAMAGS. Their behaviours are managed with the help of the PhysX Library (PhysX 2010) in the form of particle systems. The specification of a gas is quite flexible and parameterizable.

```
local smoke : Gas = constructor()

call smoke.setPosition(244,59,0)
call smoke.setParticuleImage("res/textures/smoke.jpg")
call smoke.setMaximumParticulesNumber(500)
call smoke.setParticulesSpeed(0,0,0.025)
call smoke.setEmissionAngleMargin(130.0)
call smoke.setEmissionSpeedMargin(0.0225)
call smoke.setEmissionPositionMargin(0.0025)
call smoke.setInitialParticulesRadius(0.03)
call smoke.setParticulesDissipationSpeed(0.006125)
call smoke.setMaximumParticulesRadius(0.35)
call smoke.setParticulesAcceleration(0,0,0.000635)
call smoke.setTransparency(50)
call smoke.setIsRelativeToMap(true)

call smoke.emit()
```

Fig. 15. Creation and parametration of a gas.

Figure 15 presents the code allowing for the creation of a gas representing a smoke cloud emanating from a small fire. The result is presented in Figure 16 as a series of fires.
The code presented in Figure 15 is all that is needed to create and parameterize a gas in PLAMAGS. Let us emphasize that it is usually not necessary to initialize all the parameters

of a gas, taking advantage of the default values offered by the PLAMAGS language. Different parameter sets can be used to create different kinds of gas: smoke clouds, toxic gases, stormy cloud cells, etc.

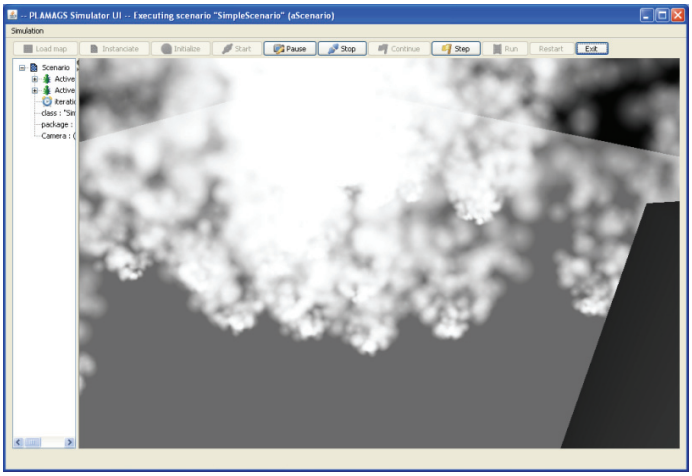


Fig. 16. Propagation of gases in PLAMAGS.

4.4 Groups

Agent groups possess a multitude of unique characteristics whose description is out of scope of this chapter. However, let us emphasize that the use of a group in PLAMAGS offers a set of functionalities, among which an important one is the possibility for a component's capability to perceive groups. The agents can also be perceived as belonging to groups (in addition to being perceived as individual agents). The groups inherit from the active agents' properties, and can be associated with a "group" behavior and all the associated rules. This does not prevent the agents belonging to a group to carry out their own behaviors.

4.5 Java classes

The extensibility of the PLAMAGS language allows for the use of any Java class in a simple way. Once a PLAMAGS type is defined from a Java class, this type is automatically defined as being a **static object** and is directly usable in a PLAMAGS specification. As a simple example, Figure 17 shows the use of the java.ioPrintWriter class in PLAMAGS.

```
private declarator JavaIO
      create type plamags.File : java.io.PrintWriter
end declarator

local file : plamags.File = constructor("data.txt")
call file.println("some text")
....
call file.close()
```

Fig. 17. Use of Java classes in PLAMAGS.

5. Language and IDE

This section is a quick overview of the general characteristics of the language (supported by the IDE), which is a complete and expressive agent-oriented language providing standard procedural and object-oriented (OO) features, as well as a bidirectional communication between PLAMAGS and Java (more details in (Garneau et al. 2010)). Although the language offers the majority of constructions available in procedural and OO languages, its power comes from its support of several structures dedicated to the specification of evolved agent behaviors. Furthermore, PLAMAGS allows a simple and transparent communication between the VGE and the agents, providing a set of tools that can be called directly from the language. This allows the user to directly integrate geographic and spatial information in the agent's decisional process.

5.1 Specification, definition, implementation and execution language

PLAMAGS is the first complete programming language for MAGS, totally dedicated to the implementation, but above all, usable in any phase of the MAGS development process (see Section 2).

Modeling, specification, definition and design
Starting with the first phases of the development process of a MAGS, PLAMAGS can be used to support the specification of the models thanks to its syntax which is both declarative and procedural. To this end, the language offers declarative syntactic blocks as well as a set of well chosen key-words.

Implementation
All the steps of the development process are supported by the language. Let us emphasize that the language, offers all the required structures to define and implement the dynamics of behavioral graphs and rules lists.

Execution and validation
The language is associated with: 1) an interpreter/compiler of the PLAMAGS syntax which translates the specification into an interpretable/compilable code and, 2) an execution engine which is responsible for managing the execution of the objects' and agents' behaviors. The execution engine is also in charge of ensuring the coherence of the executed behaviors (concurrent execution of the objectives, priorities, resources, objectives states, etc.) by doing a systematic validation of the executed objectives, of their status, etc.

Debugging
The language also supplies a set of flexible and configurable tools for debugging, notably a tracking system enabling the user to follow step by step the different decisions made by the behavioral execution engine, to get the execution result of each objective, etc.

Geo-spatial management and spatialized interactions
For an easy development of MAGSs, the language also integrates all the primitives necessary to specify the geo-spatial aspects of the simulation as well as mechanisms (syntax elements) to configure the interactions between agents, as well with the VGE.

Extensibility and versatility
Finally, to ensure that it is never limited by the syntax of the functionalities that it offers, the language allows the direct use of any external Java class or library.

5.2 IDE to support the method and the language use

The PLAMAGS language is supported by an IDE which is a complete development environment that allows for the quick development and execution of multi-agent geo-simulations. Figure 18 presents an overview of the IDE support of the PLAMAGS method.
The IDE provides: 1) a program editor (with real-time error checking); 2) a project management tree; 3) a contextual tree (describing the components of the file); 4) a language validation engine (similar to a compiler); 5) a runtime engine (an interpreter to run simulations in an interpreted mode); 6) a Java code generator and a compiler (to run simulations in a compiled mode); 7) a 3D engine to visualize the simulations; 8) a visual programming tool (to graphically develop behaviors); 9) an interface allowing to quickly create each kind of components in the shape of "stub" files; 10) an assistant allowing to create, modify, edit a component in a visual way (during development); 11) a visualization tool of the used Java classes; 12) an integrated documentation.
Moreover, to handle physics constraints (collisions, gas dispersion, repelling, friction, ground elevations, etc.) and to allow agents and objects to take into account these constraints in their decision making process, PLAMAGS uses one of the most powerful physics engine, PhysX.
All these tools are integrated in a completely transparent way into the IDE where they are correctly configured by default (compiler, interpreter, etc.). The user only has to create a project with the help of an automated assistant and to start the creation of the simulation. Then, he can, without any additional effort (no configuration is necessary), compile and execute and visualize any PLAMAGS program. Figure 19 presents a snapshot of the PLAMAGS IDE's interface.

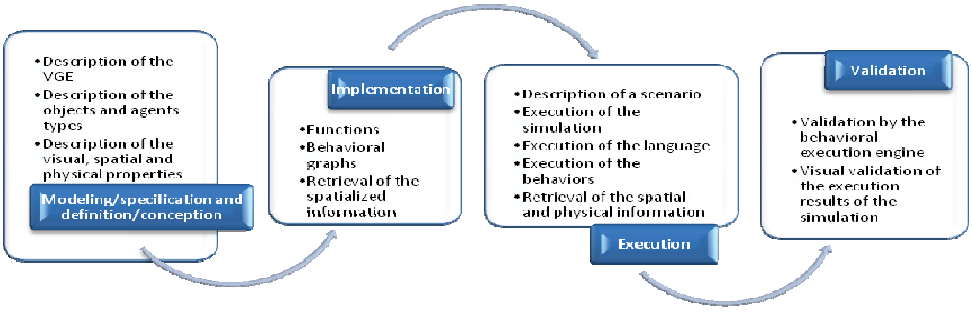


Fig. 18. PLAMAGS in all the steps of the development process of a MAGS.

6. Results and Conclusion

The PLAMAGS method distinguishes itself from the other MABS methods/tools by its framework that supports every step of the development cycle of a MAGS, from the modeling step to the validation step. This is made possible thanks to the language that

supports the modeling, the implementation and the execution of MAGS, while offering all the necessary mechanisms for the integration of the geo-spatial interactions between agents and with the VGE. In this way, PLAMAGS eliminates the transition and translation steps between the models and their implementations that are required by other MAGS specification approaches, which greatly reduces the development effort.

In order to show the large scope of PLAMAGS, we have meticulously replicated the experimentation presented in the « *Agent-based Simulation Platforms: Review and Development Recommendations* » article (Railsback et al. 2006) in the form of a 3D simulation. The results we obtained can be advantageously compared with those of the multi-agent based simulation tools used in Railsback's study: (MASON (Luke et al. 2004), SWARM (SWARM 2010), Java Swarm, Repast (North et al. 2007) and NetLogo (NetLogo 2010)). The results will be presented in a forthcoming paper, however we can already conclude that PLAMAGS could either be used for development of MABS (agents based simulations) or for the MAGS.

We used PLAMAGS in a variety of MAGS simulation projects and we have identified three main shortcomings for which we intend to find solutions in our future work. First, the definition of a simulation's initial properties (VGE, objects and agents), for a specific scenario, is a demanding task when performed at the programming level. A graphical specification tool would facilitate the user's task. Second, the JPCT library (JPCT 2010) used for 3D rendering seems to reach its performance limits when dealing with complex 3D models. Indeed, such models require too much time for graphical rendering and tend to use too much memory. This is why we use very simple models for the representation of objects and agents in our simulation examples. We may try to find a replacement library. Third, our experiments involving various simulations have shown that it would be very convenient to add a "save simulation state" function in order to save a simulation, modify one or several elements of its structure, and then resume the simulation in the state saved before the changes.

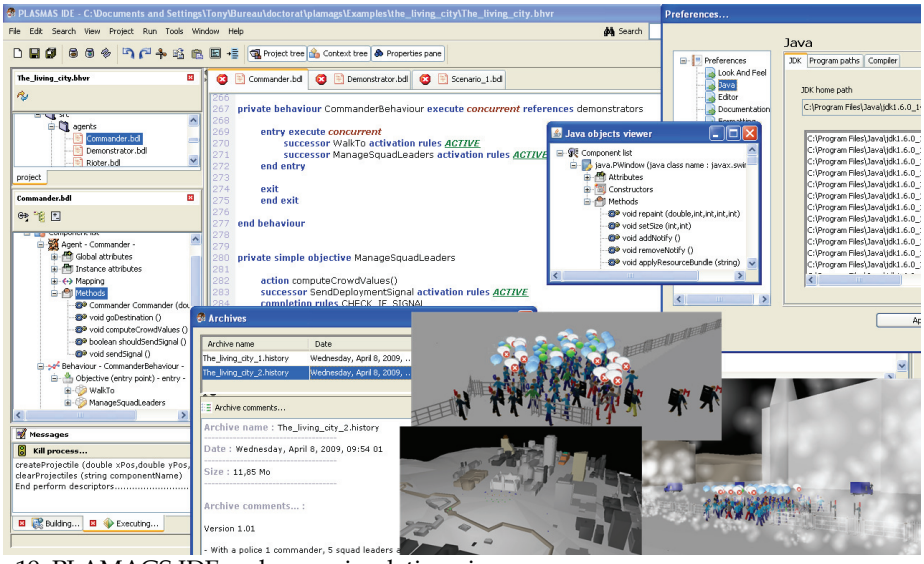


Fig. 19. PLAMAGS IDE and some simulation views.

Acknowledgments: Parts of this project have been financed by Geoide, the Canadian Network of Centers of Excellence in Geomatics. The first author also benefited from a scholarship from the Natural Sciences and Engineering Council of Canada.

7. References

AI.implant. (2009). from http://www.presagis.com/.

Benenson, I. and V. Kharbash (2005). Geographic Automata Systems:From The Paradigm to the Urban Modeling Software. AGILE 2005 and GISPlanet 2005, Estoril, Portugal.

Benenson, I. and P. M. Torrens (2004). "Geosimulation: object-based modeling of urban phenomena." Computers, Environment and Urban Systems **28**(1): 1-8.

Bourrel, E. and V. Henn (2003). Mixing micro and macro representations of traffic flow: a hybrid model based on the LWR theory. 82nd TRB Annual Meeting (Transportation Research Board), Washington, USA.

Cutumisu, M., D. Szafron, J. Schaeffer, M. McNaughton, T. Roy, C. Onuczko and M. Carbonaro (2006). "Generating Ambient Behaviors in Computer Role-Playing Games." IEEE Intelligent Systems **21**(5): 19-27.

d'Aquino, P., C. Le Page, F. Bousquet and A. Bah (2003). "Using Self-Designed Role-Playing Games and a Multi-Agent System to Empower a Local Decision-Making Process for Land Use Management: The SelfCormas Experiment in Senegal." Journal of Artificial Societies and Social Simulation **6**(3).

Donikian, S. (2001). HPTS: a behaviour modelling language for autonomous agents. International Conference on Autonomous Agents, fifth international conference on Autonomous agents, Montréal, Québec, Canada, ACM Press.

Fagiolo, G., A. Moneta and P. Windrum (2007). "A Critical Guide to Empirical Validation of Agent-Based Models in Economics: Methodologies, Procedures, and Open Problems." Computational Economics **30**(3): 195-226.

Foudil, C. and D. Noureddine (2007). "An autonomous and guided crowd in panic situations." Journal of Computer Science & Technology **2**(2): 134-140.

Fu, D., R. Houlette and S. Henke (2002). "Putting AI in Entertainment: An AI Authoring Tool for Simulation and Games." Intelligent Systems **17**(4): 81-84.

Garneau, T., B. Moulin and S. Delisle (2008). PLAMAGS: A Language and Environment to Specify Intelligent Agents in Virtual Geo-Referenced Worlds. Proceedings of the 19th IASTED International Conference on Modelling and Simulation., Quebec City, Canada.

Garneau, T., B. Moulin and S. Delisle (2010). Effective agent-based geosimulation development using PLAMAGS. Modelling, Simulation and Optimization. Intech**:** 684-708.

Gnansounou, E., S. Pierre, A. Quintero, J. Dong and A. Lahlou (2007). "Toward a Multi-Agent Architecture for Market Oriented Planning in Electricity Supply Industry." International Journal of Power and Energy Systems **27**(1): 82-91.

Guyot , P. and S. Honiden (2006). "Agent-Based Participatory Simulations: Merging Multi-Agent Systems and Role-Playing Games." Journal of Artificial Societies and Social Simulation **9**(4).

Helbing, D., A. Hennecke, V. Shvetsov and M. Treiber (2002). "Micro- and Macro-Simulation of Freeway Traffic." Mathematical and computer modelling **35**(5-5): 517-547.

JPCT. (2010). from http://www.jpct.net/.

Koch, A. (2001). Linking Multi Agent Systems And GIS - Modeling And Simulating Spatial InterActions –. Angewandte Geographische Informationsverarbeitung XII, Beiträge zum AGIT-Symposium.

Levesque, J., F. Cazzolato, J. Perron, J. Hogan, T. Garneau and B. Moulin (2008). CAMiCS: civilian activity modelling in constructive simulation. SpringSim 2008**:** 739-744.

Luke, S., C. Cioffi-Revilla, L. Panait and K. Sullivan (2004). MASON: A New Multi-Agent Simulation Toolkit. Eighth Annual Swarm Users/Researchers Conference, SwarmFest 2004, University of Michigan, Ann Arbor, Michigan USA.

Moulin, B., W. Chaker, J. Perron, P. Pelletier, J. Hogan and E. Gbei Fonh (2003). MAGS Project: Multi-Agent GeoSimulation and Crowd Simulation. Conference on Spatial Information Theory (COSIT'03), Ittingen, Switzerland, Springer-Verlag.

Müller, J.-P., C. Ratzé, F. Gillet and K. Stoffel (2005). Modeling And Simulating Hierarchies Using An Agent-Based Approach. MODSIM05 : International Congress on Modelling and Simulation. Advances and Applications for Management and Decision Making Melbourne, Australia.

NetLogo. (2010). "NetLogo." from http://ccl.northwestern.edu/netlogo/.

North, M. J., E. Tatara, N. T. Collier and J. Ozik (2007). Visual Agent-based Model Development with Repast Simphony. Agent 2007 Conference on Complex Interaction and Social Emergence, Argonne National Laboratory, Argonne, IL USA.

Papazoglou, P. M., D. A. Karras and R. C. Papademetriou (2008). On the Multi-threading Approach of Efficient Multi-agent Methodology for Modelling Cellular Communications Bandwidth Management Agent and Multi-Agent Systems: Technologies and Applications. **4953/2008**.

PATHEngine. (2009). from http://www.pathengine.com/.

PhysX. (2010). 2008, from http://www.nvidia.com/object/physx_9.09.0408_whql.html.

Railsback, S. F., S. L. Lytinen and S. K. Jackson (2006). "Agent-based Simulation Platforms: Review and Development Recommendations." SIMULATION **62**(9): 609-623.

SPR.OPS. (2009). from http://www.spirops.com/.

SWARM. (2010). "SWARM." from http://www.swarm.org/.

Torrens, P. M. and I. Benenson (2005). "Geographic Automata Systems." International Journal of Geographical Information Science **19**(4): 385-412.

Closed-form Solutions of the Cross-anisotropic Stratum Due to a Point Heat Source

Feng-Tsai Lin[1] and John C.-C. Lu[2]
[1]*Department of Naval Architecture, National Kaohsiung Marine University*
[2]*Department of Civil Engineering, Chung Hua University*
Taiwan

1. Introduction

The objective of this paper is to present the closed-form solutions of the long-term displacements and temperature change of a cross-anisotropic medium subjected to a point heat source at great depth. The medium is first assumed to be cross-anisotropic in mechanical and thermal properties. Under this assumption, the properties of the materials are different in plane of isotropy and in planes normal to it. Using Hankel and Fourier transforms, this paper presents the analytic solutions to this kind of problems, such as the repositories of nuclear wastes. The general solutions are then further simplified to cases when the material is isotropic in mechanical properties and finally totally isotropic in each property of the stratum.

Nuclear wastes are usually deposited at a great depth, such as 200 to 700 meters below ground, so that they can be isolated from the living environment of human beings. However, research efforts on the thermo-mechanical responses of soils and rocks due to the heat generated by the radioactive waste in deep underground are still very limited. Excessive thermal difference usually results in a volume change of water and solid skeleton. This change can cause an increase in excess pore water pressure, and in turn a decrease in effective stress, which can result in a thermal failure in the stratum due to the loss of shear resistance of solid skeleton.

The governing equations proposed by Biot (1941, 1955) for a fluid-saturated poroelastic solid in an isothermal quasi-static state also provided an excellent insight into a variety of mechanical phenomena. His theory was later re-formulated by Rice and Cleary (1976). Schiffman (1971) extended Biot's theory to take the thermal effects into account. The solutions to the thermal consolidation of a saturated elastic porous media around a point heat source were presented by Booker and Savvidou (1984, 1985), Savvidou and Booker (1989). In their solutions, the flow is considered to be isotropic (Booker and Savvidou, 1984, 1985) or cross-anisotropic (Savvidou and Booker, 1989) whereas the mechanical and thermal properties of the stratum are treated as being isotropic. However, it was found that the anisotropic property in the permeability of the soils has significant effects on the excess pore

water pressure generated by a heat source (Savvidou and Booker, 1989). Lu and Lin (2006) displayed transient ground surface displacement produced by a point heat source/sink through analog quantities between poroelasticity and thermoelasticity. Based on Biot's three-dimensional consolidation theory of porous media, analytical solutions of the transient thermo-consolidation deformation due to a point heat source buried in a saturated isotropic porous elastic half space were presented by Lu and Lin (2007), Lin and Lu (2009). Within the framework of the linear theory of thermoelasticity, Chao, Chen and Shen (2006) discussed the problem of circularly cylindrical layered media subjected to an arbitrary point heat source.

Soils in general are deposited through a process of sedimentation over a long period of time. Under the accumulative overburden pressure, soils display significant anisotropy on mechanical, seepage and thermal properties. Both the soil and the stratified rock masses show the nature of anisotropy. For this reason, theoretical or numerical models should be able to simulate this kind of layered soils and rocks as cross-anisotropic medium (Amadei *et al.*, 1988; Barden, 1963; Gibson, 1974; Lee & Yang, 1998; Sekowski, 1986; Sheorey, 1994).

In this paper, the soil mass is modelled as a linearly elastic medium with cross-anisotropic properties. Both the thermal flow and the mechanical properties are assumed to be cross-anisotropic. By using the Hankel and Fourier transforms, closed-form solutions of the long-term displacements and temperature change of the stratum due to a point heat source at large depth are obtained. The results are reduced to an isotropic case to provide a better understanding of the thermally induced responses of the stratum.

2. Mathematical Model

2.1 Basic Equations

Figure 1 shows a point heat source buried in a stratum at a great depth. Consider a homogeneous layer of cross-anisotropic soil or rock. For simplicity, the plane of symmetry of the stratum is in the horizontal direction. Let (r,θ,z) be a cylindrical coordinate system for this layer of solid where the plane of isotropy coincides with the horizontal (or $r-\theta$) plane. Let u_r and u_z be the displacements in the radial and vertical directions, respectively. The constitutive law for an elastic medium with linear axisymmetric deformation can thus be expressed by

$$\sigma_{rr} = A\frac{\partial u_r}{\partial r} + (A-2N)\frac{u_r}{r} + F\frac{\partial u_z}{\partial z} - \beta_r \vartheta \,, \tag{1a}$$

$$\sigma_{\theta\theta} = (A-2N)\frac{\partial u_r}{\partial r} + A\frac{u_r}{r} + F\frac{\partial u_z}{\partial z} - \beta_r \vartheta \,, \tag{1b}$$

$$\sigma_{zz} = F\frac{\partial u_r}{\partial r} + F\frac{u_r}{r} + C\frac{\partial u_z}{\partial z} - \beta_z \vartheta \,, \tag{1c}$$

$$\sigma_{rz} = L\left(\frac{\partial u_r}{\partial z} + \frac{\partial u_z}{\partial r}\right), \tag{1d}$$

where σ_{rr}, $\sigma_{\theta\theta}$, σ_{zz} and σ_{rz} are the stress components, ϑ is the temperature change of the stratum, and A, C, F, L, N are the material constants of a cross-anisotropic medium defined by Love (1944). In these equations, β_r and β_z represent the thermal expansion

factors along and normal to the symmetric plane, respectively. For symmetric problem, it can be noted that the shear stresses $\sigma_{r\theta}$, $\sigma_{\theta z}$, and circumferential displacement u_θ would vanish as the vertical z-axis is located through the point heat source.

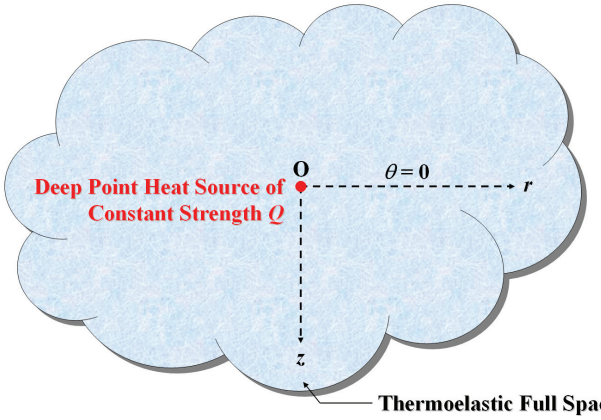


Fig. 1. Point heat source of constant heat generation rate buried deep in a cross-anisotropic stratum

Let E_r and E_z be Young's moduli with respect to directions lying in the plane of isotropy and perpendicular to it, respectively; $\nu_{r\theta}$ be Poisson's ratio for strain in the horizontal direction due to a horizontal direct stress; ν_{rz} be Poisson's ratio for strain in the vertical direction due to a horizontal direct stress; ν_{zr} be Poisson's ratio for strain in the horizontal direction due to a vertical direct stress; and G_{rz} be shear modulus for planes normal to the plane of isotropy. Equations (1a)-(1d) can then be converted to

$$\begin{Bmatrix} \dfrac{\partial u_r}{\partial r} \\ \dfrac{u_r}{r} \\ \dfrac{\partial u_z}{\partial z} \\ \dfrac{\partial u_r}{\partial z}+\dfrac{\partial u_z}{\partial r} \end{Bmatrix} = \begin{bmatrix} \dfrac{1}{E_r} & -\dfrac{\nu_{r\theta}}{E_r} & -\dfrac{\nu_{rz}}{E_r} & 0 \\ -\dfrac{\nu_{r\theta}}{E_r} & \dfrac{1}{E_r} & -\dfrac{\nu_{rz}}{E_r} & 0 \\ -\dfrac{\nu_{zr}}{E_z} & -\dfrac{\nu_{zr}}{E_z} & \dfrac{1}{E_z} & 0 \\ 0 & 0 & 0 & \dfrac{1}{G_{rz}} \end{bmatrix} \begin{Bmatrix} \sigma_{rr} \\ \sigma_{\theta\theta} \\ \sigma_{zz} \\ \sigma_{rz} \end{Bmatrix} + \begin{Bmatrix} \alpha_{sr}\vartheta \\ \alpha_{sr}\vartheta \\ \alpha_{sz}\vartheta \\ 0 \end{Bmatrix}, \tag{2}$$

where α_{sr} and α_{sz} are the linear thermal expansion coefficients of the stratum in the horizontal and vertical directions, respectively. By comparison, the mechanical and thermal constants employed in equations (1a)-(1d) and (2) are related through the following equations:

$$A=\frac{E_r\left(1-\nu_{rz}\nu_{zr}\right)}{\left(1+\nu_{r\theta}\right)\left(1-\nu_{r\theta}-2\nu_{rz}\nu_{zr}\right)}, \tag{3a}$$

$$C=\frac{E_z\left(1-\nu_{r\theta}\right)}{1-\nu_{r\theta}-2\nu_{rz}\nu_{zr}}, \tag{3b}$$

$$F=\frac{E_z\nu_{rz}}{1-\nu_{r\theta}-2\nu_{rz}\nu_{zr}}=\frac{E_r\nu_{zr}}{1-\nu_{r\theta}-2\nu_{rz}\nu_{zr}}, \tag{3c}$$

$$L=G_{rz}, \tag{3d}$$

$$N=\frac{E_r}{2\left(1+2\nu_{r\theta}\right)}, \tag{3e}$$

$$\beta_r=2\left(A-N\right)\alpha_{sr}+F\alpha_{sz}, \tag{3f}$$

$$\beta_z=2F\alpha_{sr}+C\alpha_{sz}. \tag{3g}$$

For the case of isotropy, $A=C=\lambda+2G$, $F=\lambda$, $L=N=G$, and $\beta_r=\beta_z=\left(2G+3\lambda\right)\alpha_s$. Here, λ and G are the Lame moduli of the medium and α_s is the linear thermal expansion coefficient of the solid skeleton.

For a general problem, these stresses must satisfy the equilibrium equations $\sigma_{ij,j}+f_i=0$, where f_i denotes the body forces. For axisymmetric problems and neglect the effects of body forces, the equilibrium equations can be expressed in terms of displacements u_i and temperature change of the medium ϑ as follows:

$$A\left(\frac{\partial^2 u_r}{\partial r^2}+\frac{1}{r}\frac{\partial u_r}{\partial r}-\frac{u_r}{r^2}\right)+L\frac{\partial^2 u_r}{\partial z^2}+\left(F+L\right)\frac{\partial^2 u_z}{\partial r\partial z}-\beta_r\frac{\partial\vartheta}{\partial r}=0, \tag{4a}$$

$$\left(F+L\right)\left(\frac{\partial^2 u_r}{\partial r\partial z}+\frac{1}{r}\frac{\partial u_r}{\partial z}\right)+L\left(\frac{\partial^2 u_z}{\partial r^2}+\frac{1}{r}\frac{\partial u_z}{\partial r}\right)+C\frac{\partial^2 u_z}{\partial z^2}-\beta_z\frac{\partial\vartheta}{\partial z}=0. \tag{4b}$$

Using the law of conservation of energy, the equation can be obtained as listed below.

$$-\nabla\cdot\boldsymbol{h}+q_h=0, \tag{5}$$

where $\boldsymbol{h}$ is the heat flux vector and q_h is the internal (or external) heat sources.

To describe the behavior of the heat flow in a cross-anisotropic medium, let λ_{tr} denote the horizontal thermal conductivity of heat flow in the planes of isotropy and let λ_{tz} be the corresponding vertical thermal conductivity in the plane perpendicular to the isotropic plane. Assuming that the heat flow follows Fourier's law, then

$$\boldsymbol{h}=-\lambda_{tr}\frac{\partial\vartheta}{\partial r}\boldsymbol{i}_r-\lambda_{tz}\frac{\partial\vartheta}{\partial z}\boldsymbol{i}_z, \tag{6}$$

in which $\boldsymbol{i}_r$ and $\boldsymbol{i}_z$ are unit vectors parallel to the radial and vertical directions, respectively. Consider a point heat source of strength Q located at point $(0,0)$ at great depth. Substituting (6) into (5) yields the third governing equation to relate ϑ as listed below:

$$\lambda_{tr}\left(\frac{\partial^2\vartheta}{\partial r^2}+\frac{1}{r}\frac{\partial\vartheta}{\partial r}\right)+\lambda_{tz}\frac{\partial^2\vartheta}{\partial z^2}+\frac{Q}{2\pi r}\delta(r)\delta(z)=0\,, \tag{7}$$

where $\delta(r)$ and $\delta(z)$ are the Dirac delta functions.

For a linearly elastic medium with cross-anisotropic properties, the differential equations expressed by Eqs. (4a), (4b) and (7) govern the steady state response of the medium subjected to axisymmetric and thermoelastic disturbance.

2.2 Boundary Conditions

Assume that the point heat source at great depth has no effect on the ground surface. This implies that the ground surface can be treated as a remote boundary and the stratum can be modeled as an infinite space. Thus the effect of the deep thermally disturbance vanishes at the remote boundaries $z\to\pm\infty$. In other words, the displacements in the radial and vertical directions, and the temperature change of the stratum at remote boundaries should be vanished. Then the remote boundary conditions can be expressed by

$$u_r(r,z)\to 0\,,\ u_z(r,z)\to 0\,,\text{ and }\vartheta(r,z)\to 0\ \text{ as }\ z\to\pm\infty\,. \tag{8}$$

The thermoelastic responses can be derived from the differential equations (4a), (4b) and (7) corresponding with the remote boundary conditions at $z\to\pm\infty$.

3. Analytic Solutions

3.1 Hankel Transform Solutions

The governing partial differential equations (4a), (4b) and (7) can be simplified to the ordinary differential equations by performing appropriate Hankel transforms (Sneddon, 1951) with respect to the radial coordinate r of first, zeroth and zeroth orders, respectively. Therefore, we obtain

$$-\xi^2 AU_r+L\frac{d^2U_r}{dz^2}-\xi(F+L)\frac{dU_z}{dz}+\xi\beta_r\Theta=0\,, \tag{9a}$$

$$\xi(F+L)\frac{dU_r}{dz}-\xi^2 LU_z+C\frac{d^2U_z}{dz^2}-\beta_z\frac{d\Theta}{dz}=0\,, \tag{9b}$$

$$-\xi^2\lambda_{tr}\Theta+\lambda_{tz}\frac{d^2\Theta}{dz^2}=-\frac{Q}{2\pi}\delta(z)\,, \tag{9c}$$

where

$$U_r(z;\xi)=\int_0^\infty r\,u_r(r,z)\,J_1(\xi r)\,dr\,, \tag{10a}$$

$$U_z(z;\xi)=\int_0^\infty r u_z(r,z)\,J_0(\xi r)\,dr\,, \tag{10b}$$

$$\Theta(z;\xi)=\int_0^\infty r\,\vartheta(r,z)\,J_0(\xi r)\,dr\,. \tag{10c}$$

In these equations, $J_\alpha(x)$ represents the Bessel's function of the first kind of order α. The displacements in the radial and vertical directions, and the temperature change of the stratum can then be obtained by inverting the equations (10a) to (10c), respectively, as shown below.

$$u_r(r,z)=\int_0^\infty \xi U_r(z;\xi)\,J_1(\xi r)\,d\xi\,, \tag{11a}$$

$$u_z(r,z)=\int_0^\infty \xi U_z(z;\xi)\,J_0(\xi r)\,d\xi\,, \tag{11b}$$

$$\vartheta(r,z)=\int_0^\infty \xi\Theta(z;\xi)\,J_0(\xi r)\,d\xi\,. \tag{11c}$$

Now further consideration to perform the Fourier transformations (Sneddon, 1951) with respect to the axial coordinate z on equations (9a) to (9c). The results can be expressed as

$$-\left(\xi^2 A+\omega^2 L\right)\tilde{U}_r-i\omega\xi\left(F+L\right)\tilde{U}_z+\xi\beta_r\tilde{\Theta}=0\,, \tag{12a}$$

$$i\omega\xi\left(F+L\right)\tilde{U}_r-\left(\xi^2 L+\omega^2 C\right)\tilde{U}_z-i\omega\beta_z\tilde{\Theta}=0\,, \tag{12b}$$

$$\left(\xi^2\lambda_{tr}+\omega^2\lambda_{tz}\right)\tilde{\Theta}=\frac{Q}{2\pi}\,, \tag{12c}$$

where

$$\left\{\tilde{U}_r(\xi,\omega),\tilde{U}_z(\xi,\omega),\tilde{\Theta}(\xi,\omega)\right\}=\int_{-\infty}^{\infty}\left\{U_r(z;\xi),U_z(z;\xi),\Theta(z;\xi)\right\}e^{i\omega z}\,dz\,. \tag{13}$$

The closed-form solutions of the long-term thermoelastic deformations and temperature change of the cross-anisotropic medium subjected to a deep point heat source can then be easily obtained in the integral transformed domain (ξ,ω) by solving the simultaneous algebraic equations of (12a) to (12c). The results are shown as follows:

$$\tilde{U}_r(\xi,\omega)=\frac{Q}{2\pi}\frac{\xi\left\{\xi^2 L\beta_r-\left[\left(F+L\right)\beta_z-C\beta_r\right]\omega^2\right\}}{\left(\xi^2\lambda_{tr}+\omega^2\lambda_{tz}\right)\left\{CL\omega^4+\left[AC-F\left(F+2L\right)\right]\xi^2\omega^2+AL\xi^4\right\}}\,, \tag{14a}$$

$$\tilde{U}_z(\xi,\omega)=\frac{Q}{2\pi}\frac{i\omega\left\{-\omega^2 L\beta_z+\left[\left(F+L\right)\beta_r-A\beta_z\right]\xi^2\right\}}{\left(\xi^2\lambda_{tr}+\omega^2\lambda_{tz}\right)\left\{CL\omega^4+\left[AC-F\left(F+2L\right)\right]\xi^2\omega^2+AL\xi^4\right\}}\,, \tag{14b}$$

$$\tilde{\Theta}(\xi,\omega)=\frac{Q}{2\pi}\frac{1}{\xi^2\lambda_{tr}+\omega^2\lambda_{tz}}\,. \tag{14c}$$

These solutions can also be expressed in the domain $(z;\xi)$ by applying the inverse of the Fourier transforms to Eqs. (14a) to (14c), or in mathematic language,

$$\left\{U_r(z;\xi),U_z(z;\xi),\Theta(z;\xi)\right\}=\frac{1}{2\pi}\int_{-\infty}^{\infty}\left\{\tilde{U}_r(\xi,\omega),\tilde{U}_z(\xi,\omega),\tilde{\Theta}(\xi,\omega)\right\}e^{-i\omega z}d\omega\,. \tag{15}$$

With the help of any mathematical handbooks (Erdelyi *et al.*, 1954; Gradshteyn & Ryzhik, 1980) and some calculations, the solutions can be derived in the space domain (r,z) by applying the inverse of the Hankel transforms to Eq. (15). The results are shown as following:

$$u_r(r,z)=\frac{Q}{4\pi\lambda_{tz}}\left(a_1\frac{r}{R_1^*}+a_2\frac{r}{R_2^*}+a_3\frac{r}{R_3^*}\right), \tag{16a}$$

$$u_z(r,z)=\frac{Q}{4\pi\lambda_{tz}}\left(b_1\sinh^{-1}\frac{\mu_1 z}{r}+b_2\sinh^{-1}\frac{\mu_2 z}{r}+b_3\sinh^{-1}\frac{\mu_3 z}{r}\right), \tag{16b}$$

$$\vartheta(r,z)=\frac{Q}{4\pi\lambda_{tz}}\frac{1}{\mu_3 R_3}\,. \tag{16c}$$

In these equations, $R_i=\sqrt{r^2+\mu_i^2z^2}$ and $R_i^*=R_i+\mu_i|z|$ $(i=1,2,3)$. The constants a_i and b_i $(i=1,2,3)$ are defined by

$$a_1=\frac{L\beta_r+\left[(F+L)\beta_z-C\beta_r\right]\mu_1^2}{CL\mu_1\left(\mu_1^2-\mu_2^2\right)\left(\mu_1^2-\mu_3^2\right)}, \tag{17a}$$

$$a_2=\frac{L\beta_r+\left[(F+L)\beta_z-C\beta_r\right]\mu_2^2}{CL\mu_2\left(\mu_2^2-\mu_1^2\right)\left(\mu_2^2-\mu_3^2\right)}, \tag{17b}$$

$$a_3=\frac{L\beta_r+\left[(F+L)\beta_z-C\beta_r\right]\mu_3^2}{CL\mu_3\left(\mu_3^2-\mu_1^2\right)\left(\mu_3^2-\mu_2^2\right)}, \tag{17c}$$

$$b_1=\frac{L\beta_z\mu_1^2+(F+L)\beta_r-A\beta_z}{CL\left(\mu_1^2-\mu_2^2\right)\left(\mu_1^2-\mu_3^2\right)}, \tag{17d}$$

$$b_2=\frac{L\beta_z\mu_2^2+(F+L)\beta_r-A\beta_z}{CL\left(\mu_2^2-\mu_1^2\right)\left(\mu_2^2-\mu_3^2\right)}, \tag{17e}$$

$$b_3=\frac{L\beta_z\mu_3^2+(F+L)\beta_r-A\beta_z}{CL\left(\mu_3^2-\mu_1^2\right)\left(\mu_3^2-\mu_2^2\right)}\,. \tag{17f}$$

In addition, μ_1 and μ_2 must satisfy the following characteristic equation:

$$CL\mu^4-\left[AC-F(F+2L)\right]\mu^2+AL=0\,, \tag{18}$$

and $\mu_3=\sqrt{\lambda_{tr}/\lambda_{tz}}$.

3.2 Cases of Isotropic Mechanical Behavior with Cross-anisotropic Thermal Properties

The displacement and temperature change for a stratum with cross-anisotropic properties in mechanical and heat flows are analytically solved as expressed in Eqs. (16a)-(16c) under the disturbance of a deep point heat source. For the special case when the isotropic mechanical properties of the medium is introduced, the related closed-form solutions are obtained by taking the limit conditions of $\mu_1 = \mu_2 = 1$ for Eqs. (16a)-(16c). This is carried out by using of L'Hospital's rule and careful calculations. The results are given as follows:

$$u_r(r,z) = \frac{Q}{4\pi\eta G\lambda_{tz}}\left\{ \beta_r^*\varphi_1(r,z) + \left[2\eta\beta_r^* - (2\eta-1)\beta_z^* \right]\varphi_2(r,z) \right\}, \tag{19a}$$

$$u_z(r,z) = \frac{Q}{4\pi\eta G\lambda_{tz}}\left\{ \beta_z^*\varphi_3(r,z) + \left[2\eta\beta_z^* - (2\eta-1)\beta_r^* \right]\varphi_4(r,z) \right\}, \tag{19b}$$

$$\vartheta(r,z) = \frac{Q}{4\pi\lambda_{tz}}\frac{1}{\mu_3 R_3}, \tag{19c}$$

where $\eta = (1-\nu)/(1-2\nu)$ and

$$\beta_r^* = 2G(\alpha_{sr} + \nu\alpha_{sz})/(1-2\nu), \tag{20a}$$

$$\beta_z^* = 2G\left[2\nu\alpha_{sr} + (1-\nu)\alpha_{sz} \right]/(1-2\nu). \tag{20b}$$

The functions $\varphi_i\,(i=1,2,3,4)$ in Eqs. (19a) and (19b) are functions of space variables r and z and defined as:

$$\varphi_1(r,z) = \frac{1}{4(\mu_3^2-1)}\frac{r}{R} - \frac{1}{2(\mu_3^2-1)^2}\frac{r}{R^*} + \frac{1}{2\mu_3(\mu_3^2-1)^2}\frac{r}{R_3^*}, \tag{21a}$$

$$\varphi_2(r,z) = \frac{1}{4(\mu_3^2-1)}\left(-\frac{r|z|}{RR^*} + \frac{r}{R^*} \right) + \frac{1}{2(\mu_3^2-1)^2}\frac{r}{R^*} - \frac{\mu_3}{2(\mu_3^2-1)^2}\frac{r}{R_3^*}, \tag{21b}$$

$$\varphi_3(r,z) = -\frac{1}{4(\mu_3^2-1)}\frac{z}{R} - \frac{\mu_3^2}{2(\mu_3^2-1)^2}\sinh^{-1}\frac{z}{r} + \frac{\mu_3^2}{2(\mu_3^2-1)^2}\sinh^{-1}\frac{\mu_3 z}{r} + \frac{\mu_3^2}{2(\mu_3^2-1)^2}\sinh^{-1}\frac{\mu_3 z}{r}, \tag{21c}$$

$$\varphi_4(r,z) = \frac{1}{4(\mu_3^2-1)}\frac{z}{R} + \frac{1}{2(\mu_3^2-1)^2}\sinh^{-1}\frac{z}{r} - \frac{1}{2(\mu_3^2-1)^2}\sinh^{-1}\frac{\mu_3 z}{r}, \tag{21d}$$

where $R = \sqrt{r^2+z^2}$ and $R^* = \sqrt{r^2+z^2} + |z|$.

3.3 Cases of Isotropic Mechanical and Thermal Properties

Furthermore, the closed-form solutions for the special case when the properties of the medium is isotropic in mechanics and heat flows are acquired by taking the limit conditions of $\mu_3 = 1$ for Eqs. (19a) to (19c). Applying the L'Hospital's rule and careful calculations, the results are given as below:

$$u_r(r,z) = \frac{Q}{8\pi\lambda_t}\frac{(1+\nu)\alpha_s}{1-\nu}\frac{r}{R}, \tag{22a}$$

$$u_z(r,z) = \frac{Q}{8\pi\lambda_t}\frac{(1+\nu)\alpha_s}{1-\nu}\frac{z}{R}, \tag{22b}$$

$$\vartheta(r,z) = \frac{Q}{4\pi\lambda_t}\frac{1}{R}, \tag{22c}$$

where λ_t and ν represent the thermal conductivity and Poisson's ratio of the isotropic medium, respectively. It is noted from Eqs. (22a)-(22c) that the long-term horizontal displacement, vertical displacement and temperature increment of the stratum are not directly dependent on the shear modulus of the isotropic stratum. However, horizontal displacement and vertical displacement are dependent on the shear modulus of the cross-anisotropic full space as shown in equations (16a)-(16b) or (19a)-(19b).

4. Numerical Results

To study the effect of anisotropy on displacements and temperature increment of the stratum due to a point heat source, numerical results have been obtained for different sets of thermoelastic constants appropriate for soils. The thermoelastic constants used are summarized in Table 1.

Case	$\nu_{r\theta}$	ν_{rz}	G_{rz}/E_z	E_r/E_z	α_{sr}/α_{sz}	$\lambda_{tr}/\lambda_{tz}$	Reference
Case 1	0.25	0.25	0.4	1.0	1.0*	1.0*	Booker & Carter (1986)
Case 2	0.25	0.25	0.4	1.0	10.0*	10.0*	Booker & Carter (1986)
Case 3	0	0.38	0.38	1.84	10.0*	10.0*	Lee & Rowe (1989)

Table 1. Material properties of cross-anisotropic soils (*assumed values)

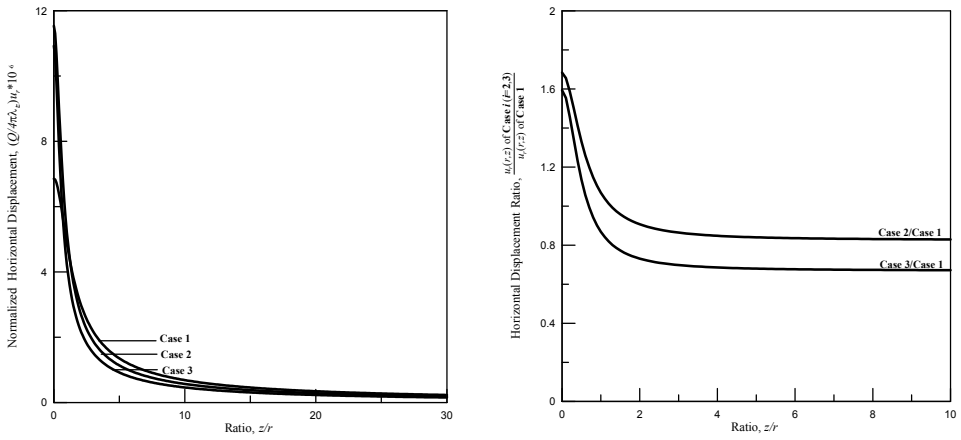
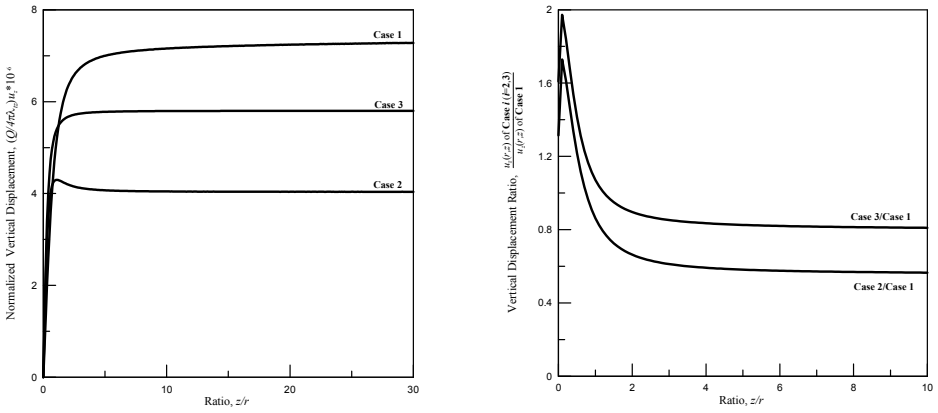


Fig. 2. Influence of anisotropy on horizontal displacement of the stratum

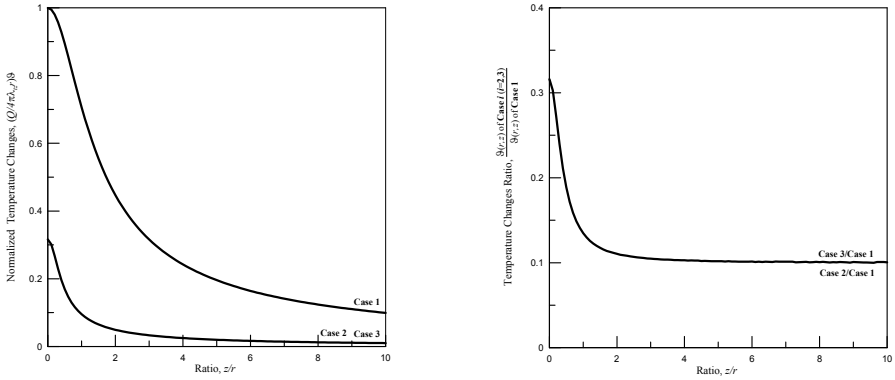
An indication of the influence of anisotropy on the thermoelastic responses are given in Figures 2-4. In the figures, the thermoelastic responses have been normalized. It is observed from Figures 2-4 that the anisotropy of the soil has significant effect on thermally elastic responses in comparison with the results obtained for an isotropic soil of case 1. For example, the vertical displacement of case 2 is reduced to about 60 percent of the corresponding value for the isotropic soil of case 1.

Figures 5-7 illustrate the horizontal displacement as effected by the anisotropy of the soil. As shown in Figure 5, the ratio E_r/E_z ranges from 0.5 to 10.0, and the effect of E_r/E_z on horizontal displacement of the stratum is secondary. Based on the available data, $\nu_{r\theta}=0.00$, $\nu_{rz}=0.38$, $G_{rz}/E_z=0.38$, and $\lambda_{tr}/\lambda_{tz}=1.0$ or 10.0, Figure 6 uses the ratio of α_{sr}/α_{sz} to display the influence on horizontal displacement of the stratum. It is shown from Figure 7 that the degree of anisotropic thermal conductivity $\lambda_{tr}/\lambda_{tz}$ has the most significant effect on horizontal displacement of the stratum due to a point heat source.



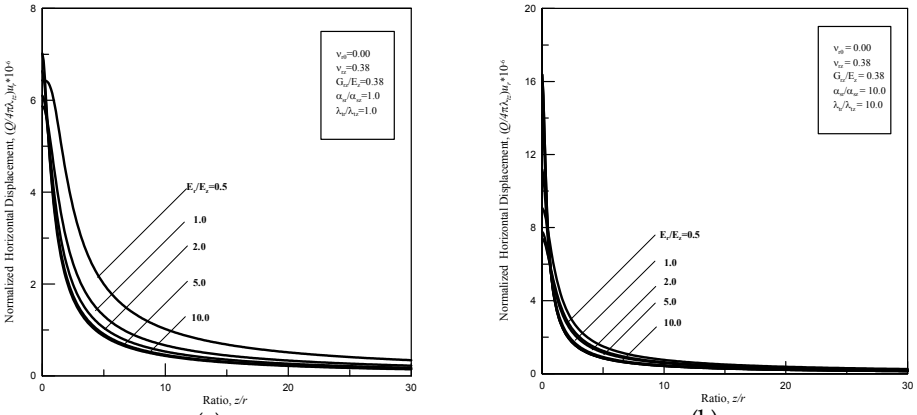
(a) (b)

Fig. 3. Influence of anisotropy on vertical displacement of the stratum



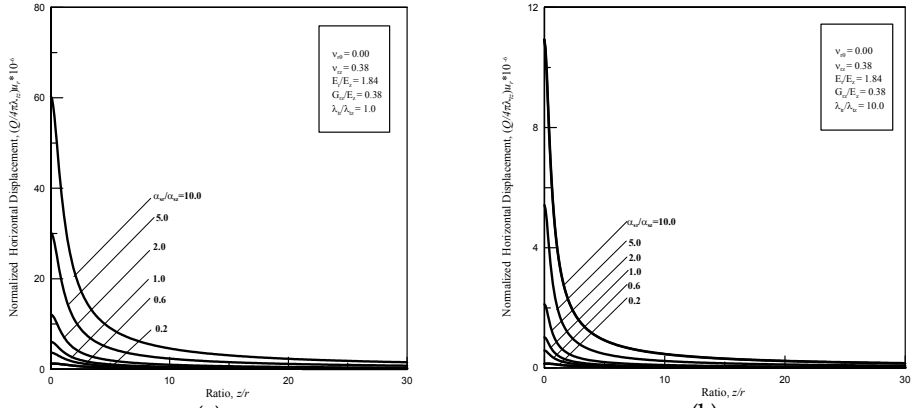
(a) (b)

Fig. 4. Influence of anisotropy on the temperature increment of the stratum



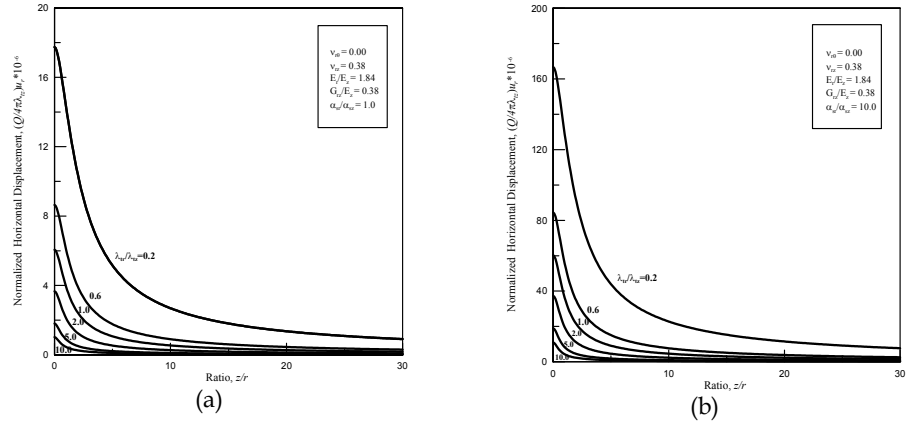
(a) (b)

Fig. 5. Influence of the degree of anisotropy E_r/E_z on horizontal displacement of the stratum



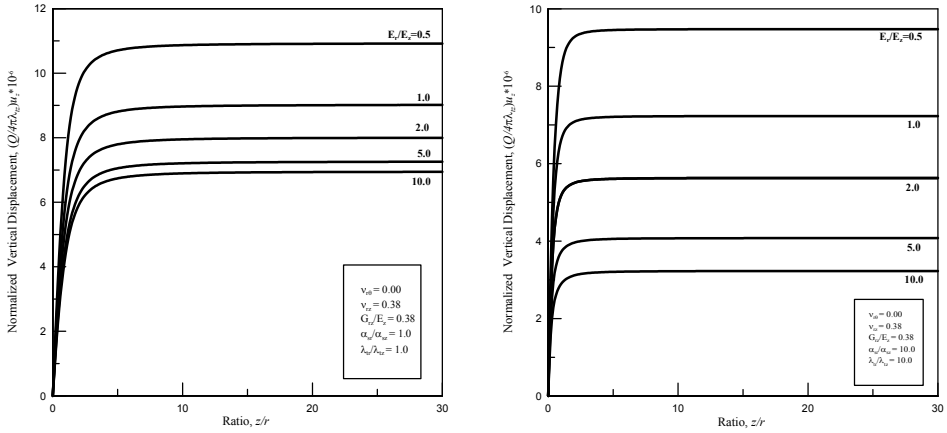
(a) (b)

Fig. 6. Influence of the degree of anisotropy α_{sr}/α_{sz} on horizontal displacement of the stratum



(a) (b)

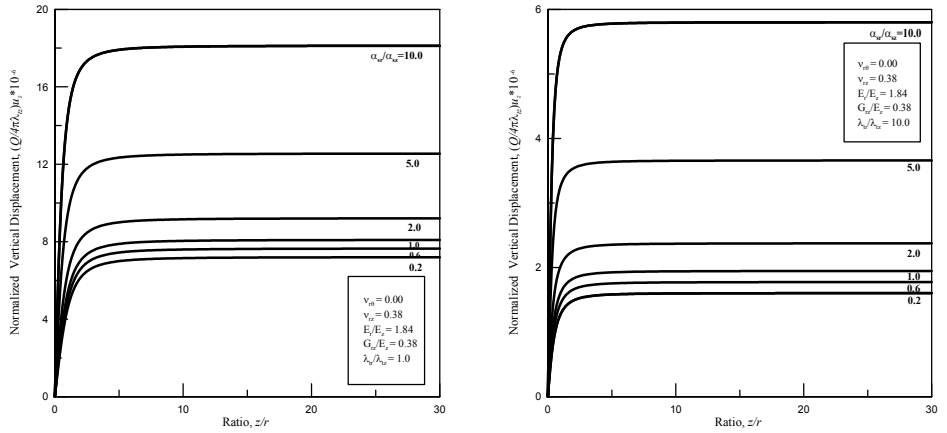
Fig. 7. Influence of the degree of anisotropy $\lambda_{tr}/\lambda_{tz}$ on horizontal displacement of the stratum



(a) (b)

Fig. 8. Influence of the degree of anisotropy E_r/E_z on vertical displacement of the stratum

Figures 8-10 show the vertical displacement as effected by the anisotropy of the soil. As shown in Figure 8, the effect of E_r/E_z on vertical displacement of the stratum is significant for the ratio E_r/E_z ranges from 0.5 to 10.0. Based on the available data, $\nu_{r\theta}=0.00$, $\nu_{rz}=0.38$, $G_{rz}/E_z=0.38$, $\lambda_{tr}/\lambda_{tz}=1.0$ or 10.0, Figure 9 uses the ratio of α_{sr}/α_{sz} to display vertical displacement of the stratum. It is shown from Figure 10 that the degree of anisotropic thermal conductivity $\lambda_{tr}/\lambda_{tz}$ has the most significant effect on vertical displacement of the stratum due to a point heat source.



(a) (b)

Fig. 9. Influence of the degree of anisotropy α_{sr}/α_{sz} on vertical displacement of the stratum

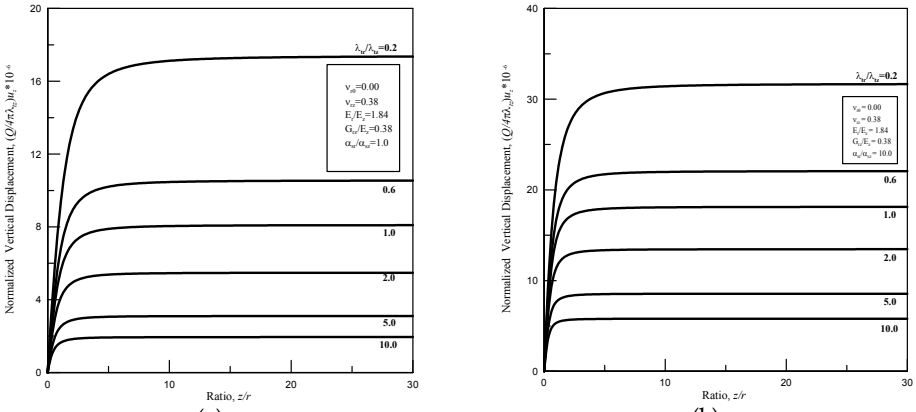


Fig. 10. Influence of the degree of anisotropy $\lambda_{tr}/\lambda_{tz}$ on vertical displacement of the stratum

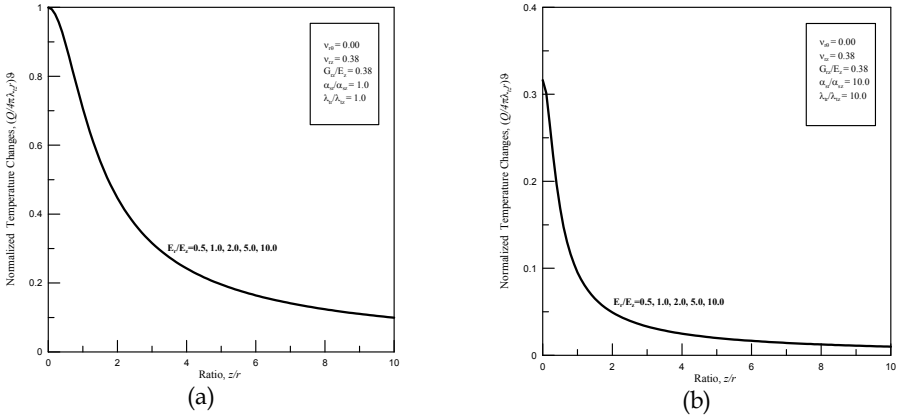


Fig. 11. Influence of the degree of anisotropy E_r/E_z on temperature increment of the stratum

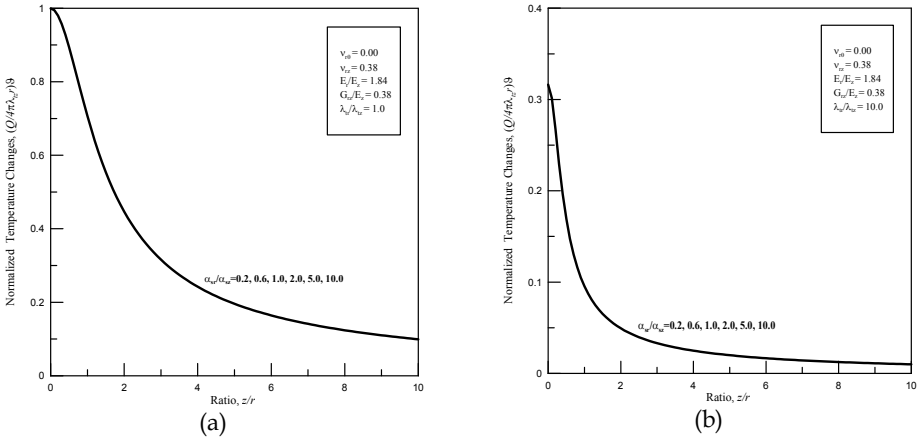


Fig. 12. Influence of the degree of anisotropy α_{sr}/α_{sz} on temperature increment of the stratum

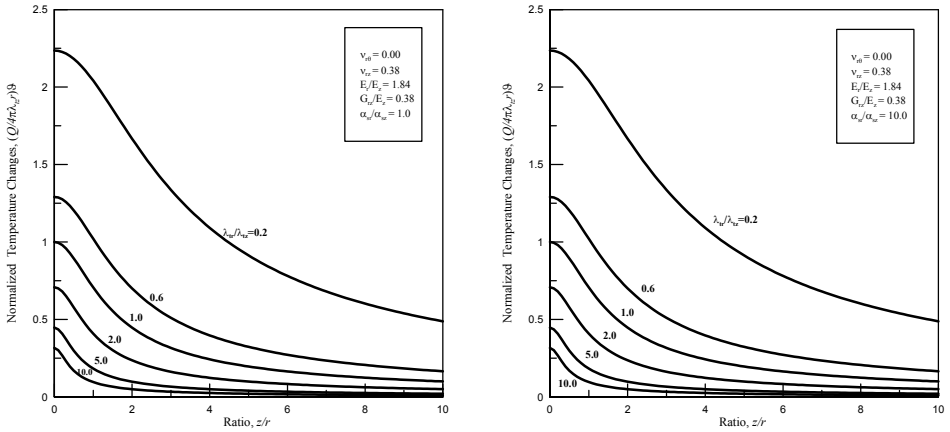


Fig. 13. Influence of the degree of anisotropy $\lambda_{tr}/\lambda_{tz}$ on temperature increment of the stratum

The temperature increments of the stratum were calculated from equation (16c) for values of various anisotropic ratio E_r/E_z , α_{sr}/α_{sz} , $\lambda_{tr}/\lambda_{tz}$ and the results are shown in Figures 11-13. Figures 11-12 display the anisotropic ratio E_r/E_z and α_{sr}/α_{sz} have no effect on the long-term temperature increment of the stratum due to a point heat source. However, Figure 13 illustrates that the ratio of anisotropic thermal conductivity $\lambda_{tr}/\lambda_{tz}$ has the most significant effect on temperature increment of the stratum. In all cases, the closer to the point heat source the larger is the temperature increment of the stratum.

5. Conclusions

This paper presents the mathematical modelling of a deep point heat source, such as the repositories of nuclear wastes. Using Hankel and Fourier integral transformations, analytic solutions of long-term thermo-mechanical responses of cross-anisotropic homogeneous soils or rocks due to a deep point heat source are obtained. The closed-form solutions of displacements and temperature change of the stratum are presented. The following conclusions were drawn based on the numerical results obtained for the available soil properties in Table 1:

1. It is noted from Eqs. (22a)-(22c) that the long-term horizontal displacement, vertical displacement and temperature increment of the stratum are not directly dependent on the shear modulus of the isotropic stratum. However, horizontal displacement and vertical displacement are dependent on the shear modulus of the cross-anisotropic full space as shown in equations (16a)-(16b) or (19a)-(19b).
2. The influence of anisotropy E_r/E_z on horizontal displacement due to a point heat source is secondary while the effects of thermoelastic anisotropy α_{sr}/α_{sz} or $\lambda_{tr}/\lambda_{tz}$ has primary effect on the horizontal displacement.
3. The influence of anisotropy E_r/E_z , α_{sr}/α_{sz} and $\lambda_{tr}/\lambda_{tz}$ on vertical displacement subjected to a point heat source is of appreciable effect on the vertical displacement.

Figures 11-12 show that the anisotropic ratio E_r/E_z and α_{sr}/α_{sz} have no effect on the long-term temperature increment of the stratum due to a point heat source. However, Figure 13 illustrates that the ratio of anisotropic thermal conductivity $\lambda_{tr}/\lambda_{tz}$ has the most significant effect on temperature increment of the stratum.

6. Acknowledgements

This work is supported by the National Science Council of Republic of China through grant NSC-98-2815-C-216-007-E, and also by the Chung Hua University under grant CHU-98-2815-C-216-007-E.

7. References

Amadei, B., Swolfs, H.S. & Savage, W.Z. (1988). Gravity-induced stresses in stratified rock masses, *Rock Mechanics and Rock Engineering*, Vol. 21, No. 1, pp. 1-20.

Barden, L. (1963). Stresses and displacements in a cross-anisotropic soil, *Geotechnique*, Vol. 13, No. 2, pp. 198-210.

Biot, M.A. (1941). General theory of three-dimensional consolidation, *Journal of Applied Physics*, Vol. 12, No. 2, pp. 155-164.

Biot, M.A. (1955). Theory of elasticity and consolidation for a porous anisotropic solid, *Journal of Applied Physics*, Vol. 26, No. 2, pp. 182-185.

Booker, J.R. & Savvidou, C. (1984). Consolidation around a spherical heat source, *International Journal of Solids and Structures*, Vol. 20, No. 11/12, pp. 1079-1090.

Booker, J.R. & Savvidou, C. (1985). Consolidation around a point heat source, *International Journal for Numerical and Analytical Methods in Geomechanics*, Vol. 9, No. 2, pp. 173-184.

Booker, J.R. & Carter, J.P. (1986). Analysis of a point sink embedded in a porous elastic half space, *International Journal for Numerical and Analytical Methods in Geomechanics*, Vol. 10, No. 2, pp. 137-150.

Chao, C.K., Chen, F.M. & Shen, M.H. (2006). Green's functions for a point heat source in circularly cylindrical layered media, *Journal of Thermal Stresses*, Vol. 29, No. 9, pp. 809 – 847.

Erdelyi, A.; Magnus, W., Oberhettinger, F. & Tricomi, F.G. (1954). *Tables of Integral Transforms*, McGraw-Hill, New York.

Gibson, R.E. (1974). The analytical method in soil mechanics, *Geotechnique*, Vol. 24, No. 2, pp. 115-140.

Gradshteyn, I.S. & Ryzhik, I.M. (1980). *Table of Integrals, Series, and Products*, Academic Press, 1160p.

Lee, S.L. & Yang, J.H. (1998). Modeling of effective thermal conductivity for a nonhomogeneous anisotropic porous medium, *International Journal of Heat and Mass Transfer*, Vol. 41, No. 6-7, pp. 931-937.

Lee, K.M. & Rowe, R.K. (1989). Deformations caused by surface loading and tunnelling: The role of elastic anisotropy, *Geotechnique*, Vol. 39, No. 1, pp. 125-140.

Lin, F.-T. & Lu, J. C.-C. (2009). Analysis of transient ground surface displacements due to an instantaneous point heat source, *Proceedings of the 20th IASTED International Conference on Modelling and Simulation*, pp. 59-64, Banff, Alberta, Canada.

Love, A.E.H. (1944). *A Treatise on the Mathematical Theory of Elasticity*, Dover Publications, New York, 643p.

Lu, J. C.-C. & Lin, F.-T. (2006). The transient ground surface displacements due to a point sink/heat source in an elastic half-space, *Geotechnical Special Publication No. 148*, *ASCE*, pp. 210-218.

Lu, J. C.-C. & Lin, F.-T. (2007). Thermo-consolidation due to a point heat source buried in a poroelastic half space, *Proceedings of the 3rd IASTED International Conference on Environmental Modelling and Simulation*, pp. 48-53, Honolulu, Hawaii, U.S.A.

Rice, J.R. & Cleary, M.P. (1976). Some basic stress diffusion solutions for fluid-saturated elastic porous media with compressible constituents, *Reviews of Geophysics and Space Physics*, Vol. 14, No. 2, pp. 227-241.

Savvidou, C. & Booker, J.R. (1989). Consolidation around a heat source buried deep in a porous thermoelastic medium with anisotropic flow properties, *International Journal for Numerical and Analytical Methods in Geomechanics*, Vol. 13, No. 1, pp. 75-90.

Schiffman, R.L. (1971). A thermoelastic theory of consolidation, *Environmental and Geophysical Heat Transfer*, C.J. Cremers, *et al.*, (eds.), ASME, Vol. 4, New York, pp. 78-84.

Sekowski, J. (1986). Stratified subsoil modelled by a cross-anisotropic elastic half-space, *International Journal for Numerical and Analytical Methods in Geomechanics*, Vol. 10, No. 4, pp. 407-414.

Sheorey, P.R. (1994). A theory for in situ stresses in isotropic and transversely isotropic rock, *International Journal of Rock Mechanics and Mining Sciences and Geomechanics Abstracts*, Vol. 31, No. 1, pp. 23-34.

Sneddon, I.N. (1951). *Fourier Transforms*, McGraw-Hill, New York, 542p.

8. Notation of Symbols

$a_i\,(i=1,2,3)$	Constants defined in equations (17a)-(17c) ($^oC^{-1}$)
A , C , F , L , N	Material constants defined by Love (*Pa*)
$b_i\,(i=1,2,3)$	Constants defined in equations (17d)-(17f) ($^oC^{-1}$)
E_r , E_z	Young's modulus in horizontal/vertical direction (*Pa*)
$f_i\,(i=r,\theta,z)$	Body forces of the stratum (N/m^3)
G	Shear modulus of the isotropic stratum (*Pa*)
G_{rz}	Modulus of shear deformation in vertical plane (*Pa*)
$\boldsymbol{h}$	Heat flux vector (J/sm^2)
$\boldsymbol{i}_r$, $\boldsymbol{i}_z$	Unit vector parallel to the radial/vertical direction (Dimensionless)
$J_v(x)$	First kind of the Bessel function of order v (Dimensionless)
q_h	Internal (or external) heat sources (J/sm^3)
Q	Strength of the point heat source (*J/s*)
(r,θ,z)	Cylindrical coordinates system (*m*, *radian*, *m*)
R	Parameter, $R=\sqrt{r^2+z^2}$ (m)
$R_i\,(i=1,2,3)$	Parameter, $R_i=\sqrt{r^2+\mu_i^2 z^2}$ (m)

R^*	Parameter, $R^* = \sqrt{r^2 + z^2} + \lvert z \rvert$ (m)
$R_i^* \left(i = 1,2,3\right)$	Parameter, $R_i^* = R_i + \mu_i \lvert z \rvert$ (m)
$u_i \left(i = r,\theta,z\right)$	Displacement components of the stratum (m)
U_r , U_z	Hankel transforms of u_r and u_z , Eqs. (10a)-(10b)
$\tilde{U}_r$, $\tilde{U}_z$	Fourier transforms of U_r and U_z , Eq. (13)
α_s	Linear thermal expansion coefficient of the isotropic stratum ($^oC^{-1}$)
α_{sr} , α_{sz}	Linear thermal expansion coefficient of the cross-anisotropic stratum in horizontal/vertical direction ($^oC^{-1}$)
β_r , β_z	Thermal expansion factors of the cross-anisotropic stratum $(Pa/^oC)$
β_r^* , β_z^*	Thermal expansion factors of the isotropic stratum $(Pa/^oC)$
$\delta\left(x\right)$	Dirac delta function (m^{-1})
η	Parameter, $\eta = \left(1-\nu\right)/\left(1-2\nu\right)$ (Dimensionless)
ϑ	Temperature change of the stratum (^oC)
Θ	Hankel transform of ϑ , Eq. (10c)
$\tilde{\Theta}$	Fourier transform of Θ , Eq. (13)
λ	Lame constant of the isotropic stratum (Pa)
λ_t	Thermal conductivity of the isotropic thermoelastic medium $(J/sm°C)$
λ_{tr} , λ_{tz}	Thermal conductivity of the cross-anisotropic thermoelastic medium in the horizontal/vertical direction $(J/sm°C)$
μ_1 , μ_2	Characteristic roots of characteristic equation (18) (Dimensionless)
μ_3	Characteristic root, $\mu_3 = \sqrt{\lambda_{tr}/\lambda_{tz}}$ (Dimensionless)
ν	Poisson's ratio of the isotropic stratum (Dimensionless)
ν_{rz}	Poisson's ratio for strain in the vertical direction due to a horizontal direct stress (Dimensionless)
$\nu_{r\theta}$	Poisson's ratio for strain in the horizontal direction due to a horizontal direct stress (Dimensionless)
ν_{zr}	Poisson's ratio for strain in the horizontal direction due to a vertical direct stress (Dimensionless)
ξ	Hankel transform parameter (m^{-1})
$\sigma_{ij} \left(i,j = r,\theta,z\right)$	Thermal stress components of the stratum (Pa)
$\varphi_i \left(i = 1,2,3,4\right)$	Functions defined in Eqs. (21a)-(21d) (Dimensionless)
ω	Fourier transform parameter (m^{-1})

14

Modelling of Transient Ground Surface Displacements Due to a Point Heat Source

Feng-Tsai Lin[1] and John C.-C. Lu[2]
[1]*Department of Naval Architecture, National Kaohsiung Marine University*
[2]*Department of Civil Engineering, Chung Hua University*
Taiwan

1. Introduction

Using Laplace-Hankel integral transformations, transient closed-form solutions of the thermally induced ground surface displacements, excess pore water pressure and temperature increment due to an instantaneous point heat source buried in an isothermal permeable half space are presented and discussed. The basic formulations of the governing equations are on the basis of Biot's three-dimensional consolidation theory of porous media. Numerical results show that the maximum ground surface horizontal displacement is around 38.5% of the maximum ground surface vertical displacement. The study concludes that the thermally induced horizontal displacement is significant. The solutions can be used to test numerical models and numerical simulations of the thermoelastic processes near the heat sources.

Heat source buried in the stratum leads to thermo-mechanical responses of fluid saturated porous medium. The heat source such as a canister of radioactive waste can cause temperature rise in the soil. The solid skeleton and pore fluid expand due to the heat source, and the volume increase of pore fluid is greater than that of the voids of solid matrix. This leads to an increase in pore fluid pressure and a reduction in effective stress. Therefore, thermal failure of soil will occur as a result of losing shear resistance due to the decrease in effective stress.

Attention is focused on the analytical solutions of the transient thermoelastic responses of an isotropic stratum due to an instantaneous point heat source. The responses of the stratum were satisfactorily modeled by assuming it as a thermoelastic porous continuum (Booker & Savvidou, 1985). It suggested that linear theory was adequate for a repository design based on technical conservatism. For example, Hueckel and Peano (1987) indicated that European guidelines require that temperature increments in the soil close to the heat source should not exceed 80°C while the temperature increments at the ground surface are limited to less than 1°C. Given these modest temperature increments, Hollister *et al.* (1981) observed that any significant non-linear behavior and/or plastic deformation of the soil would be confined to a relatively small volume of soil around the waste canister itself. In this case, a linear model can provide a reasonable approximation to the assessment of a proposed design (Smith & Booker, 1996). Hudson *et al.* (2005) given advices on how to incorporate thermo-hydro-

mechanical coupled processes into performance and safety assessments and design studies for radioactive waste disposal in geological formations.
Governing equations of a fluid-saturated poroelastic solid in an isothermal quasi-static state were developed by Biot (1941, 1955). Lu and Lin (2006) displayed transient ground surface displacement produced by a point heat source/sink through analog quantities between poroelasticity and thermoelasticity. Booker and Savvidou (1984, 1985), Savvidou and Booker (1989) derived an extended Biot theory including the thermal effects and presented solutions of thermo-consolidation around the spherical and point heat sources. In their solutions, the isotropic or transversely isotropic flow properties are considered, whereas the isotropic elastic and thermal properties of the soils are introduced.
Based on Biot's three-dimensional consolidation theory of porous media, analytical solutions of the transient thermo-consolidation deformation due to a point heat source of constant strength buried in a saturated isotropic poroelastic half space were presented by Lu and Lin (2007). In this paper, instantaneous point heat source induced transient ground surface displacements are derived by using Laplace-Hankel integral transforms. The soil mass is modeled as a homogeneous isotropic saturated elastic half space of porous medium. Case of isothermal permeable half space boundary is investigated. Results are illustrated and compared to provide better understanding of the time dependent thermoelastic responses due to an instantaneous point heat source. The solutions can be used to test numerical models and the detailed numerical simulations of the thermoelastic processes near the buried heat sources.

2. Mathematical Model

2.1 Basic Equations

Figure 1 shows an instantaneous point heat source buried in a saturated porous stratum at depth h. The porous soil mass is considered as a homogeneous isotropic thermoelastic half space. The constitutive behaviours of the elastic soil skeleton are presented as:

$$\sigma_{ij} = 2G\varepsilon_{ij} + \frac{2G\nu}{1-2\nu}\varepsilon\delta_{ij} - \frac{2G(1+\nu)\alpha_s}{1-2\nu}\vartheta\delta_{ij} - p\delta_{ij} . \tag{1}$$

Here, σ_{ij}, ε_{ij} and ϑ are the total stress components, strain components and temperature increment measured from the reference state of the porous medium, respectively; ε is the volume strain of the porous medium; δ_{ij} is the Kronecker delta. The excess pore water pressure p is positive for compression. The constants ν, G and α_s are Poisson's ratio, shear modulus, and linear thermal expansion coefficient of the skeletal materials, respectively.
The strains ε_{ij} and displacement components u_i are given by the linear law:

$$\varepsilon_{ij} = \tfrac{1}{2}\left(u_{i,j} + u_{j,i}\right) . \tag{2}$$

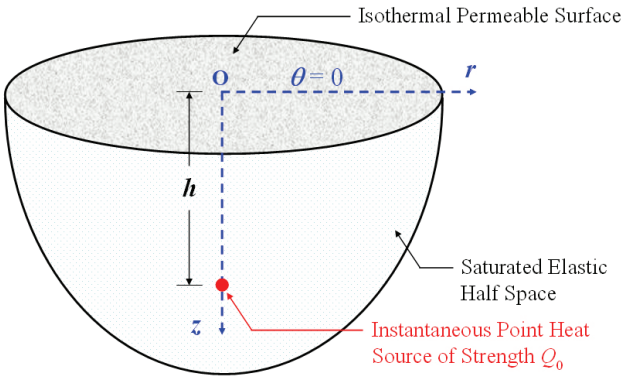


Fig. 1. Instantaneous point heat source buried in a homogeneous poroelastic half space

The total stress components satisfy the following equilibrium equations:

$$\sigma_{ij,j} + b_i = 0\,, \tag{3}$$

where b_i denote the body forces. From Eqs. (1) and (2), the equilibrium equations (3) for axially symmetric problem without body forces b_i can be expressed in terms of displacements u_i, excess pore water pressure p and temperature change ϑ of the thermoelastic half space in cylindrical coordinates (r,θ,z) as below:

$$G\nabla^2 u_r + \frac{G}{1-2\nu}\frac{\partial \varepsilon}{\partial r} - G\frac{u_r}{r^2} - \frac{\partial p}{\partial r} - \frac{2G(1+\nu)\alpha_s}{1-2\nu}\frac{\partial \vartheta}{\partial r} = 0\,, \tag{4a}$$

$$G\nabla^2 u_z + \frac{G}{1-2\nu}\frac{\partial \varepsilon}{\partial z} - \frac{\partial p}{\partial z} - \frac{2G(1+\nu)\alpha_s}{1-2\nu}\frac{\partial \vartheta}{\partial z} = 0\,, \tag{4b}$$

where the volume strain of the porous medium ε can be denoted as $\varepsilon = \partial u_r/\partial r + u_r/r + \partial u_z/\partial z$, while the Laplacian operator $\nabla^2 = \partial^2/\partial r^2 + 1/r\,\partial/\partial r + \partial^2/\partial z^2$.
According to Darcy's law, the governing equation of the conservation of mass can be expressed as

$$-\frac{k}{\gamma_w}\nabla^2 p + \frac{\partial \varepsilon}{\partial t} + n\beta\frac{\partial p}{\partial t} + 3\alpha_u\frac{\partial \vartheta}{\partial t} = 0\,, \tag{5}$$

where k and n are the permeability and porosity of the porous medium, respectively; β and γ_w are the compressibility and unit weight of pore water, respectively; $\alpha_u = (1-n)\alpha_s + n\alpha_w$, in which α_w is the coefficient of linear thermal expansion of the pore water.

For an instantaneous point heat source of strength Q_0 buried at point $(0,h)$, the uncoupled governing equation in axially symmetry is obtained from the conservation of energy and heat conduction law as following:

$$-\lambda_t \nabla^2 \vartheta + m\frac{\partial \vartheta}{\partial t} - \frac{Q_0}{2\pi r}\delta(r)\delta(z-h)\,\delta(t) = 0\,, \tag{6}$$

in which λ_t is the heat conduction coefficient of the porous stratum; the symbol $m = (1-n)\rho_s c_s + n\rho_w c_w$, c_s and c_w are the specific heats of the skeletal materials and pore water, while ρ_s, ρ_w are their densities, respectively; $\delta(r)$ or $\delta(t)$ is Dirac delta function. Eqs. (4a), (4b), (5) and (6) constitute the basic governing equations of the time-dependent axially symmetric thermoelastic responses of a saturated porous medium.

2.2 Boundary Conditions and Initial Conditions

The half space surface, $z = 0$, is treated as a traction-free, isothermal and permeable boundary for all time $t \geq 0$. Its mathematical statements of the boundary conditions are:

$$\sigma_{rz}(r,0,t) = 0,\ \sigma_{zz}(r,0,t) = 0,\ p(r,0,t) = 0,\ \text{and}\ \ \vartheta(r,0,t) = 0\,. \tag{7a}$$

It's reasonable to assume that the instantaneous point heat source has no effect on the far boundary at $z \to \infty$ for all time. Hence

$$\lim_{z\to\infty}\{u_r(r,z,t), u_z(r,z,t), p(r,z,t), \vartheta(r,z,t)\} = \{0,0,0,0\}\,. \tag{7b}$$

Assuming no initial change in displacements, temperature increment and seepage for the poroelastic medium, then the initial conditions at time $t = 0$ of the mathematical model due to an instantaneous point heat source can be treated as:

$$u_r(r,z,0) = 0,\ u_z(r,z,0) = 0,\ p(r,z,0) = 0,\ \text{and}\ \vartheta(r,z,0) = 0\,. \tag{8}$$

The transient ground surface displacements can be derived from the differential equations (4a), (4b), (5) and (6) corresponding with the boundary conditions at $z = 0$, $z \to \infty$, and initial conditions at time $t = 0$.

3. Analytic Solutions

3.1 Laplace-Hankel Transformations

Applying initial conditions of Eq. (8), the governing partial differential equations (4a), (4b), (5) and (6) are reduced to ordinary differential equations by performing appropriate Laplace-Hankel transforms (Sneddon, 1951) with respect to the time variable t and the radial coordinate r:

$$\left(\frac{d^2}{dz^2}-2\eta\xi^2\right)\tilde{u}_r-(2\eta-1)\xi\frac{d\tilde{u}_z}{dz}+\frac{1}{G}\xi\tilde{p}+\frac{2(1+\nu)\alpha_s}{1-2\nu}\xi\tilde{\vartheta}=0\,, \tag{9a}$$

$$(2\eta-1)\xi\frac{d\tilde{u}_r}{dz}+\left(2\eta\frac{d^2}{dz^2}-\xi^2\right)\tilde{u}_z-\frac{1}{G}\frac{d\tilde{p}}{dz}-\frac{2(1+\nu)\alpha_s}{1-2\nu}\frac{d\tilde{\vartheta}}{dz}=0\,, \tag{9b}$$

$$-\frac{k}{\gamma_w}\left(\frac{d^2}{dz^2}-\xi^2\right)\tilde{p}+s\left(\xi\tilde{u}_r+\frac{d\tilde{u}_z}{dz}\right)+n\beta s\tilde{p}+3\alpha_u s\tilde{\vartheta}=0\,, \tag{9c}$$

$$-\lambda_t\left(\frac{d^2}{dz^2}-\xi^2\right)\tilde{\vartheta}+ms\tilde{\vartheta}-\frac{Q_0}{2\pi}\delta(z-h)=0\,, \tag{9d}$$

where ξ and s are Hankel and Laplace transform parameters, respectively; $\eta=(1-\nu)/(1-2\nu)$; and the symbols $\tilde{u}_r$, $\tilde{u}_z$, $\tilde{p}$, $\tilde{\vartheta}$ are defined as:

$$\tilde{u}_r(z;\xi,s)=\int_0^\infty\int_0^\infty ru_r(r,z,t)exp(-st)J_1(\xi r)dtdr\,, \tag{10a}$$

$$\tilde{u}_z(z;\xi,s)=\int_0^\infty\int_0^\infty ru_z(r,z,t)exp(-st)J_0(\xi r)dtdr\,, \tag{10b}$$

$$\tilde{p}(z;\xi,s)=\int_0^\infty\int_0^\infty rp(r,z,t)exp(-st)J_0(\xi r)dtdr\,, \tag{10c}$$

$$\tilde{\vartheta}(z;\xi,s)=\int_0^\infty\int_0^\infty r\vartheta(r,z,t)exp(-st)J_0(\xi r)dtdr\,, \tag{10d}$$

in which $J_\alpha(x)$ represents the first kind of Bessel function of order α.
The general solutions of equations (9a)-(9d) are obtained as below:

$$\begin{aligned}\tilde{u}_r(z;\xi,s)=&(A_1+A_2z)e^{\xi z}+(A_3+A_4z)e^{-\xi z}+A_5e^{\sqrt{\xi^2+\frac{s}{c_1}}z}+A_6e^{-\sqrt{\xi^2+\frac{s}{c_1}}z}+A_7e^{\sqrt{\xi^2+\frac{s}{c_2}}z}+A_8e^{-\sqrt{\xi^2+\frac{s}{c_2}}z}\\&+\frac{c_2Q_0}{8\pi\eta G\lambda_t}\left(-\frac{c_a}{s}e^{-\xi|z-h|}-\frac{c_b}{s}\xi\sqrt{\xi^2+\frac{s}{c_1}}^{-1}e^{-\sqrt{\xi^2+\frac{s}{c_1}}|z-h|}+\frac{c_c}{s}\xi\sqrt{\xi^2+\frac{s}{c_2}}^{-1}e^{-\sqrt{\xi^2+\frac{s}{c_2}}|z-h|}\right),\end{aligned} \tag{11a}$$

$$\begin{aligned}\tilde{u}_z(z;\xi,s)=&\left[-A_1+\frac{1+(2\eta+1)n\beta G}{1+(2\eta-1)n\beta G}\frac{1}{\xi}A_2-A_2z\right]e^{\xi z}+\left[A_3+\frac{1+(2\eta+1)n\beta G}{1+(2\eta-1)n\beta G}\frac{1}{\xi}A_4+A_4z\right]e^{-\xi z}\\&-\frac{1}{\xi}\sqrt{\xi^2+\frac{s}{c_1}}A_5e^{\sqrt{\xi^2+\frac{s}{c_1}}z}+\frac{1}{\xi}\sqrt{\xi^2+\frac{s}{c_1}}A_6e^{-\sqrt{\xi^2+\frac{s}{c_1}}z}-\frac{1}{\xi}\sqrt{\xi^2+\frac{s}{c_2}}A_7e^{\sqrt{\xi^2+\frac{s}{c_2}}z}+\frac{1}{\xi}\sqrt{\xi^2+\frac{s}{c_2}}A_8e^{-\sqrt{\xi^2+\frac{s}{c_2}}z}\\&\pm\frac{c_2Q_0}{8\pi\eta G\lambda_t}\left(-\frac{c_a}{s}e^{-\xi|z-h|}-\frac{c_b}{s}e^{-\sqrt{\xi^2+\frac{s}{c_1}}|z-h|}+\frac{c_c}{s}e^{-\sqrt{\xi^2+\frac{s}{c_2}}|z-h|}\right),\end{aligned} \tag{11b}$$

$$\tilde{p}(z;\xi,s)=\frac{2G}{1+(2\eta-1)n\beta G}\left(-A_2e^{\xi z}+A_4e^{-\xi z}\right)$$

$$-2\eta G\frac{1}{\xi}\frac{s}{c_1}\left(A_5e^{\sqrt{\xi^2+\frac{s}{c_1}}z}+A_6e^{-\sqrt{\xi^2+\frac{s}{c_1}}z}\right)-2\eta G\frac{1}{\xi}\frac{c_b}{c_c}\frac{s}{c_1}\left(A_7e^{\sqrt{\xi^2+\frac{s}{c_2}}z}+A_8e^{-\sqrt{\xi^2+\frac{s}{c_2}}z}\right)$$

$$+\frac{Q_0}{4\pi\lambda_t}\frac{c_bc_2}{c_1}\left(\sqrt{\xi^2+\frac{s}{c_1}}^{-1}e^{-\sqrt{\xi^2+\frac{s}{c_1}}|z-h|}-\sqrt{\xi^2+\frac{s}{c_2}}^{-1}e^{-\sqrt{\xi^2+\frac{s}{c_2}}|z-h|}\right), \tag{11c}$$

$$\tilde{\vartheta}(z;\xi,s)=2\eta G\frac{1}{\xi}\frac{1}{c_c}\frac{s}{c_2}\left(A_7e^{\sqrt{\xi^2+\frac{s}{c_2}}z}+A_8e^{-\sqrt{\xi^2+\frac{s}{c_2}}z}\right)+\frac{Q_0}{4\pi\lambda_t}\sqrt{\xi^2+\frac{s}{c_2}}^{-1}e^{-\sqrt{\xi^2+\frac{s}{c_2}}|z-h|}, \tag{11d}$$

in which

$$c_a=\frac{c_1}{c_3}-\frac{2G(1+\nu)\alpha_s}{1-2\nu}, \tag{12a}$$

$$c_b=\frac{c_1^2}{c_3(c_2-c_1)}, \tag{12b}$$

$$c_c=\frac{c_1c_2}{c_3(c_2-c_1)}-\frac{2G(1+\nu)\alpha_s}{1-2\nu}, \tag{12c}$$

where $c_a+c_b=c_c$ and

$$c_1=\frac{k}{\gamma_w}\frac{2\eta G}{2\eta Gn\beta+1}, \tag{13a}$$

$$c_2=\frac{\lambda_t}{m}, \tag{13b}$$

$$c_3=\frac{k}{\gamma_w}\frac{1-\nu}{3(1-\nu)\alpha_u+(1+\nu)\alpha_s}. \tag{13c}$$

The constants $A_i\,(i=1,2,\cdots,8)$ in Eqs. (11a)-(11d) are functions of the transformed variables ξ and s which must be determined from the transformed mechanical, flow and thermal boundary conditions. The upper and lower signs in equation (11b) are for the conditions of $(z-h)\geq 0$ and $(z-h)<0$, respectively.

3.2 Transformed Boundary Conditions

Taking Laplace-Hankel transforms for the boundary conditions at $z=0$, the Eq. (7a), yields the transformed boundary conditions as following:

$$\frac{d\tilde{u}_r(0;\xi,s)}{dz}-\xi\tilde{u}_z(0;\xi,s)=0,\ \eta\frac{d\tilde{u}_z(0;\xi,s)}{dz}+(\eta-1)\xi\tilde{u}_r(0;\xi,s)=0,$$

$$\tilde{p}(0;\xi,s)=0,\text{ and }\tilde{\vartheta}(0;\xi,s)=0. \tag{14a}$$

In this manipulation, the boundary conditions at $z\to\infty$ are used to perform the integral transformations as below:

$$\lim_{z\to\infty}\left\{\tilde{u}_r\left(z;\xi,s\right),\tilde{u}_z\left(z;\xi,s\right),\tilde{p}\left(z;\xi,s\right),\tilde{\vartheta}\left(z;\xi,s\right)\right\}=\left\{0,0,0,0\right\}. \tag{14b}$$

Here, $\tilde{u}_r$, $\tilde{u}_z$, $\tilde{p}$ and $\tilde{\vartheta}$ follow the definitions of Eqs. (10a)-(10d).

The constants $A_i\left(i=1,2,\cdots,8\right)$ of the general solutions can be determined by the transformed half space boundary conditions at $z=0$ and the remote boundary conditions at $z\to\infty$. Finally, the desired quantities u_r, u_z, p and ϑ are obtained by applying appropriate inverse Laplace-Hankel transformations with the help of mathematical handbook (Erdelyi *et al.*, 1954).

3.3 Expressions for Ground Surface Displacements

The study is focused on horizontal and vertical displacements of the ground surface, $z=0$, due to an instantaneous point heat source. The transformed transient ground surface displacements $\tilde{u}_r\left(0;\xi,s\right)$ and $\tilde{u}_z\left(0;\xi,s\right)$ due to an instantaneous point heat source are derived from the transformed general solutions (11a)-(11b) and mechanical boundary conditions at $z=0$ and $z\to\infty$, the Eqs. (14a)-(14b), as below:

$$\tilde{u}_r\left(0;\xi,s\right)=\frac{Q_0}{2\left(2\eta-1\right)\pi Gm}\left[-\frac{c_a}{s}exp\left(-\xi h\right)-\frac{c_b}{s}exp\left(-\sqrt{\xi^2+\frac{s}{c_1}}h\right)+\frac{c_c}{s}exp\left(-\sqrt{\xi^2+\frac{s}{c_2}}h\right)\right], \tag{15a}$$

$$\tilde{u}_z\left(0;\xi,s\right)=\frac{Q_0}{2\left(2\eta-1\right)\pi Gm}\left[\frac{c_a}{s}exp\left(-\xi h\right)+\frac{c_b}{s}exp\left(-\sqrt{\xi^2+\frac{s}{c_1}}h\right)-\frac{c_c}{s}exp\left(-\sqrt{\xi^2+\frac{s}{c_2}}h\right)\right]. \tag{15b}$$

The Laplace-Hankel inversion formulae for displacements are defined as following:

$$u_r\left(r,z,t\right)=\frac{1}{2\pi i}\int_{\alpha-i\infty}^{\alpha+i\infty}\int_0^\infty \xi\tilde{u}_r\left(z;\xi,s\right)J_1\left(\xi r\right)exp\left(st\right)d\xi ds, \tag{16a}$$

$$u_z\left(r,z,t\right)=\frac{1}{2\pi i}\int_{\alpha-i\infty}^{\alpha+i\infty}\int_0^\infty \xi\tilde{u}_z\left(z;\xi,s\right)J_0\left(\xi r\right)exp\left(st\right)d\xi ds. \tag{16b}$$

Using integral transform handbook (Erdelyi *et al.*, 1954) and integral inversions listed in Eqs. (16a)-(16b), the transient horizontal displacement $u_r\left(r,0,t\right)$ and vertical displacement $u_z\left(r,0,t\right)$ of the ground surface due to an instantaneous point heat source of strength Q_0 are obtained as follows:

$$u_r\left(r,0,t\right)=\frac{Q_0}{2\left(2\eta-1\right)\pi Gm}\left\{-\frac{c_a r}{\left(h^2+r^2\right)^{3/2}}-\frac{c_b}{c_1}\int_0^{c_1 t}\frac{c_1 hr}{16\tau^3}e^{-\frac{2h^2+r^2}{8\tau}}\left[I_0\left(\frac{r^2}{8\tau}\right)-I_1\left(\frac{r^2}{8\tau}\right)\right]d\tau\right.$$
$$\left.+\frac{c_c}{c_2}\int_0^{c_2 t}\frac{c_2 hr}{16\tau^3}e^{-\frac{2h^2+r^2}{8\tau}}\left[I_0\left(\frac{r^2}{8\tau}\right)-I_1\left(\frac{r^2}{8\tau}\right)\right]d\tau\right\}, \tag{17a}$$

$$u_z(r,0,t)=\frac{Q_0}{2(2\eta-1)\pi Gm}\left\{\frac{c_a h}{\left(h^2+r^2\right)^{2/3}}+\frac{c_b}{c_1}\left[\frac{c_1 h}{h^2+r^2}\frac{1}{\sqrt{\pi c_1 t}}e^{-\frac{h^2+r^2}{4c_1 t}}+\frac{c_1 h}{\left(h^2+r^2\right)^{3/2}}erfc\left(\frac{\sqrt{h^2+r^2}}{2\sqrt{c_1 t}}\right)\right]\right.$$
$$\left.-\frac{c_c}{c_2}\left[\frac{c_2 h}{h^2+r^2}\frac{1}{\sqrt{\pi c_2 t}}e^{-\frac{h^2+r^2}{4c_2 t}}+\frac{c_2 h}{\left(h^2+r^2\right)^{3/2}}erfc\left(\frac{\sqrt{h^2+r^2}}{2\sqrt{c_2 t}}\right)\right]\right\}, \tag{17b}$$

where $erfc(x)$ denotes the complementary error function; $I_\alpha(x)$ is known as the modified Bessel function of the first kind of order α. The transient ground surface horizontal and vertical displacements shown in Eqs. (17a)-(17b) vanished when $t\to\infty$ in this linear elastic model.

The maximum ground surface horizontal displacement $u_{r\,max}$ of the half space due to an instantaneous point heat source is derived from Eq. (17a) by letting $r=h/\sqrt{2}\approx 0.707h$. After doing so, we have

$$u_{r\,max}=u_r\left(h/\sqrt{2},0,0^+\right)=-\frac{\sqrt{3}c_a Q_0}{9(2\eta-1)\pi Gmh^2}, \tag{18}$$

in which the value $r=h/\sqrt{2}$ is derived when $du_r\left(r,0,0^+\right)/dr$ is equal to zero.

The maximum ground surface vertical displacement $u_{z\,max}$ of the isothermal permeable half space due to an instantaneous point heat source is derived from Eq. (17b) by letting $r=0$. Hence

$$u_{z\,max}=u_z\left(0,0,0^+\right)=\frac{c_a Q_0}{2(2\eta-1)\pi Gmh^2}. \tag{19}$$

The absolute value of the ratio of $u_{r\,max}/u_{z\,max}$ can be derived from Eqs. (18) and (19) as below:

$$\left|\frac{u_{r\,max}}{u_{z\,max}}\right|\times 100\%=\frac{2\sqrt{3}}{9}\times 100\%\cong 38.5\%. \tag{20}$$

The above result shows the maximum ground surface horizontal displacement is around 38.5% of the maximum vertical displacement for the isothermal permeable ground surface due to an instantaneous point heat source.

3.4 Expressions for Excess Pore Water Pressure and Temperature Increment of the Stratum

The study also addressed the excess pore water pressure and temperature increment of the poroelastic half space due to an instantaneous point heat source. The transformed excess pore water pressure and temperature increment are obtained from Eqs. (11c)-(11d) with the

help of transformed hydraulic the thermal boundary conditions in equations (14a)-(14b) and can be expressed as following:

$$\tilde{p}(z;\xi,s)=\frac{Q_0}{4\pi\lambda_t}\frac{c_b c_2}{c_1}\left\{\sqrt{\xi^2+\frac{s}{c_1}}^{-1}\left[exp\left(-\sqrt{\xi^2+\frac{s}{c_1}}\left|z-h\right|\right)-exp\left(-\sqrt{\xi^2+\frac{s}{c_1}}\left(z+h\right)\right)\right]\right.$$
$$\left.-\sqrt{\xi^2+\frac{s}{c_2}}^{-1}\left[exp\left(-\sqrt{\xi^2+\frac{s}{c_2}}\left|z-h\right|\right)-exp\left(-\sqrt{\xi^2+\frac{s}{c_2}}\left(z+h\right)\right)\right]\right\}, \tag{21a}$$

$$\tilde{\vartheta}(z;\xi,s)=\frac{Q_0}{4\pi\lambda_t}\sqrt{\xi^2+\frac{s}{c_2}}^{-1}\left[exp\left(-\sqrt{\xi^2+\frac{s}{c_2}}\left|z-h\right|\right)-exp\left(-\sqrt{\xi^2+\frac{s}{c_2}}\left(z+h\right)\right)\right]. \tag{21b}$$

The Laplace-Hankel inversion formulae for $\tilde{p}(z;\xi,s)$ and $\tilde{\vartheta}(z;\xi,s)$ are defined as below:

$$p(r,z,t)=\frac{1}{2\pi i}\int_{\alpha-i\infty}^{\alpha+i\infty}\int_0^\infty \xi\tilde{p}(z;\xi,s)J_0(\xi r)exp(st)d\xi ds\,, \tag{22a}$$

$$\vartheta(r,z,t)=\frac{1}{2\pi i}\int_{\alpha-i\infty}^{\alpha+i\infty}\int_0^\infty \xi\tilde{\vartheta}(z;\xi,s)J_0(\xi r)exp(st)d\xi ds\,. \tag{22b}$$

The transient excess pore water pressure $p(r,z,t)$ and temperature increment $\vartheta(r,z,t)$ of the saturated isothermal permeable half space due to an instantaneous point heat source are obtained as following:

$$p(r,z,t)=\frac{Q_0}{8\pi\lambda_t}\frac{c_1c_2}{c_3(c_2-c_1)}\left\{\frac{1}{\sqrt{\pi c_1t^3}}\left[exp\left(-\frac{r^2+(z-h)^2}{4c_1t}\right)-exp\left(-\frac{r^2+(z+h)^2}{4c_1t}\right)\right]\right.$$
$$\left.-\frac{1}{\sqrt{\pi c_2t^3}}\left[exp\left(-\frac{r^2+(z-h)^2}{4c_2t}\right)-exp\left(-\frac{r^2+(z+h)^2}{4c_2t}\right)\right]\right\}, \tag{23a}$$

$$\vartheta(r,z,t)=\frac{Q_0}{8\pi\lambda_t}\frac{1}{\sqrt{\pi c_2t^3}}\left[exp\left(-\frac{r^2+(z-h)^2}{4c_2t}\right)-exp\left(-\frac{r^2+(z+h)^2}{4c_2t}\right)\right]. \tag{23b}$$

4. Numerical Results

Following Ma and Hueckel (1992, 1993), Bai and Abousleiman (1997), the selected representative parameters are listed in Table 1 to verify the proposed solutions. The constants $c_i\,(i=1,2,3,a,b,c)$ are derived as shown in Table 2 by using the parameters listed in Table 1, Eqs. (12a)-(12c) and Eqs. (13a)-(13c).

Symbol	Value	Unit
Shear modulus of the skeletal material, G	100	*MPa*
Poisson's ratio of the skeletal material, ν	0.23	Dimensionless
Porosity of the porous medium, n	0.2	Dimensionless
Coefficient of heat conduction of the porous medium, λ_t	1.69	*J/sm°C*
Linear thermal expansion coefficient of the pore water, α_w	3.33×10^{-6}	$^oC^{-1}$
Linear thermal expansion coefficient of the skeletal material, α_s	3.33×10^{-7}	$^oC^{-1}$
Specific heat of the pore water, c_w	500	J/kg^oC
Specific heat of skeletal material, c_s	200	J/kg^oC
Density of the pore water, ρ_w	1,000	kg/m^3
Density of the skeletal material, ρ_s	2,000	kg/m^3
Permeability of porous medium, k	1×10^{-11}	*m/s*
Unit weight of the pore water, γ_w	9,810	N/m^3
Compressibility of the pore water, β	5×10^{-10}	Pa^{-1}

Table 1. Selected representative parameters (Ma and Hueckel, 1992, 1993; Bai and Abousleiman, 1997)

Symbol	Value	Unit
c_1	2.826×10^{-7}	m^2/s
c_2	4.024×10^{-6}	m^2/s
c_3	3.062×10^{-10}	$^oCm^4/Ns$
c_a	7.714×10^{2}	$N/^oCm^2$
c_b	6.974×10^{1}	$N/^oCm^2$
c_c	8.411×10^{2}	$N/^oCm^2$

Table 2. Values of $c_i\,(i = 1, 2, 3, a, b, c)$

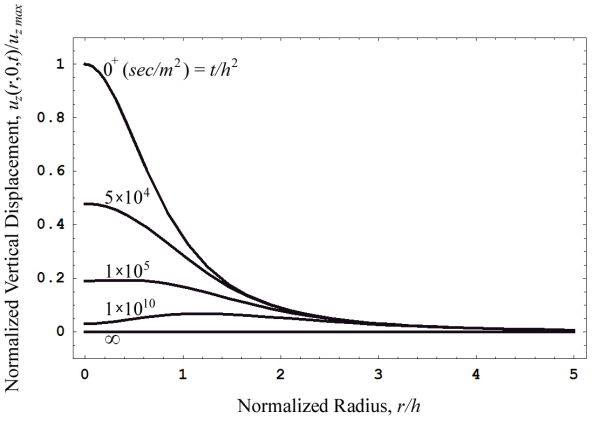


Fig. 2. Vertical displacement profile at the ground surface $z = 0$

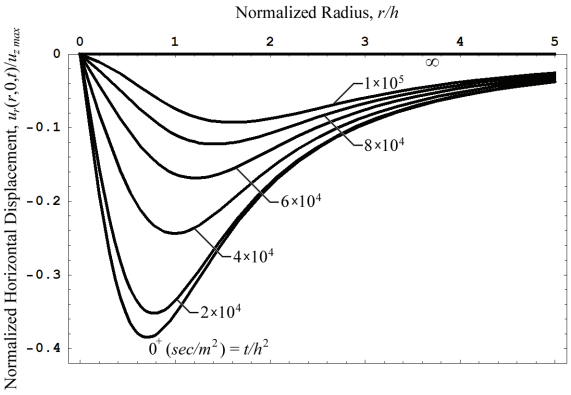


Fig. 3. Horizontal displacement profile at the ground surface $z = 0$

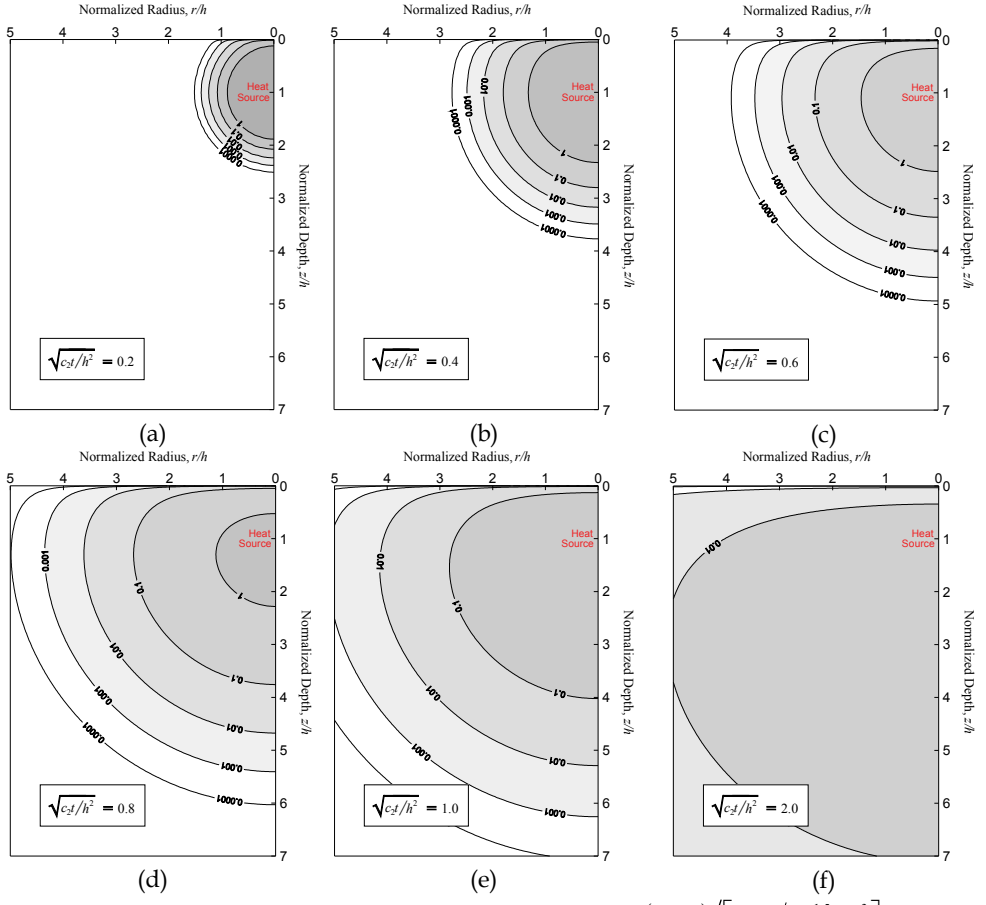


Fig. 4. Distribution of normalized temperature increments $\vartheta(r,z,t)/\left[c_2 Q_0/8\pi^{1.5}\lambda_i h^3\right]$

The profiles of vertical and horizontal displacements at the ground surface $z = 0$ are normalized by $u_{z\,max}$ as shown in Figures 2 and 3, respectively. The results shown in Figures 2 and 3 indicate that the ground surface displacements due to instantaneous point heat source can reach its extreme values initially, and then the displacements decreases gradually. Figure 3 shows that the maximum ground surface horizontal displacement is around 38.5% of the maximum ground surface vertical displacement. Figures 2 and 3 also concluded that the long-term thermoelastic ground surface deformations due to an instantaneous point heat source vanished in this linear elastic model.

From Eq. (23b), the profiles of normalized temperature increment $\vartheta(r,z,t)/\left[c_2 Q_0/8\pi^{1.5}\lambda_t h^3\right]$ of isothermal permeable half space at six different dimensionless time factor $\sqrt{c_2 t/h^2} = 0.2$, 0.4, 0.6, 0.8, 1.0 and 2.0 are illustrated in Figures 4(a)-(f), respectively. The changes in temperature increment have positive value of ϑ which is caused by the heating of instantaneous point heat source. It's observed that the positive temperature change increases to a wider region of the half space initially and then gradually decreased. The stratum temperature rise caused by instantaneous point heat finally disappeared, and the elastic deformations due to instantaneous point heat source fully recovered as the temperature increment vanished.

The presented closed-form solutions can be used to test numerical models for thermoelastic processes. It can also be used in more detailed numerical simulations of the processes near the buried heat sources.

5. Conclusions

Using Laplace-Hankel transformations, the transient closed-form solutions of the thermoelastic consolidation due to an instantaneous point heat source in an isothermal permeable half space are obtained. The results show:

1. The maximum ground surface horizontal displacement is around 38.5% of the maximum ground surface vertical displacement of the isothermal permeable half space at $r = h/\sqrt{2} \approx 0.707h$.
2. It's observed that the positive temperature change increases to a wider region of the half space initially and then gradually decreased. The stratum temperature rise caused by instantaneous point heat finally disappeared, and the elastic deformations due to instantaneous point heat source fully recovered as the temperature increment vanished.

6. Acknowledgements

This work is supported by the National Science Council of Republic of China through grant NSC-98-2815-C-216-007-E, and also by the Chung Hua University under grant CHU-98-2815-C-216-007-E.

7. References

Bai, M. & Abousleiman, Y. (1997). Thermoporoelastic coupling with application to consolidation, *International Journal for Numerical and Analytical Methods in Geomechanics*, Vol. 21, No. 2, pp. 121-132.

Biot, M.A. (1941). General theory of three-dimensional consolidation, *Journal of Applied Physics*, Vol. 12, No. 2, pp. 155-164.

Biot, M.A. (1955). Theory of elasticity and consolidation for a porous anisotropic solid, *Journal of Applied Physics*, Vol. 26, No. 2, pp. 182-185.

Booker, J.R. & Savvidou, C. (1984). Consolidation around a spherical heat source, *International Journal of Solids and Structures*, Vol. 20, No. 11/12, pp. 1079-1090.

Booker, J.R. & Savvidou, C. (1985). Consolidation around a point heat source, *International Journal for Numerical and Analytical Methods in Geomechanics*, Vol. 9, No. 2, pp. 173-184.

Erdelyi, A.; Magnus, W., Oberhettinger, F. & Tricomi, F.G. (1954). *Tables of Integral Transforms*, McGraw-Hill, New York.

Hollister, C.D., Anderson, D.R. & Health, G.R. (1981). Subseabed disposal of nuclear wastes, *Science*, Vol. 213, No. 4514, pp. 1321-1326.

Hudson, J.A., Stephansson, O. & Andersson, J. (2005). Guidance on numerical modelling of thermo-hydro-mechanical coupled processes for performance assessment of radioactive waste repositories, *International Journal of Rock Mechanics and Mining Sciences*, Vol. 42, No. 5/6, pp. 850-870.

Hueckel, T. & Peano, A. (1987). Some geotechnical aspects of radioactive waste isolation in continental clays, *Computers and Geotechnics*, Vol. 3, No. 2/3, pp. 157-182.

Lu, J. C.-C. & Lin, F.-T. (2006). The transient ground surface displacements due to a point sink/heat source in an elastic half-space, *Geotechnical Special Publication No. 148*, *ASCE*, pp. 210-218.

Lu, J. C.-C. & Lin, F.-T. (2007). Thermo-consolidation due to a point heat source buried in a poroelastic half space, *Proceedings of the 3rd IASTED International Conference on Environmental Modelling and Simulation*, pp. 48-53, Honolulu, Hawaii, U.S.A.

Ma, C. & Hueckel, T. (1992). Stress and pore pressure in saturated clay subjected to heat from radioactive waste: A numerical simulation, *Canadian Geotechnical Journal*, Vol. 29, No. 6, pp. 1087-1094.

Ma, C. & Hueckel, T. (1993). Thermomechanical effects on adsorbed water in clays around a heat source, *International Journal for Numerical and Analytical Methods in Geomechanics*, Vol. 17, No. 3, pp. 175-196.

Savvidou, C. & Booker, J.R. (1989). Consolidation around a heat source buried deep in a porous thermoelastic medium with anisotropic flow properties, *International Journal for Numerical and Analytical Methods in Geomechanics*, Vol. 13, No. 1, pp. 75-90.

Smith, D.W. & Booker, J.R. (1996). Boundary element analysis of linear thermoelastic consolidation, *International Journal for Numerical and Analytical Methods in Geomechanics*, Vol. 20, No. 7, pp. 457-488.

Sneddon, I.N. (1951). *Fourier Transforms*, McGraw-Hill, New York, pp. 48-70.

8. Notation of Symbols

$b_i\left(i=r,\theta,z\right)$	Body forces of the poroelastic half space (N/m^3)
c_1	Parameter, $c_1=2\eta Gk\big/\left(2\eta Gn\beta+1\right)\gamma_w$ (m^2/s)
c_2	Parameter, $c_2=\lambda_t/m$ (m^2/s)
c_3	Parameter, $c_3=\left(1-\nu\right)k\big/\left[3\left(1-\nu\right)\alpha_u+\left(1+\nu\right)\alpha_s\right]\gamma_w$ ($^oCm^4/Ns$)
c_a	Parameter, $c_a=\dfrac{c_1}{c_3}-\dfrac{2G\left(1+\nu\right)\alpha_s}{1-2\nu}$ ($N/^oCm^2$)
c_b	Parameter, $c_b=\dfrac{c_1^2}{c_3\left(c_2-c_1\right)}$ ($N/^oCm^2$)
c_c	Parameter, $c_c=\dfrac{c_1c_2}{c_3\left(c_2-c_1\right)}-\dfrac{2G\left(1+\nu\right)\alpha_s}{1-2\nu}$ ($N/^oCm^2$)
c_s	Specific heat of the skeletal material (J/kg^oC)
c_w	Specific heat of the pore water (J/kg^oC)
$erfc\left(x\right)$	Complementary error function (Dimensionless)
G	Shear modulus of the isotropic poroelastic half space (Pa)
h	Buried depth of instantaneous point heat source (m)
$I_\nu\left(x\right)$	Modified Bessel function of the first kind of order ν (Dimensionless)
$J_\nu\left(x\right)$	First kind of the Bessel function of order ν (Dimensionless)
k	Permeability of the isotropic poroelastic half space (m/s)
m	Thermal parameter, $m=\left(1-n\right)\rho_sc_s+n\rho_wc_w$ ($J/^oCm^3$)
n	Porosity of the poroelastic half space (Dimensionless)
p	Excess pore fluid pressure of the isotropic poroelastic half space (Pa)
$\tilde{p}$	Laplace-Hankel transforms of p, Eq. (10c)
Q_0	Strength of instantaneous point heat source (J)
$\left(r,\theta,z\right)$	Cylindrical coordinates system (m, *radian*, m)
s	Laplace transform parameter (s^{-1})
t	Time (s)
$u_i\left(i=r,z\right)$	Displacement components of the poroelastic half space (m)
$u_{r\,max}$, $u_{z\,max}$	Maximum ground surface horizontal/vertical displacement of the poroelastic half space (m)
$\tilde{u}_i\left(i=r,z\right)$	Laplace-Hankel transforms of u_i, Eqs. (10a)-(10b)
α_s	Linear thermal expansion coefficient of skeletal of the stratum ($^oC^{-1}$)
α_u	Linear thermal expansion factor, $\alpha_u=\left(1-n\right)\alpha_s+n\alpha_w$ ($^oC^{-1}$)
α_w	Linear thermal expansion coefficient of pore water ($^oC^{-1}$)
β	Compressibility of pore water (Pa^{-1})
γ_w	Unit weight of pore water (N/m^3)

$\delta(t)$	Dirac delta function (s^{-1})
$\delta(x)$	Dirac delta function (m^{-1})
δ_{ij}	Kronecker delta (Dimensionless)
ε	Volume strain of the poroelastic half space (Dimensionless)
$\varepsilon_{ij}(i,j=r,\theta,z)$	Strain components of the poroelastic half space (Dimensionless)
η	Parameter, $\eta=(1-\nu)/(1-2\nu)$ (Dimensionless)
ϑ	Temperature change of the poroelastic half space (^oC)
$\tilde{\vartheta}$	Laplace-Hankel transform of ϑ, Eq. (10d)
λ_t	Thermal conductivity of the poroelastic half space ($J/sm°C$)
ν	Poisson's ratio of the isotropic poroelastic half space (Dimensionless)
ξ	Hankel transform parameter (m^{-1})
ρ_s	Density of skeletal material (kg/m^3)
ρ_w	Density of pore water (kg/m^3)
$\sigma_{ij}(i,j=r,\theta,z)$	Total stress components of the poroelastic half space (Pa)
∇^2	Laplacian operator, $\nabla^2=\partial^2/\partial r^2+1/r\,\partial/\partial r+\partial^2/\partial z^2$ (m^{-2})

15

Assessment of seismic risk and reliability of road network

Salvatore Cafiso
University of Catania
Italy

1. Introduction

Essential services for road users as well as for every kind of human activity are strongly dependent from road network that is considered a "lifeline" as one of the essential linear infrastructures for human life. When a catastrophic event strikes a wide area, it is necessary that the infrastructure system is designed with a high redundancy and/or low risk of failure to maintain network function to give access for the rescue service.
In the case of damage produced by seismic events, the effects of an interruption to the road network and the consequent reduction of what remains available profoundly affect the overall performance of the system (increasing travelling time, distance and costs).
The road network must be reliable, that is, it must (Wakabayashi, Idia, 1992) "…. provide a safe and not fluctuating service for the traffic and offer the users alternative routes, even when some parts of the system are not available due to road accidents, maintenance or natural disasters".
If network Reliability is able to measure the ability of the system to maintain its performance due to the vulnerability to suffer damages of some of its components, Risk assessment is able to consider other aspects in addition to the overall functionality of the network system such as the consequences in terms of victims that can derive from the reduction of its functionality.
In the chapter a comprehensive methodology framework for the evaluation of the seismic risk and reliability of rural road networks will be presented. These original methodologies make it possible to identify beforehand critical parts of the road network as regards to possible structural damage and the importance of the connection related to the number of inhabitants that could suffer a delay in the emergency services. The analyses were carried out considering bridges as the "weak" element of the road infrastructure in case of seismic events, but the procedure could also be applied to different types of element (trenches, embankments, culverts, etc).
The methodology has been designed for applications based on Geographic Information System (GIS). With reference to different seismic emergency scenarios (road network, seismic intensity, O/D routes), case studies are presented to highlight the possibility of a preventive estimation of towns and links that present different levels of Risk.

2. Risk assessment

Risk assessment aims to define a measure of the risk. Since risk management is subject to large costs and variable benefits, proper risk assessment and management are crucial to make successful actions.
Risk can be defined as: *the combination of the likelihood of an occurrence of a hazardous event and the severity of the consequences (human, social and economical losses) that can be caused by the event*. Therefore, if the measurement of uncertainty refers only to the probabilities of the event occurrence, the measure of risk requires to carry out both the probability for outcomes and the related losses.
Based on this definition, risk assessment of road network can be carried out as the product of three independent factors (Cafiso et al. 2005, Cafiso et al. 2008, Cafiso 2009):
1) **Exposure**, given by the number of people (and/or goods) that can be damaged by the event.
2) **Hazard**, linked to the probability that in a certain place there will be an event of a certain intensity with a given return time;
3) **Vulnerability**, which defines the propensity of an infrastructural element to undergo damage during the event.
In the following a complete definition of each term and the methodology to measure it will be provided.

2.1 Exposure

Seismic exposure represents the *extension, the quantity and quality of the various anthropic elements that make up the territorial context (population, buildings, infrastructure, etc.) whose conditions and operation could be damaged by a seismic event*. The population is the main category at risk and the potential number of injured or dead people is considered as a measure of exposure.
In a seismic risk assessment it is fundamental to consider the "**direct exposure**" of the users of the transportation network beyond the "classical" one of the resident population of the urban buildings. In fact, specially in urban area, on road infrastructures during much of the day a great number of people are exposed to risk as well as those who are inside buildings. It should be remembered that in the Loma Prieta earthquake (USA, 17-10-1989, Magnitude M_{SW} 7.1) the collapse of the viaducts of the busy Cypress Street, in the City of Oakland, caused the highest number of fatalities (42 on a total of 56 deaths) (Figure 1).

Fig. 1. Loma Prieta, sections of the Cypress viaduct (H.G. Wilshire, U.S. Geological Survey)

The "direct exposure" can be related to the number of users present along the infrastructure during the event or to the property value or to the mission of the element in the transportation system. The expected vehicle density (vehicles per unit length of road) can be used as measure of exposure.
Table 1 shows threshold direct exposure levels used in CAPTA (NCHRP report 525, 2009) to determine if a transportation asset exceeds the threshold and will be included in further analysis as a high-consequence asset. The distinct differentiation between potentially exposed populations, property, and mission is highlighted within the equation box.

Asset	Potentially Exposed Population Equation	Property Equation	Mission Equation
Road Bridges	Separated into primary direction and secondary direction - for each, if vehicles/lane > 2400, assume 40 vehicles/1000 ft. Otherwise assume 7.5 vehicles/1000 ft	$20,000/feet	(ADT) x (detour length) 75th, 85th, 95 th percentile thresholds relative to typical bridge inventory (Example is based on the National Bridge Inventory)
Road Tunnels	Separated into primary direction and secondary direction - for each, il vehicles/lane > 2400, assume 40 vehicles/1000 ft. Otherwisw assume 7.5 vehicles/1000 ft	$100,000/feet	User input for criticality
Transit/Rail Station	4 x (maximum capacity of rail cars)	Below ground = critical	User input if transfer station is critical
Transit/Rail Bridge	2 x (maximum capacity of rail cars)	$15,600/feet	User input percentage of ridership that regularly use this transit/rail transportation asset
Transit/Rail Tunnel	2 x (maximum capacity of rail cars)	$40,000/feet	User input percentage of ridership that regularly use this transit/rail transportation asset

Table 1. Threshold consequence determination (NCHRP REPORT 525, 2009)

In the emergency phases that follow a seismic event, the transportation network has the task of making assistance accessible to the stricken area so that aids can be quick and efficient. If road infrastructure efficiency has been compromised due to the effects of the quake, to reach the stricken areas would be impossible or really slow and difficult. Many strong earthquakes have tragically shown the essential role that transportation network has to provide timely emergency services after a seismic event. In the Kobe earthquake (Japan, 17-01-1995, Magnitude M_{SW} 6.92) interruption of the access routes prevented emergency services from reaching the devastated areas for many hours (Figure 2). As consequence, damage from fires indirectly caused by the quake determined consequences comparable to that caused directly by the quake (totally 5.500 fatalities).
Based on these considerations an original definition of "**indirect exposure**" on single stretches of the road network can be defined in relation to the number of people who would experience delays in the arrival of emergency services due to an interruption of that given

stretch of the network. The analysis of indirect seismic exposure consists in a study of estimations of damage or injury to the population that can result from a road network or of a part of it that does not function correctly. It is, obviously, a study of exposure because the object is the analysis of damage or injury suffered by people as an indirect consequence of the bad efficiency of road networks after the event.

Fig. 2. Kobe earthquake, interrupted access route and fires caused by the quake

The level of damage caused by this kind of situation is different according to the dimensions and characteristics of the city that has been hit by the quake. Each town, in fact, has to be considered as a generator of demand for assistance that is proportional to the resident population. Therefore, indirect exposure of a stretch of the road network increases with the growth of the expected number of resident population stroked by the earthquake in the towns linked to it.
The road system as to be analyzed as a network composed by links (i.e. the road stretches) between two nodes representing the intersections. The assignation of indirect exposure to a link of the network follows the following three steps:
1) to each town (Di) destination of the emergency services, an "direct exposure" value is assigned equal to the number of its inhabitants multiplied by the 'index of seismic risk' ranging between 0 to 0.8 that in Italy is defined proportionally to the percentage of expected losses after an earthquake (Ord.P.C.M. n. 2788 del 12/06/1998);
2) defining the route from the origin (O) to the destination (Di), to each of the links along this route, an "indirect exposure" (Ei) is assigned equal to the direct exposure value of the town of destination.
3) the overall indirect exposure of each link of the road network is carried out as sum of the Ei values related to all the destinations (Di) using that link in the O/Di routes.

2.2 Hazard

Seismic hazard defines *the probability of occurrence of a seismic event of certain intensity, in a given area and in a given period of time*. The evaluation of seismic hazard of an area is based on the study of historic seismology and on the analysis of the geological seismologic and seismogenetic characteristics of the site.
The historical studies are aimed at the definition of the principal geophysical characteristics (epicenter, magnitude, ground acceleration etc.) of seismic events that have struck in the past the area under examination, in such a way as to predict the effects of earthquakes expected for different return times in terms of horizontal force, acceleration, etc.

Peak Ground Acceleration (PGA) is used as the seismic hazard measurement parameter. In Italy, this parameter is provided by the National Institute of Geophysics and Volcanology (I.N.G.V.) in terms of PGA with an 81%, 63%, 50%, 39%, 22%, 10%, 5%, 2% probability of exceeding this in 50 years, corresponding to a return time period of 30, 50, 72, 100, 200, 475, 975 and 2475 years.

2.3 Vulnerability

Seismic vulnerability is defined as *the propensity of an element, simple or complex, to suffer damage, collapse or modification during a seismic event*. Seismic vulnerability is an intrinsic characteristic of each construction, that is independent from any kind of external factor. For example, the vulnerability of a bridge depends on the construction technologies adopted, on the materials employed, on its structural configuration, on its age, on its state of maintenance, on the quality of the original project and so on. All these factors are independent from the localization of the object and from the probability that a seismic event can take place there, which has been already evaluated in the study of seismic hazard.
To define the vulnerability of a road segment, it should be considered that each stretch could be composed by a series of components (bridges, embankments, trenches, tunnels…) with different vulnerability characteristics and evaluation models not always comparable.
In the following tables (tables 2a-2d) a selection of the principal elements that need to be considered to evaluate the structural vulnerability for each component, is presented (Cafiso et al., 2005).

Element	**Min. Vulnerability ←**		**→ Max. Vulnerability**
Design criteria	Constructed according to seismic criteria		Constructed without seismic criteria
Construction type	Continuous structures		Discontinuous structures
Regularity of geometry and rigidity	Regular structures		Irregular structures
Pier type	Single		Multiple
Abutment height	Low		High
Soil-foundation system	good		Bad and with liquefaction problems
Condition of the construction	Good state of conservation		Bad state of conservation
Alignment	Low angles of deviation		High angles of deviation
Type of bearing support	If longitudinal and transversal movement is allowed and there are systems to prevent girder fall		Simple support
Expansion joints	None	Present with seismic criteria	Present with a short base
Building material	steel	Reinforced concrete	masonry

Table 2a. Factors for the evaluation of seismic vulnerability of bridges and viaducts

If more than one vulnerable component is present on the same network link, to characterise with only one indicator of vulnerability a stretch of road between to nodes, many criteria of aggregation could be used, among which one of the simplest and of immediate application

consists in giving to the whole stretch the maximum value from among the indicators of structural vulnerability of the components that make it up.

Element	**Min. Vulnerability ←**		**→ Max. Vulnerability**	
Design criteria	Constructed according to seismic criteria		Constructed without seismic criteria	
Height	Low		High	
Geometrical condition of the site	Low inclination of ground	High inclination of ground with seismic wall	High inclination of ground with no seismic wall	High inclination of ground without wall
Embankment geometry	Slope inclination < 2/3		Slope inclination > 2/3	
Soil support characteristics	Good		Bad	
Condition of the structure	Good state of conservation		Bad state of conservation	
Slope protection	Presence of slope protection		No slope protection	

Table 2.b. – Factors for the evaluation index of seismic vulnerability of embankments

Element	**Min. Vulnerability ←**		**→ Max. Vulnerability**
Design criteria	Constructed according to antiseismic criteria		Constructed without antiseismic criteria
Geological and geometric condition of slope	Dynamic safety coefficient Fs ≥ 1.3		Dynamic safety coefficient Fs < 1.3
Length - Height	Trenches low and short		Trenches long and high
Rock fall	Impossible	Possibility of rock fall with slope protection	Possibility of rock fall without slope protection
Retaining structures	built according to antiseismic laws		Not built according to antiseismic laws

Table 2.c. – Factors for the evaluation index of seismic vulnerability of trenches

Element	**Min. Vulnerability ←**	**→ Max. Vulnerability**
Design criteria	Constructed according to antiseismic criteria	Constructed without antiseismic criteria
Geostructural condition of the mass	Good: no earth pressure - absence of discontinuity	Bad: earth pressure - presence of landslides and discontinuities
Deformation joint	Presence of deformation joints	Absence of deformation joints
Location	Deep tunnel	Superficial tunnel
Section area	Small	Large
Section type	Closed section	Open section
Condition of the structure	Good state of conservation	Bad state of conservation

Table 2.d. – Factors for the evaluation index of seismic vulnerability of tunnels

Therefore, basing on experience acquired after a strong Earthquake, a first level macro analysis of large road networks can be conducted considering the vulnerability of only bridges as the weak road structural element when an earthquake strikes (Figure 3).

Fig. 3. Example of bridge collapse after an earthquake

To evaluate the seismic vulnerability of the bridges forming part of a road network, a model (Buckle, Kim, 1995) was selected from the literature because it is particularly effective for the proposed approach:
- it minimizes the arbitrariness of subjective judgement;
- all the parameters indicated by the procedure can be easily identified in the bridges of the area under investigation;
- the way of determining damage as a product of hazard and vulnerability is suitable for the risk evaluation method adopted;
- the model provides a numerical damage index.

In the model (Buckle, Kim, 1995), the level of vulnerability (V) is obtained using a linear regression of the damage indicators recorded during seismic events and related to evaluation parameters present in the model:

$$V = \Sigma i\ \beta i \times Xi \tag{1}$$

where: Xi (i=1,…,12) is the value assumed by the model parameters (Intensity of Peak Ground Acceleration, Design Specification, Type of Superstructure, Shape of Superstructure, Internal Hinge, Type of Pier, Type of Foundation, Material of Substructure, Irregularity in Geometry or in Stiffness, Site Condition, Effect of Liquefaction, Seat Lenght); βi is the weighting factors for each attribute.

2.4 Damage

Once the vulnerability (V) and the Hazard (PGA) are defined, it is possible to obtain the damage index for each bridge by means of the following relation:

$$D = XPGA \times V \quad (2)$$

where XPGA is the hazard index assuming the values shown in Table 3.

PGA	XPGA
PGA < 0,1 g	1
0,1 g < PGA < 0,2 g	2
0,2 g < PGA < 0,3 g	3
PGA > 0,3 g	4

Table 3. hazard index XPGA

From equations (1) and (2), the damage D can assume values of between 0 and 9 (0=no damage, 9=maximum damage). Increasing the bridge damage a reduction in the residual traffic capacity (transitability) of the link can be expected. Usually, for low level of damage (only pavement cracking) without structural failure a low speed traffic can be permitted, for medium level of damage (large cracking, join fault) without structural collapse a limited and controlled traffic (only emergencies services) can be permitted, for high damage (deck unseating) till to the collapse of the structure, traffic is not permitted. Therefore, the loss of transitability Index, ranging from 0 to 10 (0=no traffic limitation, 10=traffic not permitted), is associated to the level of damage with the relation shown in figure 4.

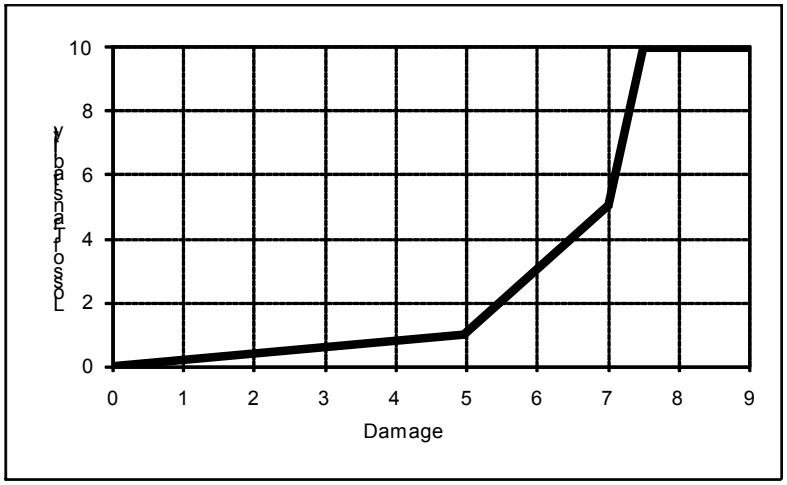


Fig. 4. damage Vs. loss of transitability index Tr

2.5 Risk

The risk factor is associated to the access of emergency services to the towns after an earthquake. Therefore, the road network risk assessment can be defined by associating the Damage (Hazard x Vulnerability) and Exposure factors to the links of the road network.

The procedure comprises the following 5 steps (Cafiso et al, 2008) which refer to different phases of the process and different items of the network system:
Step 1 –Hazard and Vulnerability (item: bridge)
To each bridge in the road network values of vulnerability (V) and XPGA are associated depending to the structural features of the bridge (equation 1) and to the seismic scenario expected in the location of the element (Table 3).
For each bridge an index of Damage is calculated as product of Hazard and Vulnerability (equation 2). Moreover, the loss of Transitability index can be associated to damage (figure 4).
Step 2 - damage and Loss of Transitability (item: link)
When dealing with road links where there are no bridges or overpasses then both Damage and Loss of Transitability were taken as being equal to 0. For those stretches where there is one bridge or overpass the Damage and Loss of Transitability indexes were assigned on the basis of values carried out in step 1. Finally, if there are more than one bridge and/or overpass then the Damage and Loss of Transitability indexes of the stretch were considered as being equal to the maximum of the values attributed to the different bridges or overpasses.
Step 3 - O/D routes (item: town)
After having defined the origin (O) and the destination (D) in the earthquake scenario, the routes connecting each O/D connection can be defined using different criteria (e.g. length, travel time, encountered bridge damage or Loss of Transitability).
Step 4 - Indirect exposure factor (item: link)
An indirect exposure value is assigned to each link of the road network constituting part of the O/D route equal to the number of inhabitants in the town of destination multiplied by its seismic risk index (exposure of the town).
Once all the O/Di routes has been identified for all the "i" destinations, an indirect exposure value can be associated to each link of the network equal to the sum of the values attributed to this link in each of the O/Di.
Therefore, some stretches of the network have a nil exposure, because they have never been used for O/D routes. Others have an exposure value based on a single destination, while those which have been used a number of times in order to reach different destinations have an exposure value equal to the sum of the exposures of the towns for which the stretch is used for that type of route.
Step 5 - Risk evaluation (item: link)
When for each link of the road network the damage value (Step 2) and the indirect exposure (step 4) are carried out, it is possible to obtain the risk value of the link as product of damage and indirect exposure values.

$$\text{Risk} = \text{Indirect exposure} \times \text{Damage}$$

3. Case Study

Referring to a high seismic-risk area of eastern Sicily (Italy) as case study, it was possible to verify the effectiveness of the proposed procedure. The methodology has been designed for applications based on Geographic Information System (GIS). In particular, implementing the method using the GIS made it possible to draw up maps which identify the most critical stretches for different earthquake scenarios and emergency service origins.

3.1 Area of investigation

The province of Catania has an area of about 3,552 Km² in which there are 58 towns/cities with an overall population of 1,054,778 inhabitants.
The GIS contains all the data necessary for an analysis of risk and emergency management, organized in shape-files and relational data bases. All the bridges and overpasses present on the road network were positioned within the GIS using 1:10,000 scale maps of the province. 321 bridges and overpasses situated on stretches of the road network within the study zone were localized. For each bridge "j" a visual inspection was carried out and data were obtained from the Department of Transport and local road Agencies to evaluate the vulnerability parameters Xji of model (1). As result each bridge localized in the GIS has as attribute a vulnerability index Vj

$$Vj = \Sigma i\ \beta i \times Xji \tag{3}$$

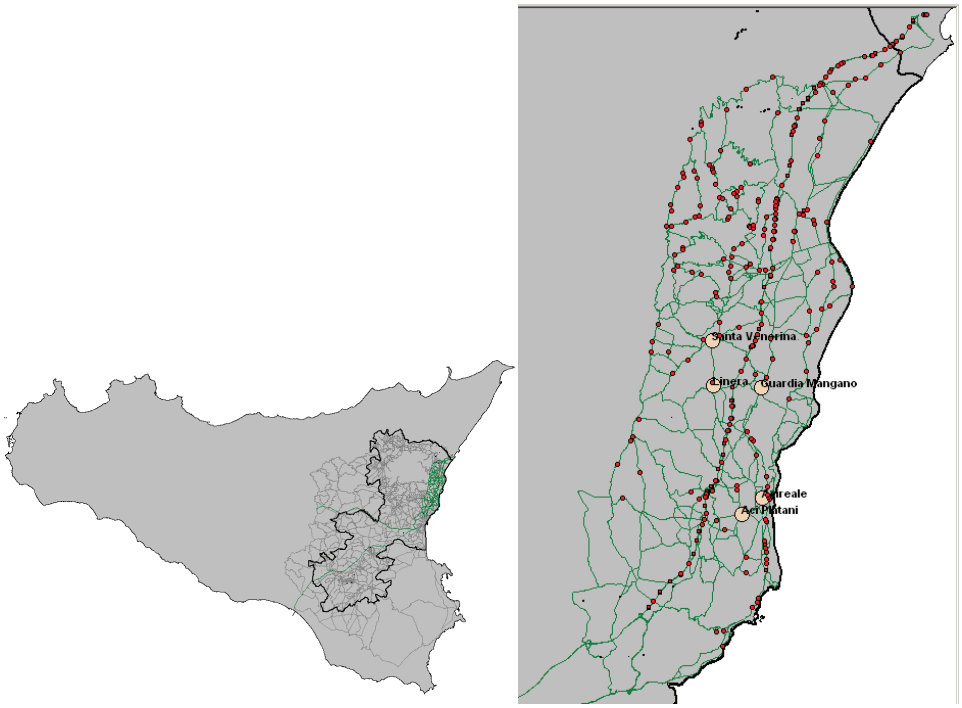


Fig. 5. Area of investigation, road network and bridges localization

Based on the history of seismic events in the investigated area, three levels of hazard (PGA values) were chosen, characterized by different return times:
1) events having a return time period of 50 years for the most frequent shakes (PGA with 50 % probability of exceeding the value in the next 50 years) (Figure 6.a);
2) events having a return time period of 100 years for not particularly severe and localized earthquakes (PGA with 39 % probability of exceeding the value in the next 50 years) (Figure 6.b);

3) events having a return time period of 475 years which corresponds to the strongest seismic events taken into consideration by building regulations (PGA with 10 % probability of exceeding the value in the next 50 years) (Figure 6.c);

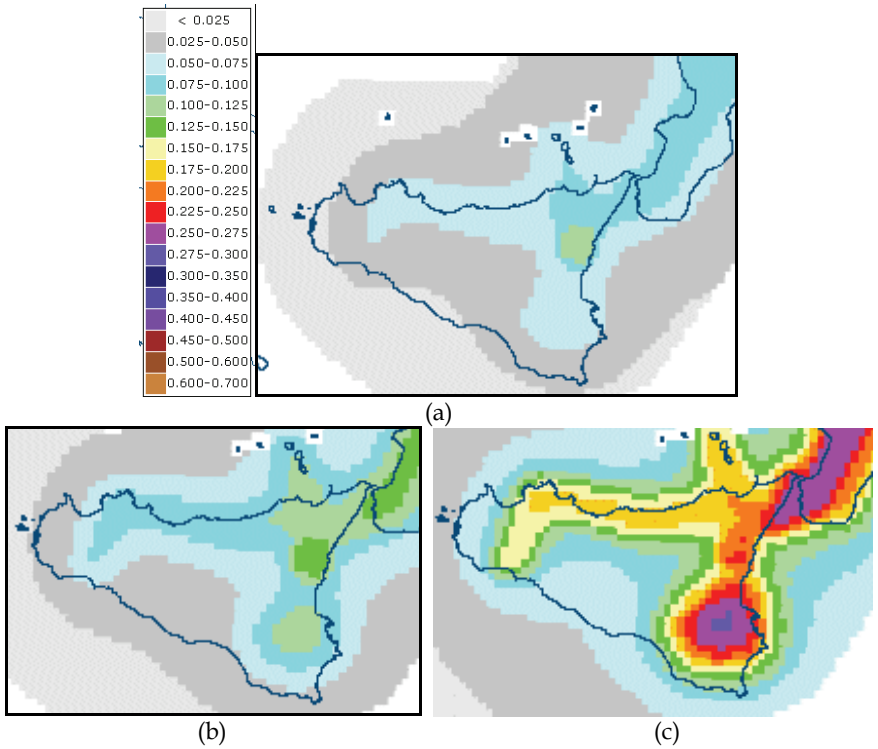


Fig. 6. Maps of PGA with return times of 50 years (a), 100 years (b) and 475 years (c) (Istituto Nazionale di Geofisica e Vulcanologia (I.N.G.V.), Italy)

3.2 Risk assessment

The methodology has been designed for applications based on Geographic Information System (GIS). The case study presentation is subdivided in four phases that represents the different activities carried out in the process to asses the seismic risk of the road network.

Phase 1 - damage and Loss of Transitability index

Using specific GIS tools, it was possible to attribute a specific PGA value to each bridge previously localized in the area (Figure 7) using an "overlay analysis" procedure on the area values drown from the seismic maps of I.N.G.V..

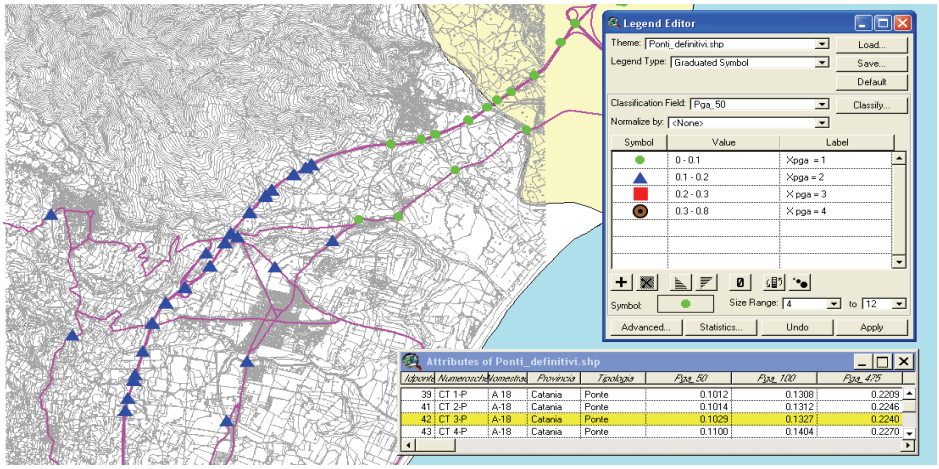


Fig. 7: 50, 100 and 475-year PGA maps of bridges on the road network

Once the vulnerability (Vj) was defined and the Hazard (XjPGA) noted, the adopted model made it possible to obtain the expected damage index for each bridge by means of the following relation:

$$Dj = XjPGA \times Vj$$

From the relation showed in figure 4, also a value of Loss of Transitability can be associated as attribute to each bridge.

When dealing with stretches where there are no bridges or overpasses then both Damage and Loss of Transitability were taken as being equal to 0 for any earthquake scenario (50, 100 and 475 years). For those stretches where there is one bridge or overpass the Damage and Loss of Transitability indexes were assigned on the basis of values taken from the previously-illustrated model for the various earthquake scenarios at 50, 100 and 475 years (figure 5). Finally, if there is more than one bridge and/or overpass then the Damage and Loss of Transitability indexes of the stretch were considered as being equal to the maximum of the values attributed to the different bridges or overpasses.

Phase 2 - O/D routes

Two different emergency service origins were chosen:

1) Origin North (O_N): the motorway from the town of Messina that represent the connection of Sicilia Island to the continental part of Italy;

2) Origin West (O_W): the interchange in the Catania urban Freeway, relating to emergency services coming from eastern Sicily and the southern part of the province of Catania.

As regards destinations, 5 towns in the study area were considered:

Acireale (D_1: 33,010 inhabitants), Santa Venerina (D_2: 4,056 inhabitants), Aciplatani (D_3: 3,269 inhabitants), Linera (D_5: 2,781 inhabitants), Guardia Mangano (D_6: 2,457 inhabitants).

The shape-file relating to the roads present in the province of Catania and the segmentation of the network as connected links makes it possible to use a GIS tool able to define the best route from the Origin (O_N or O_W) to the Destination (D_i). The best rout is the one among all

the alternatives which minimize the cost of transport obtained as sum of the cost attributes of each link composing the route (figure 8).
After having defined the origin (O) and the destination (D) in the earthquake scenario, 4 different routes can be identified for each O/D connection using as cost function Length, Time, Damage and Transitability attributes at 50, 100 and 475 years previously assigned to the links of the road network.

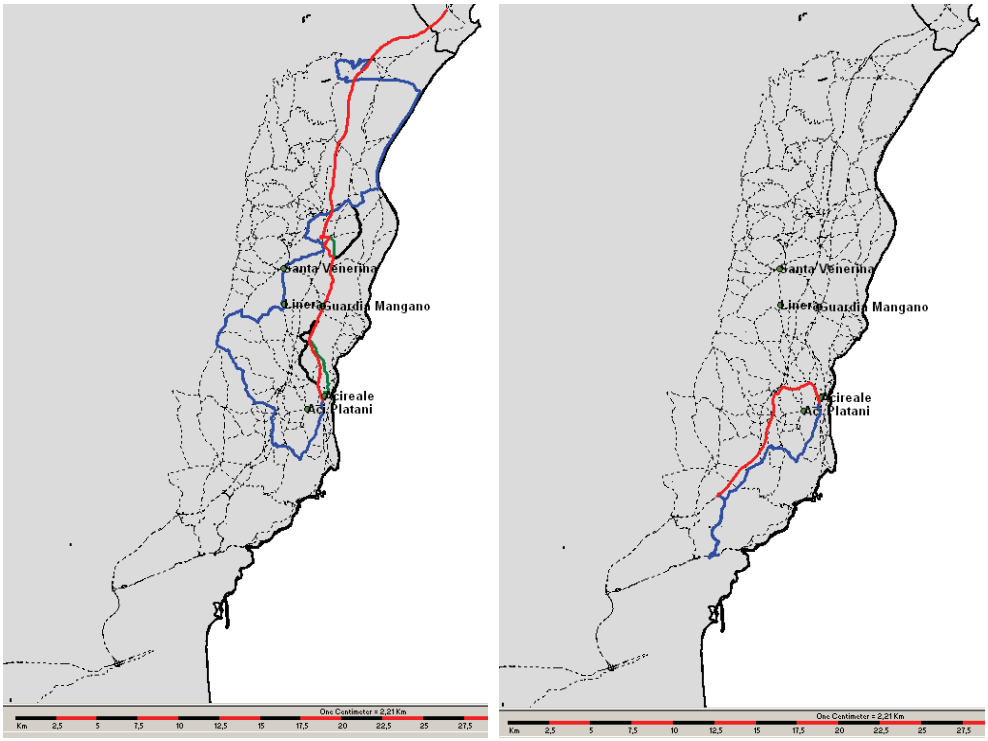


Fig. 8. Origin Nord and West: minimum length routes (red line), minimum time (green), minimum damage (blue) and minimum loss of transitability (black)

Phase 3 - Indirect exposure factor
An indirect exposure value is assigned to each link of the road network constituting part of the O/D route equal to the number of inhabitants in the town of destination (D_i) multiplied by its seismic risk index (exposure of the town).
Once all the O/Di routes of the same type has been identified for all the "i" destinations, an overall indirect exposure value can be associated to each link of the network equal to the sum of the values attributed to the link in each of the O/Di.
Therefore, some stretches of the network have a nil exposure, because they have never been used for O/D routes. Others have an exposure value based on a single destination, while those which have been used a number of times in order to reach different destinations have an exposure value equal to the sum of the exposures of the towns for which the stretch is used for that type of route (figure 9).

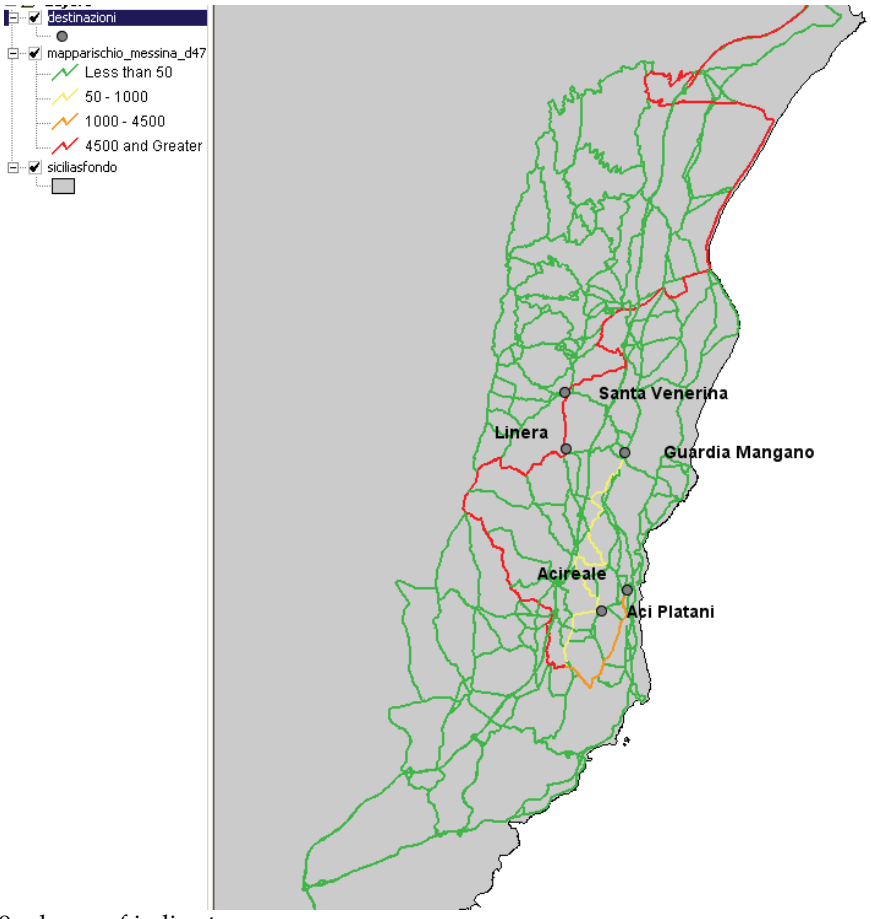


Fig. 9. classes of indirect exposure

Phase 4 - Risk evaluation of the links

When the damage value (Phase 2) and the indirect exposure of each single link of the network (Phase 3) are carried out it is possible to obtain the risk value relating to that particular route, by multiplying the damage value by the exposure value.

$$\text{Risk} = \text{Indirect exposure} \times \text{Damage}$$

In GIS environment, risk maps can be drawn up for each of the origins. The thematic maps graphically highlight those road network stretches having the highest risk index (Figures 10, 11)

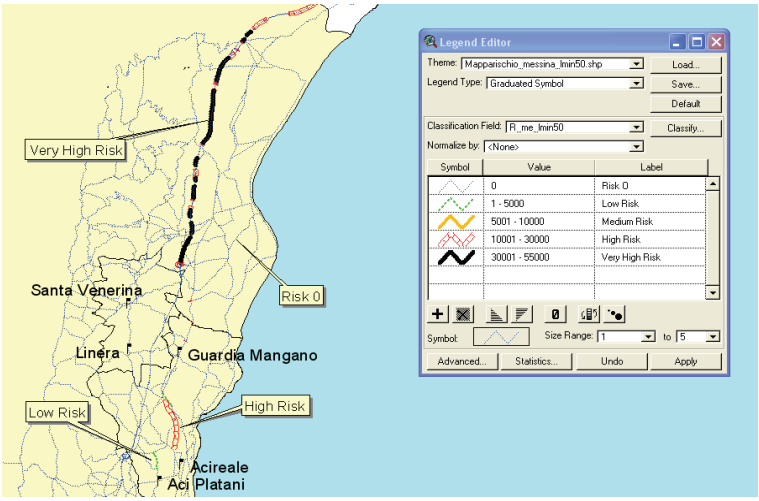


Fig. 10. Thematic risk map for the minimum length route (Cafiso et al, 2008)

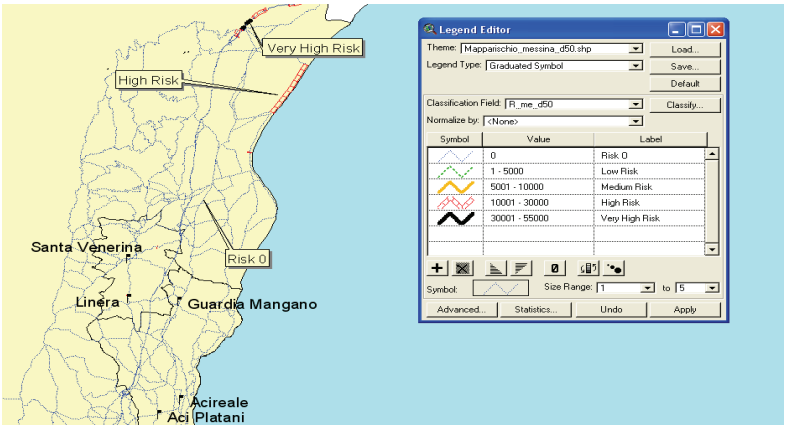


Fig. 11. Thematic risk map for the route with minimum damage (Cafiso et al, 2008)

4. Lifeline Reliability

In general terms, Reliability can be defined as the "*the probability of a device performing its purpose adequately for the period of time intended under the operating conditions encountered*" [1].
A road network, in particular, will be reliable if "…. provides a safe and not fluctuating service for the traffic and offers the users alternative routes, even when some parts of the system are not available due to road accidents, maintenance or natural disasters"
The transport network is affected by two different phenomena that can modify its reliability:
1) Variation in what is offered for transportation;
2) Variation in the demand on the transport services.
In the case of damage produced by seismic events the effects of the interruption of the local network and the consequence reduction in what remains available affect the overall

performance of the system (increase in travel time, distance and costs). In some cases it could be of primary importance that the journey finishes in a determined period of time, while in other cases it is more important to evaluate if there are interruptions along the route that could obstruct to reach the destination [Selçuk, Yücemen, 1999 - Du, Nicholson, 1997).

With this aim the following two different terms of reliability can be defined:

- **Terminal Reliability** is "the probability that nodes are connected, i.e. it is possible to reach the destination" and this is surely the parameter that is easier to evaluate.
- **Encountered Reliability** is "the probability of not encountering a link degradation on the path with least (expected) cost".

Another concept that is complimentary to the previous, is the reliability of the time and cost of the journey, commonly defined as "the probability that a trip can be successfully finished within a specified time interval".

4.1. Encountered Reliability

For the attribution of an Encountered Reliability the travel-length and -time of the route from an origin (Oj) to a destination (Di) which minimize the overall cumulative expected damage after an earthquake (post event route) are compared with the original best route with the minimum travel time and length (travel cost) without considering any interruption (pre event route).

For each destination, the index of Encountered Reliability can be obtained by the formula [10]:

$$E(O_j)_{R,Di} = \frac{l(O_j)_{\min,Di}^{\ 2}}{l(O_j)_{E\min,Di}^{\ 2}} \tag{4}$$

where:

Di = Destination town for emergency services

l(Oj)min,Di = minimum cost (in terms of time or length) from the origin (Oj) to the destination Di (pre event route);

l(Oj)Emin,Di = travel cost (in terms of time or length) related to the route, from the origin (Oj) to the destination Di, along which there are the minimum level of expected bridge damages (post event route).

Using this approach $E(Oj)_{R,Di}$ is always less or equal to 1 and, therefore, the most reliable routes are characterized by higher values of $E(Oj)_{R,Di}$ in as much as the alternative for the emergency services is not much longer than the direct route.

Once the values of $E(Oj)_{R,Di}$ have been defined from equation (4) for each origin for the emergency services (Nord, West, etc.), it is possible to establish a general value for the Encountered Reliability for each town $ER_{TOT}(Di)$, as an average weighted on the itineraries with respect to the different origins:

$$E_{RTOT(Di)} = \frac{\sum_j p_j E(O_j)_{RDi}}{\sum_j p_j} \tag{5}$$

where:

Oj= J=1, .. n and represents the n origins of the emergency services that can serve the destination Di;
pj= weight factor related to the importance of Origin Oj;
From the relationship between the direct exposure values for each town (Di) and the relative total value of $E_{RTOT(Di)}$ it is possible to obtain an Encountered Risk factor RE(Di) related to each town (Destination) of the area under investigation (Figure 12):

$$RE(D_i) = \frac{\text{Direct exposure(Di)}}{\text{E}_{\text{RTOT,Di}}} \tag{6}$$

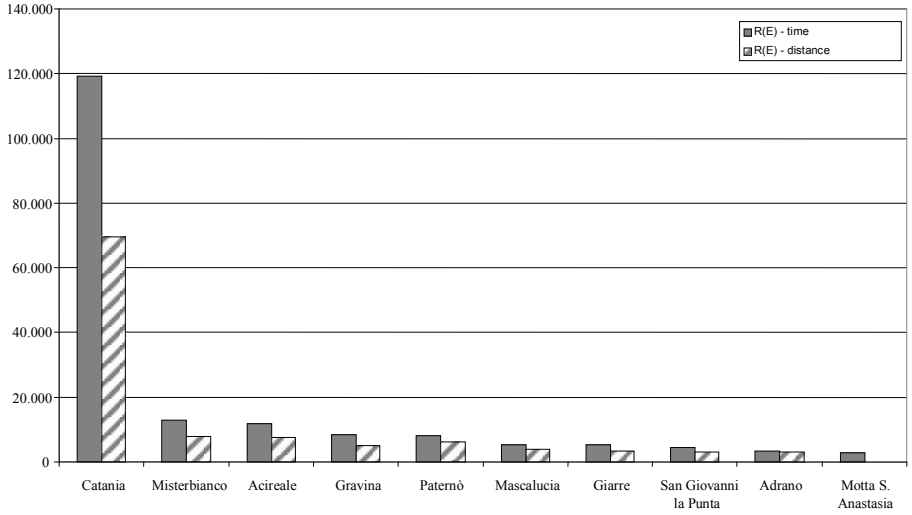


Fig. 4. Risk factors of Encountered Reliability related to different destinations (towns in East of Sicily, IT) (Cafiso et al., 2004)

4.2. Terminal Reliability

For the definition of Terminal Reliability the routes that lead, from the Origin (Oj), to the destination (Di), bypassing the bridges with defined level of damage are considered. The iterative procedure stops either when there are no more bridges to by pass along the alternative route and the destination is reached or it is no longer possible to reach the destination.
The index of Terminal Reliability, for the entire road network can be obtained by the formula:

$$T(O_j)_{R,Di} = \frac{l(O_j)_{\min,Di}^{\ 2}}{l(O_j)_{T\min,Di}^{\ 2}} \tag{7}$$

where:
Di = Destination town for emergency services

l(Oj)min,Di = minimum cost (in terms of time or length) from the origin (Oj) to the destination Di (pre event route);
l (Oj)Tmin,Di = travel cost (in terms of time or length) related to the route, from the origin (Oj) to the destination Di, (post event route) along which there are only bridges with damage lower than a defined value.
Using this approach $T(Oj)_{R,Di}$ is always less or equal to 1 and, therefore, the most reliable routes are characterized by higher values of $T(Oj)_{R,Di}$.
Once the values of $T(Oj)_{R,Di}$ have been defined from equation (7) for each origin for the emergency services (Nord, West, etc.), it is possible to establish a general value for the Terminal Reliability for each town $T_{RTOT}(Di)$, as an average weighted on the itineraries with respect to the different origins:

$$T_{RTOT(Di)} = \frac{\sum_j p_j T(O_j)_{RDi}}{\sum_j p_j} \tag{8}$$

where:
Oj= J=1, .. n and represents the n origins of the emergency services that can serve the destination Di;
pj= weight factor related to the importance of Origin Oj;

From the relationship between the direct exposure values for each town (D) and the relative total value of $T_{RTOT,Di}$ it is possible to obtain the risk factor R (Figure 13):

$$RT(D_i) = \frac{\text{Direct exposure}(D_i)}{T_{RTOT,Di}} \tag{9}$$

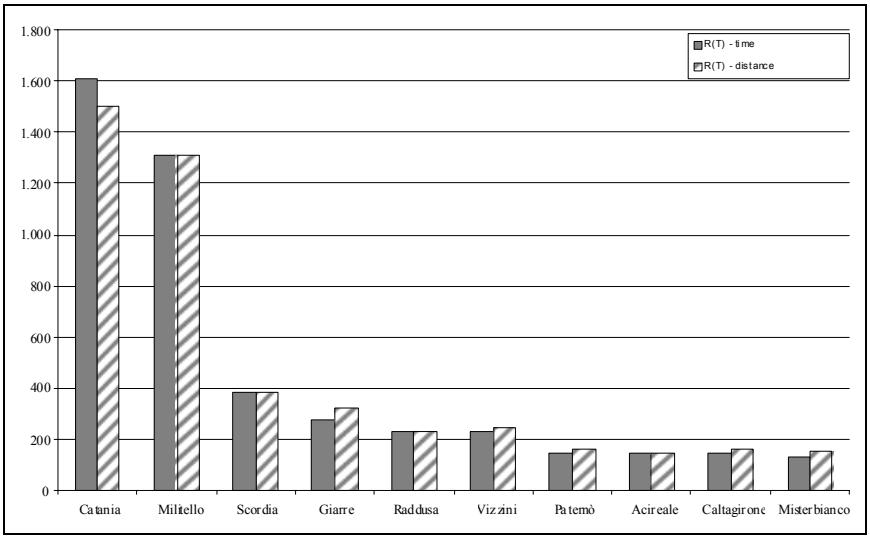


Fig. 13. Risk factors connected to Terminal Reliability (Cafiso et al., 2004)

If the previous definition of Risk (paragraph 3) is associated to the network links, Encountered and Terminal Reliability Risk factors are related to the town of destination given an evaluation of the probability to encounter an obstruction in the route to reach the destination.

5. Conclusions

The maintenance of an efficient road network after an earthquake is fundamental if emergency services from outside the area have to reach the struck towns as easily and quickly as possible. Therefore, risk and reliability assessment of road network are indispensable to evaluated beforehand, so as to program seismic retrofitting works to the links which are strategic to the efficient functioning of the road network.
An original methodology to conduct risk assessment is presented, which makes it possible to identify the links of the road network with a higher level of risk both as regards to possible structural damage and the importance of the connection related to the number of inhabitants that can be reached by the emergency services. The analyses were carried out considering bridges as the "weak" element of the road infrastructure in cases of seismic events, but the procedure could also be applied to different types of element (trenches, embankments, culverts, etc).
Using a high seismic-risk area of eastern Sicily as a case study, it was possible to verify the effectiveness of the proposed procedure. In particular, implementing the method using a GIS software made it possible to draw up maps which identify the most critical stretches for different earthquake scenarios (return times of 50, 100, 475 years) and emergency service origins.
Also the concepts of Encountered and Terminal Reliability can be used to identify the routes that lead to specific destination from the origins of emergency services, crossing the minimum level of expected damage of bridges, both in terms of length and time to cover the given distance. These values can be referred to the towns of destination for the emergency services allowing the definition of a risk index relative to the accessibility of the town in case of earthquakes.
If the previous definition of Risk (paragraph 3) is associated to the network links, Encountered and Terminal Reliability Risk factors are related to the town of destination given an evaluation of the probability to encounter an obstruction in the route to reach the destination.
This information are useful in order to identify those parts of the road network where more resources should be employed both to program retrofitting work on structures and for a more in-depth analysis of the system.

6. References

Buckle I.G., Kim S.H., A vulnerability assesment for highway bridge. ASCE. *Lifeline Earthquake Engineering*, 1995.

Cafiso S., Condorelli A., Cutrona G., Mussumeci G. A Seismic Network Reliability Evaluation on GIS Environment - A Case Study on Catania Province. *Risk Analysis IV*, 2004, pagg. 131-140 - WIT Press, – ISBN 1-85312-736-1.

Cafiso S., Condorelli A., D'Andrea A.. Methodological Considerations for the Evaluation of Seismic Risk on Road Network. *Pure and Applied Geophysics*, Vol. 162, n. 4, 2005. pagg.767-782.

Cafiso S., La Bruna A., La Cava G. "Seismic Risk Assessment of Rural Road Network". *Risk Analysis, Simulation and Hazard Mitigation V*. pp.91-100. WITpress, Southampton, United Kingdom, 2008. ISBN: 978-1-84564-104-7.

Cafiso S. "Seismic Risk and Reliability for Rural Road Network". *IASTED International Conference on Modelling, Simulation and Identification*, 2009, Pechino, Cina.

Du Z.P., Nicholson A.J., Degradable Transportation System: Sensitivity and Reliability Analysis, *Transportation Research B*, 31(3), 225-237, 1997.

Istituto Nazionale di Geofisica e Vulcanologia (I.N.G.V.). Mappa di pericolosità sismica del territorio nazionale PCM n. 3519, All. 1b. April 2006 in http://zonesismiche.mi.ingv.it/

Ord.P.C.M. n. 2788 del 12/06/1998, Individuazione delle zone ad elevato rischio sismico del territorio nazionale, 1998.

NCHRP REPORT 525 - Costing Asset Protection: An All Hazards Guide for Transportation Agencies (CAPTA), Vol. 15, ISBN 978-0-309-11763-0, TRB 2009

Selçuk A.S., Yücemen M.S., Reliability of lifeline networks under seismic hazard. Reliability Engineering and System Safety, pp. 213-227, 1999

Wakabayashi H., Idia Y., Upper and Lower bounds of terminal reliability of road networks: an efficient method with Boolean Algebra. Journal of Natural Disaster Science 14, pp. 29-44, 1992.

16

Modelling and simulation of the dynamic behavior of an oil wave journal bearing

Nicoleta M. Ene, Florin Dimofte and Abdollah A. Afjeh
The University of Toledo
U.S.A.

1. Introduction

The dynamic stability of journal bearings is an important problem for rotating machinery because the dynamic properties of the bearing have a direct influence on machine stability and safety.
Because of their simplicity and large capacity, the plain journal bearings are frequently used in rotating machinery; however they can become unstable under small loads and have the tendency to generate a whirl motion with a frequency of about one half the rotational frequency of the shaft. The radius of the whirl orbit can rapidly increase so that the shaft and the sleeve can come into contact, a phenomenon that would damage the bearing. Therefore the study of the dynamic behavior of the journal bearings, especially after the instability occurs, is of theoretical and practical importance.
The bearing stability can be improved by adding grooves and holes or by reshaping the bearing surface from perfectly circular to one that incorporates lobes, offsets, tilting pads, etc. The main disadvantage of these methods is that any gain in the bearing stability may reduce the maximum load carrying capacity.

2. The Wave Bearing Concept

An alternative method to improve journal bearing stability was proposed by Dimofte (Dimofte 1995 a; Dimofte 1995 b). His concept called "wave bearing" circumscribed a continuous waved profile onto the non-rotating bearing surface. The wave amplitude is usually a fraction of the nominal bearing clearance. To exemplify the concept, a comparison between a wave bearing having circumscribed a three–wave profile and a plain journal bearing is presented in Fig. 1. In order to visualize the concept, the wave amplitude and the clearance are greatly exaggerated in Fig. 1.
The most important parameters of a wave bearing are presented in Fig. 2. In Fig. 2, a wave bearing having a three-wave profile is presented. The radial clearance C of the wave bearing is defined as the difference between the radius of the mean circle of the waves R_{med} and the radius, R, of the shaft:

$$C = R_{med} - R \tag{1}$$

The radial clearance is usually less than one thousandth of the journal radius.

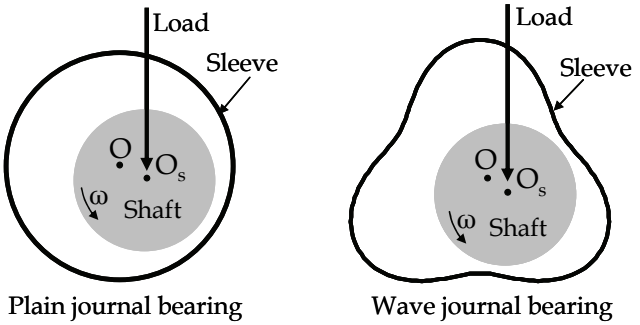


Fig. 1. Comparison between the plain journal bearing and the wave journal bearing

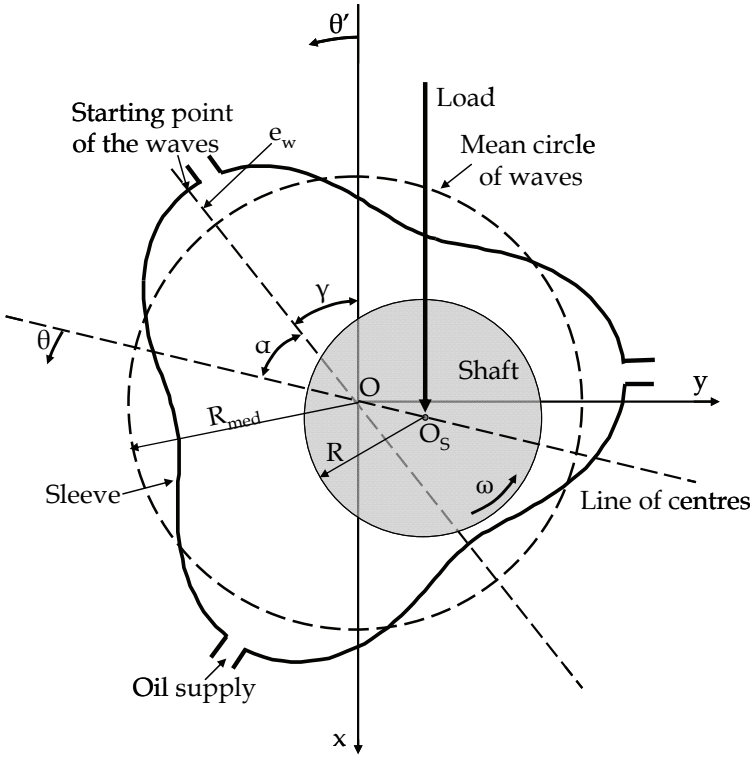


Fig. 2. The geometry of a wave journal bearing

For computational purposes, the wave amplitude is usually non-dimensionalised by dividing it by clearance:

$$\varepsilon_w = \frac{e_w}{C} \tag{2}$$

The ratio ε_w is usually called the wave amplitude ratio. The wave amplitude ratio is one of the most important geometrical characteristics of a wave bearing because the performance of the wave bearing is strongly influenced by this ratio (Ene et all., 2008). The wave amplitude ratio has values that generally range between 0.033 and 0.5.
The performance of a wave bearing also depends on the number of the waves, n_w, and on the position of the waves relative to the direction of the load, W. Theoretical and experimental studies indicate that the best performance is obtained by a three-wave bearing having one of the points with maximum wave amplitude on the direction of the load (Dimofte, 1995 a, Dimofte, 1995 c).
The load capacity of a wave bearing is due to the rotation of the shaft and to the variation of film thickness along the circumference. With some geometrical considerations, it can be shown that in a system of coordinates rotating with the line of centres the fluid film thickness h is given by:

$$h = C + e\cos\theta + e_w \cos\left[n_w(\theta + \alpha)\right] \tag{3}$$

where θ is the angular coordinate starting from the line of centres and α is the angle between the starting point of the waves and the line of centres. The film thickness can be also expressed in a system of reference Oxy fixed with respect to the sleeve:

$$h = C + e_w \cos\left[n_w(\theta' - \gamma)\right] + x_s \cos\theta' + y_s \sin\theta' \tag{4}$$

where θ' is the angular coordinate starting from the negative Ox axis, γ is the angle between the starting point of the waves and the vertical axis and (x_s, y_s) are the coordinates of the rotor centre.
Because the shape of a wave bearing is very close to the shape of a plain journal bearing, the loss of load capacity of the wave bearing compared to the plain journal bearing is minimal (Dimofte, 1995 a).
The wave bearing concept also includes a number of supply pockets equal to the number of the waves (see Fig. 2). They are situated near the points where the waves have maximum values.

3. Methods for Simulating the Dynamic Behavior of a Wave Journal Bearing

Since the early paper of Newkirk and Taylor (Newkirk & Taylor, 1925), a considerable number of papers have been devoted to the study of the dynamic stability of the journal bearings. Two types of approaches can be identified:
- critical mass approaches based on small-perturbation theories;
- transient approaches based on linear or non-linear theories.

The critical mass approach has been very popular because of its simplicity and limited computational requirements. The main disadvantage of this method is that no bearing information can be obtained after the appearance of an unstable whirl. The bearing dynamic behavior for unstable conditions can be predicted only by using transient methods. The main disadvantage of the transient methods is that they require a large amount of computational time.

Calculation of the critical mass implies in the first stage the computation of the dynamic coefficients. Two methods can be used to compute the dynamic coefficients:
- numerical differentiation of the pressure distributions with respect to small perturbations of displacements and velocities of the journal centre;
- perturbation methods.

The numerical differentiation method requires correct identification of the values of the small perturbations so that they are small enough to remain within the limits of the linear theory, but large enough to produce significant perturbations with respect to numerical errors and approximations. (Frêne et al., 1997). The numerical differentiation method can be used to calculate the dynamic coefficients for different types of journal bearings: circular journal bearings (Orcutt & Arvas, 1967; Parkins, 1979), tilting pad journal bearings (White & Chan, 1992), etc.

The problem of correct selection of the small perturbation values can be eliminated by using a perturbation method. The perturbation method was first introduced by Lund (Lund & Thomsen, 1978; Lund 1984; Kilt & Lund, 1986). It consists of solving five partial differential equations, one corresponding to the steady-state pressure distribution and the other four to the pressure gradients. The dynamic coefficients are then calculated by integrating the pressure gradients distributions. Lund's approach has been used by many authors to compute the dynamic characteristics of different types of bearings. For example, Lund's method was used by Kostrzewsky et al. (Kostrzewsky et al., 1998) to compute the dynamic characteristics of highly preloaded three-lobe journal bearings. In order to reduce the computation time, the authors assumed a polynomial form for the axial pressure distribution. Kakoty and Majumdar (Kakoty & Majumdar, 1999; Kakoty & Majumdar, 2000) studied the effect of fluid inertia on the stability of oil film journal bearings also using a linear perturbation analysis. Rao and Sawicki adapted Lund's infinitesimal procedure so that the film content at cavitation rupture and deformation boundaries is taken into consideration for both a steady-state pressure distribution and dynamic pressure gradients. Rao and Sawicki used this method to investigate the stability characteristics of journal bearings (Rao & Sawicki, 2002) and herringbone grooved journal bearings (Rao & Sawicki, 2004).

Many papers have investigated the dynamic stability of journal bearings using a transient approach. One of the first transient analyses of a journal bearing was performed by Kirk and Gunter (Kirk & Gunter, 1976a; Kirk & Gunter, 1976b). Because of the computation limitations, they determined the fluid film forces by using a short bearing approximation.

Monmousseau and Fillon (Monmousseau & Fillon, 1999) used a non-linear transient approach to analyze the dynamic behavior of a tilting–pad journal bearing submitted to a synchronous and a non-synchronous load. The authors showed that the amplitude of the shaft orbit is maximum when the loading frequency is near the critical frequency.

San Andres (San Andres, 1997) compared the transient responses of a rigid rotor supported on externally pressurized, turbulent fluid film bearings obtained using two different models: an approximate model based on constant rotordynamic coefficients and a full nonlinear model. He concluded that the approximate model provided accurate results only for small amplitude loadings and for operating conditions far enough from the stability margin of the rotor bearing system.

Tieu and Qiu (Tieu & Qiu, 1995) presented a comparison between the journal centre trajectories of a journal bearing computed using both non-linear and linear theory. Both

methods provided the same critical speed. However, under large dynamic excitations, the trajectories obtained with the two methods were significantly different.
Tichy and Bou-Said (Tichy & Bou-Said, 1991) and Hashimoto and Wada (Hashimoto & Wada, 1990) used transient methods to emphasize the effects of turbulence on the dynamic response of rotors supported in journal bearings.
Vijayaraghavan and Brewe (Vijayaraghavan & Brewe, 1992) predicted the trajectories of a hydrodynamic journal bearing when a unidirectional external periodic load is applied. The authors reported that under periodic loads, the journal centre almost always attains a stable limit cycle. When the loading frequency is half of the journal frequency, the fluid film forces become very large and the journal centre whirls at large eccentricities.
In this paper both a critical mass approach and a transient method will be used to predict the dynamic behavior of a three-wave bearing in the absence of any external load.

4. Critical Mass Approach

4.1 Governing equations

The small perturbation theory assumes that the shaft motion is stable and is limited to small perturbations around the static equilibrium. Suppose that in the fixed system of reference Oxy (see Fig. 2), the journal centre position is characterized in the steady-state regime by coordinates (x_{s0}, y_{s0}). The corresponding film thickness will be denoted by h_0. Because the shaft motion is limited to small perturbations, the shaft position in the dynamic regime (x_s, y_s) can be described by small amplitudes from the static equilibrium:

$$\begin{aligned} x_s &= x_{s0} + \Delta x \\ y_s &= y_{s0} + \Delta y \end{aligned} \tag{5}$$

By combining Eqs. 4 and 5, a relation between the film thicknesses corresponding to the steady state (h_0) and dynamic regimes (h) is obtained:

$$h = h_0 + \Delta x \cos\theta' + \Delta y \sin\theta' \tag{6}$$

Similarly, the dynamic pressure p can be expressed as a first order Taylor expansion around the steady-state pressure p_0:

$$p = p_0 + p_x \Delta x + p_y \Delta y + p_{\dot{x}} \Delta \dot{x} + p_{\dot{y}} \Delta \dot{y} \tag{7}$$

The pressure distribution in the wave journal bearing is described by the Reynolds equation:

$$\frac{1}{R^2}\frac{\partial}{\partial\theta}\left(\frac{h^3}{k_\theta \mu}\frac{\partial p}{\partial\theta}\right) + \frac{\partial}{\partial z}\left(\frac{h^3}{k_z \mu}\frac{\partial p}{\partial z}\right) = \frac{\partial h}{\partial t} + \frac{\omega}{2}\frac{\partial h}{\partial\theta} \tag{8}$$

where p is the dynamic pressure, h - the fluid film thickness, R - the bearing radius, ω – the rotational speed, μ – lubricant viscosity, t-time, θ –angular coordinate, z – axial coordinate, and k_θ, k_z are correction coefficients for turbulent flow. The Reynolds equation can be

obtained from the general Navier-Stocks equations by considering that the dimension upon the film thickness is very small compared to the two other directions.
By introducing the pressure expansion (Eq. 7) and the film thickness equation (Eq. 6) into the Reynolds equation (Eq. 8), developing a series and retaining only first order terms, five partial differential equations are obtained:

$$\frac{1}{R^2}\frac{\partial}{\partial\theta'}\left(\frac{h^3}{k_\theta\mu}\frac{\partial p_0}{\partial\theta'}\right)+\frac{\partial}{\partial z}\left(\frac{h^3}{k_z\mu}\frac{\partial p_0}{\partial z}\right)=\frac{\omega}{2}\frac{\partial h}{\partial\theta'} \tag{9a}$$

$$\frac{1}{R^2}\frac{\partial}{\partial\theta'}\left(\frac{h^3}{k_\theta\mu}\frac{\partial p_x}{\partial\theta'}\right)+\frac{\partial}{\partial z}\left(\frac{h^3}{k_z\mu}\frac{\partial p_x}{\partial z}\right)=-\frac{\omega}{2}\left(\sin\theta'+3\frac{\cos\theta'}{h}\frac{\partial h}{\partial\theta'}\right)-\frac{h^3}{4\mu R^2}\frac{\partial p_0}{\partial\theta'}\frac{\partial}{\partial\theta'}\left(\frac{\cos\theta'}{h}\right) \tag{9b}$$

$$\frac{1}{R^2}\frac{\partial}{\partial\theta'}\left(\frac{h^3}{k_\theta\mu}\frac{\partial p_y}{\partial\theta'}\right)+\frac{\partial}{\partial z}\left(\frac{h^3}{k_z\mu}\frac{\partial p_y}{\partial z}\right)=\frac{\omega}{2}\left(\cos\theta'-3\frac{\sin\theta'}{h}\frac{\partial h}{\partial\theta'}\right)-\frac{h^3}{4\mu R^2}\frac{\partial p_0}{\partial\theta'}\frac{\partial}{\partial\theta'}\left(\frac{\sin\theta'}{h}\right) \tag{9c}$$

$$\frac{1}{R^2}\frac{\partial}{\partial\theta'}\left(\frac{h^3}{k_\theta\mu}\frac{\partial p_{\dot{x}}}{\partial\theta'}\right)+\frac{\partial}{\partial z}\left(\frac{h^3}{k_z\mu}\frac{\partial p_{\dot{x}}}{\partial z}\right)=\cos\theta' \tag{9d}$$

$$\frac{1}{R^2}\frac{\partial}{\partial\theta'}\left(\frac{h^3}{k_\theta\mu}\frac{\partial p_{\dot{y}}}{\partial\theta'}\right)+\frac{\partial}{\partial z}\left(\frac{h^3}{k_z\mu}\frac{\partial p_{\dot{y}}}{\partial z}\right)=\sin\theta' \tag{9e}$$

For the above equations, the Reynolds boundary conditions at the film rupture zones are:

$$p_0=\frac{\partial p_0}{\partial\theta'}=\frac{\partial p_0}{\partial z}=0 \qquad p_x=p_y=p_{\dot{x}}=p_{\dot{y}} \tag{10}$$

The boundary conditions at the axial ends of the bearing are:

$$p_0=p_{ext} \qquad p_x=p_y=p_{\dot{x}}=p_{\dot{y}} \tag{11}$$

where p_{ext} is the external pressure. Similarly, the boundary conditions corresponding to the supply pockets are:

$$p_0=p_s \qquad p_x=p_y=p_{\dot{x}}=p_{\dot{y}} \tag{12}$$

where p_s is the supply pressure.
It can be seen that Eq. 9a corresponds to the steady-state regime. If the external force is vertical, then the equilibrium equations are:

$$\begin{bmatrix} -W \\ 0 \end{bmatrix} = R \int_{-\frac{L}{2}}^{\frac{L}{2}} \int_{0}^{2\pi} p_0 \begin{bmatrix} \cos\theta' \\ \sin\theta' \end{bmatrix} d\theta' dz \tag{13}$$

The solutions of Eqs. 9b-d allow one to determine the stiffness and damping coefficients of the bearing:

$$\begin{bmatrix} K_{xx} & K_{xy} \\ K_{yx} & K_{yy} \end{bmatrix} = R \int_{-\frac{L}{2}}^{\frac{L}{2}} \int_{0}^{2\pi} \begin{bmatrix} p_x \cos\theta' & p_y \cos\theta' \\ p_x \sin\theta' & p_y \sin\theta' \end{bmatrix} d\theta' dz \tag{14}$$

$$\begin{bmatrix} B_{xx} & B_{xy} \\ B_{yx} & B_{yy} \end{bmatrix} = R \int_{-\frac{L}{2}}^{\frac{L}{2}} \int_{0}^{2\pi} \begin{bmatrix} p_{\dot{x}} \cos\theta' & p_{\dot{y}} \cos\theta' \\ p_{\dot{x}} \sin\theta' & p_{\dot{y}} \sin\theta' \end{bmatrix} d\theta' dz \tag{15}$$

The upper limit of the stability is given by the critical mass:

$$m_{cr} = \frac{K_s}{\gamma_s^2} \tag{16}$$

where K_s is the effective bearing stiffness:

$$K_s = \frac{B_{xx}K_{yy} + B_{yy}K_{xx} - B_{xy}K_{yx} - B_{yx}K_{xy}}{B_{xx} + B_{yy}} \tag{17}$$

and γ_s is the instability whirl frequency:

$$\gamma_s = \sqrt{\frac{(K_{xx} - K_s)(K_{yy} - K_s) - K_{xy}K_{yx}}{B_{xx}B_{yy} - B_{xy}B_{yx}}} \tag{18}$$

The critical mass delimitates two possible equilibrium cases:
1. The rotor mass is smaller than the critical mass ($m < m_{cr}$). The rotor centre returns to its static equilibrium position. Particularly, in absence of any external load, the rotor centre rotates with a small radius around the bearing centre. The radius depends on the shaft run-out. In this case, the operating point is stable.
2. The rotor mass is greater than the critical mass ($m > m_{cr}$). The rotor centre leaves its static position and the equilibrium point is unstable. In this case, the method does not allow one to predict the motion of the journal centre.

4.2 Turbulence model

Constantinescu's model of turbulence (Constantinescu et all., 1985; Frêne & Constantinescu, 1975), which is based on Prandtl mixing length hypothesis, was chosen to model the turbulence effects. According to this model, the first signs of turbulence appear when the mean Reynolds number, Re_m, is equal to the critical Reynolds number, Re_{cr}:

$$Re_m = Re_{cr} \tag{19}$$

where:

$$Re_m = \frac{2\rho q}{\mu} \tag{20}$$

$$Re_{cr} = \min\left(41.2\sqrt{\frac{R}{h}}, 2000\right) \tag{21}$$

and q is the total flow. The flow becomes turbulent when:

$$Re_m = 2\,Re_{cr} \tag{22}$$

With these assumptions, the coefficients for turbulent flow are given by:

$$\begin{aligned} k_\theta &= 12 + 0.0136\,Re_{eff}^{0.9} \\ k_z &= 12 + 0.0044\,Re_{eff}^{0.9} \end{aligned} \tag{23}$$

where:

$$Re_{eff} = \begin{cases} 0 & Re_m < Re_{cr} \\ \left(\frac{Re_m}{Re_{cr}} - 1\right)\frac{\rho R\omega h}{\mu} & Re_{cr} \le Re_m \le 2Re_{cr} \\ \frac{\rho R\omega h}{\mu} & Re_m > 2Re_{cr} \end{cases} \tag{24}$$

4.3 Model for thermal effects

Due to computational time considerations, a constant mean temperature is assumed throughout the film. The value of the mean temperature is obtained from a global energy balance on the bearing. An adiabatic model is considered for the energy balance. According to the adiabatic model, all the energy dissipated in the fluid film is convected away by the lubricant. Consequently, the heat generated by friction causes only an increase of the lubricant temperature. Therefore, the increase of the lubricant temperature (the difference

between the temperature of the lubricant entering the film and the constant mean temperature of the film) is given by:

$$\Delta T = \frac{F_f R\omega}{\rho c_v q_{lat}} \tag{25}$$

where c_v is the lubricant specific heat, q_{lat} is the rate of lateral flow and F_f is the friction force. The friction force can be obtained by integrating the friction stresses on the bearing surfaces:

$$(\tau)_{0,h} = \frac{\mu V}{h}\left(1 + 0.0012\,\mathrm{Re}_{eff}^{0.94}\right) \pm \frac{h}{2R}\frac{\partial p}{\partial \theta} \tag{26}$$

4.4 Numerical approach

The first problem that must be solved before evaluating the critical mass is to determine the equilibrium position. At the equilibrium, in absence of any external force, the fluid film force must be vertical and equal to the bearing weight. The fluid film force can be calculated by integrating the steady-state pressure distribution (Eq. 13) where the steady-state pressure distribution is described by the steady-state Reynolds equation (Eq. 9a). In the present paper the steady-state Reynolds equation is successively solved for different positions of the shaft until the fluid film force is vertical and equal to the shaft weight. A bisection algorithm was developed for this purpose.

For every position of the shaft, the turbulence correction coefficients are determined by successive iterations. Thus, for each journal centre position, at the first iteration the pressure distribution is determined by assuming that the flow is laminar (i.e., the effective Reynolds number is zero and the correction coefficients are 12). From the computed pressure distribution, new values of the correction coefficients at every grid point are determined. Also, the new mean film temperature (Eqs. 25-26) and the new lubricant properties are determined. Then the steady-state Reynolds equation (Eq. 9a) is integrated again for the new values of the correction coefficients and the new mean film temperature. The iterative process is repeated until the relative errors for the correction coefficients are smaller than a prescribed value (10^{-5}).

The steady-state Reynolds equation (Eq. 9a) is discretized with a finite difference scheme. The resultant system of equations is solved with a successive over-relaxation method (the Gauss–Seidel method).

Having the equilibrium position of the shaft and the correction coefficients corresponding to this position, the pressure gradients can now be determined by integrating the differential equations corresponding to the pressure gradients (Eqs. 9b-9e) with a finite difference scheme. The dynamic coefficients are then calculated by integrating the pressure gradients over the entire film (Eqs. 14-15) and next the critical mass is obtained with the equations Eqs. 16-18.

5. Model for simulating the bearing dynamic behavior with a transient method

5.1 Governing equations

Without any external force, the equations of motion along and perpendicular to the line of centres are:

$$\begin{aligned} m\left[\frac{d^2e}{dt^2}-e\left(\frac{d\phi}{dt}\right)^2\right] &= F_r + mg\cos\phi + m\omega^2\rho\cos(\omega t-\phi) \\ m\left(e\frac{d^2\phi}{dt^2}+2\frac{d\phi}{dt}\frac{de}{dt}\right) &= F_\phi - mg\sin\phi + m\omega^2\rho\sin(\omega t-\phi) \end{aligned} \tag{27}$$

where F_r and F_ϕ are the radial and tangential components of the fluid force, ρ - the shaft run-out, 2m- the rotor mass, and ω - the rotational velocity. The components of the fluid film force can be determined by integrating the pressure distribution:

$$\begin{aligned} F_r &= R\int_{-L/2}^{L/2}\int_0^{2\pi} p\cos\theta\, d\theta\, dz \\ F_\phi &= R\int_{-L/2}^{L/2}\int_0^{2\pi} p\sin\theta\, d\theta\, dz \end{aligned} \tag{28}$$

The pressure distribution in this case is obtained by solving the Reynolds equation written in the following form:

$$\frac{1}{R^2}\frac{\partial}{\partial\theta}\left(\frac{1}{k_\theta}\frac{h^3}{\mu}\frac{\partial p}{\partial\theta}\right)+\frac{\partial}{\partial z}\left(\frac{1}{k_z}\frac{h^3}{\mu}\frac{\partial p}{\partial z}\right)=\mu\left(\dot{e}\cos\theta+e\dot{\phi}\sin\theta+\frac{\omega}{2}\frac{\partial h}{\partial\theta}\right) \tag{29}$$

The above form of the Reynolds equation was obtained by introducing the expression for the wave bearing film thickness (Eq. 3) into Eq. 8. For Eq. 29 the Reynolds boundary conditions at the film rupture zones are:

$$p=\frac{\partial p}{\partial\theta}=\frac{\partial p}{\partial z}=0 \tag{30}$$

Constantinescu's model of turbulence is also used for this approach to model the turbulent flow.

5.2 Numerical approach

The trajectory of the journal centre is obtained by integrating in time the equations of motion (Eqs. 27). At each time step, a pressure distribution corresponding to the motion parameters (e, ϕ, $\dot{e}$, $\dot{\phi}$) and correction coefficients for turbulent flow (k_θ and k_z) from the previous moment of time is first obtained. The pressure distribution is found by integrating the

Reynolds equation (Eq. 29) using a central difference scheme combined with a Gauss-Seidel method.
Then a new set of correction coefficients (Eqs. 23) corresponding to the new pressure distribution is calculated. Next an energy balance is performed and a new mean film temperature is obtained, Eq. (25). The lubricant properties (viscosity, density and specific heat) are then updated for the new mean film temperature. Afterwards, the Reynolds equation is integrated again for the new values of the correction coefficients and lubricant viscosity. The iterative process is repeated until the relative errors for the correction coefficients and for mean temperature are smaller than prescribed values. Furthermore, the fluid film forces are calculated by integrating the final pressure distribution over the entire film (Eqs. 28). All the parameters of the equations of motion (Eqs. 27) are now known so they can be integrated to determine the new position of the journal centre. A fourth order Runge–Kutta algorithm is used to integrate the motion equations. The algorithm is repeated until the orbit of the journal centre is completed.

6. Numerical simulations

Both the critical mass and the transient approaches are used to study the dynamic behavior of a three-wave bearing having a length of 27.5 mm, a radius of the mean circle of the waves of 15 mm, and a clearance of 35 microns. The rotor mass corresponding to one bearing is 0.825 kg. Synthetic turbine oil Mil-L-23699 was used as a lubricant. The numerical predictions are compared to experimental data (Dimofte et al., 2004).
In real machinery, the rotor always has a small unbalance. This unbalance can be modeled as a small run-out. For this reason, two types of transient simulations were performed: the "ideal case" with zero run-out and "the real case" with a small run-out of 2 microns. For the critical mass approach, the shaft unbalance can not be taken into consideration.
Different wave amplitude ratios, oil supply pressures and inlet temperatures were considered for simulations. The experimental studies showed that for a wave amplitude ratio of 0.305 the wave bearing is stable even at speeds of 60000 rpm and supply pressures of 0.152 MPa. For example, the FFT analysis and the wave shape of the signal from one of the proximity probes corresponding to a rotational speed of 60000 rpm are presented in Fig. 3. FFT analysis shows the presence of an amplitude peak only at the synchronous frequency. In addition, the regular shape of the proximity probe signals indicates also a harmonic motion. The same conclusion can be drawn from the numerical simulations. The variation of the critical mass with the rotational speed, as it was predicted by the small perturbation theory, is shown in Fig. 4. It can be seen that the critical mass is greater than the bearing mass for speeds up to 60000 rpm. Consequently, the bearing is stable for speeds up to 60000 rpm. The trajectory of the shaft centre is predicted with the transient approach. For example, the trajectory of the shaft centre for a rotational speed of 60000 rpm and a zero run-out is presented in Fig. 5. It can be seen from Fig. 5 that the journal centre approaches very rapidly to the bearing centre and orbits around it with a very small radius. When a run-out is considered, the journal centre rotates with one frequency around the bearing centre on a closed orbit having the radius approximately equal to the run-out. For example, the trajectory of the journal centre for a rotational speed of 60000 rpm and a run-out of 2 microns is shown in Fig. 6. The shaft centre motion in horizontal direction (Fig. 7) indicates a

harmonic motion. The FFT analysis of the shaft centre motion in horizontal direction (Fig.8) shows that the frequency of the journal centre motion is equal to the rotor speed.

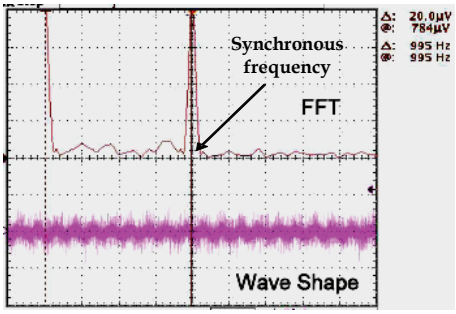


Fig. 3. FFT analysis and wave shape of the experimental signal for $\varepsilon_w = 0.305$, n=60000 rpm, p_s=0.152 MPa

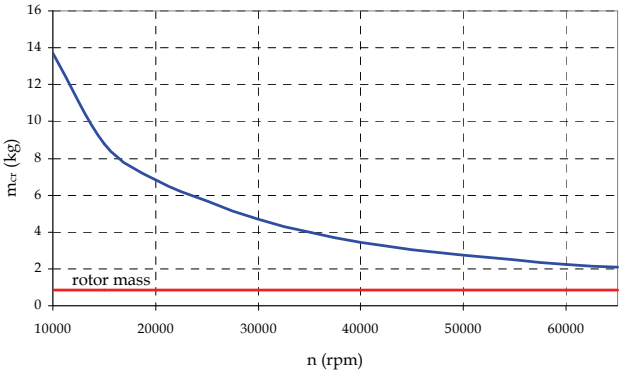


Fig. 4. Critical mass as function of running speed for $\varepsilon_w = 0.305$ and p_s=0.152 MPa

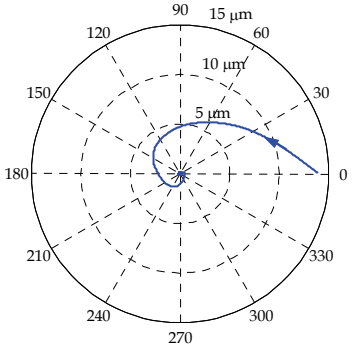


Fig. 5. Trajectory of the journal centre for $\varepsilon_w = 0.305$, n=60000 rpm, p_s=0.152 MPa , and zero run-out

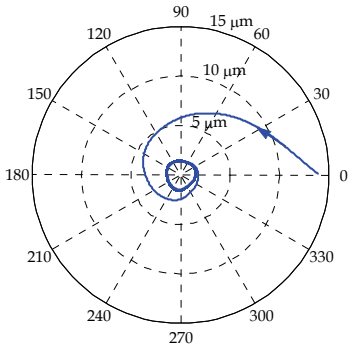


Fig. 6. Trajectory of the journal centre for $\varepsilon_w = 0.305$, n=60000 rpm, p_s=0.152 MPa , and 2 microns run-out

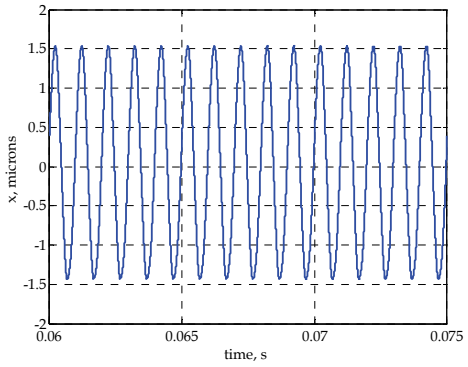


Fig. 7. The position of the shaft centre in the horizontal direction for $\varepsilon_w = 0.305$, n=60000 rpm, p_s=0.152 MPa , and 2 microns run-out

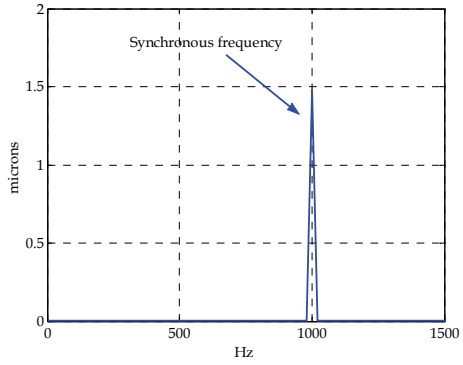


Fig. 8. FFT analysis of the motion in the horizontal direction for $\varepsilon_w = 0.305$, n=60000 rpm, p_s=0.152 MPa , and 2 microns run-out

For wave amplitude ratios smaller than 0.305, the experiments and the numerical simulations show that the rotor centre of the analyzed wave bearing can experience an unstable motion at rotational speeds that depend on the wave amplitude ratio and oil supply pressure. For example, the variation of the critical mass with the rotational speed for a wave amplitude ratio of 0.075, a supply pressure of 0.276 MPa at an oil temperature inlet of 126° C is presented in Fig. 9. It can be seen that the critical mass is greater than the mass of the shaft related to one bearing for speeds smaller than 39000 rpm. The critical mass is very close to the rotor mass around 39000 rpm and then it becomes smaller than the rotor mass. Consequently, it may be concluded that the fluid film of the wave bearing is unstable for rotational speeds greater than 39000 rpm.

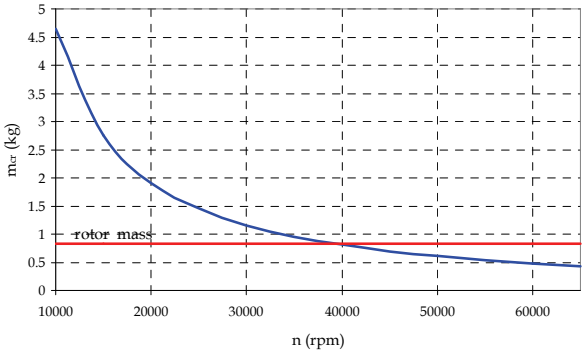


Fig. 9. Critical mass as function of running speed for $\varepsilon_w = 0.075$ and p_s=0.276 MPa

The transient analysis allows for the examination of the post whirl orbit details. The stable trajectories of the journal centre rotating at 36000 rpm with zero and 2 microns run-out are presented in Figs. 10 and 11. The FFT analyses of the numerical predicted motion (Fig. 12) and of the experimental signals from the proximity probes (Fig. 13) indicate the presence of only the synchronous frequency. The wave shapes of the motion in the horizontal direction obtained experimentally (Fig. 13) and from numerical simulations (Fig. 14) also suggest a harmonic motion.

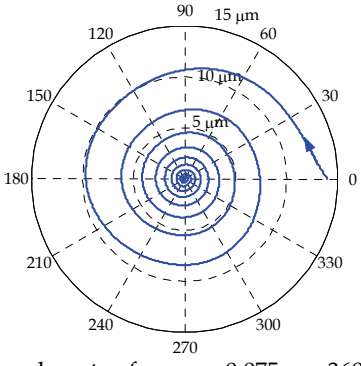


Fig. 10. Trajectory of the journal centre for $\varepsilon_w = 0.075$, n=36000 rpm, p_s=0.276 MPa , and zero run-out

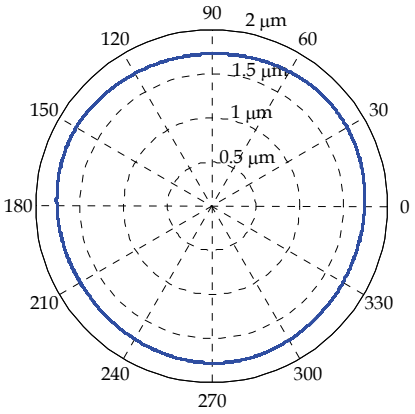


Fig. 11. Trajectory of the journal centre for $\varepsilon_w = 0.075$, n=36000 rpm, p_s=0.276 MPa , and 2 microns run-out

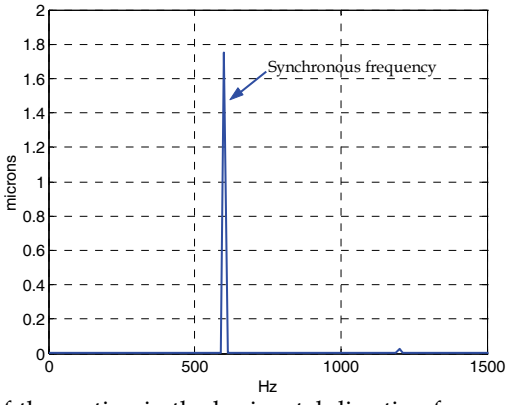


Fig. 12. FFT analysis of the motion in the horizontal direction for $\varepsilon_w = 0.075$, n=36000 rpm, ps=0.276 MPa , and 2 microns run-out

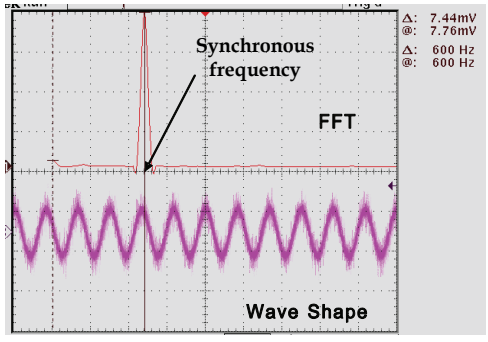


Fig. 13. FFT analysis and wave shape of the experimental signal for $\varepsilon_w = 0.075$, n=36000 rpm, p_s=0.276 MPa

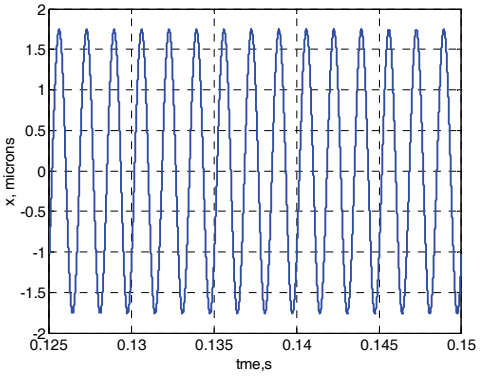


Fig. 14. The position of the shaft centre in the horizontal direction for $\varepsilon_w = 0.075$, n=36000 rpm, p_s=0.276 MPa , and 2 microns run-out

If the speed is increased to the stability threshold (39000 rpm, in this case) an incipient sub-synchronous motion can be detected. The FFT analysis and the wave shape of the signal from the proximity probes are presented in Fig. 15. In this case, both the synchronous and sub-synchronous frequencies can be identified. However, the synchronous frequency is still dominant. The simulated journal centre motion for zero unbalance is presented in Fig. 16. In this case the journal centre rotates on a closed orbit. The FFT analysis of the motion predicts only the sub-synchronous frequency (Fig. 17). The presence of the synchronous frequency can be predicted by the numerical simulations only if the rotor unbalance is taken into consideration. When the run-out is introduced in simulations, the journal centre rotates on a limit cycle with two frequencies (Fig. 18). The motion of the journal centre in the horizontal direction (Fig. 19) and the FFT analysis of the motion (Fig. 20) indicate the existence of both synchronous and sub-synchronous frequencies. They are very similar to those predicted by experiments (Fig. 15).

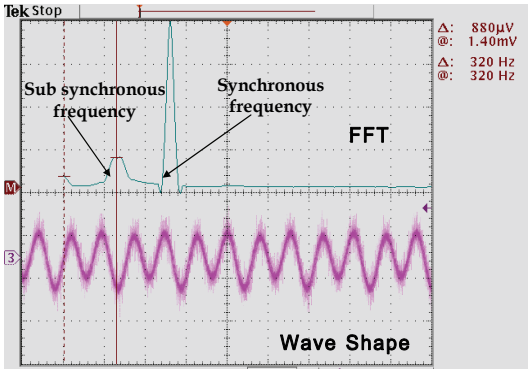


Fig. 15. FFT analysis and wave shape of the experimental signal for $\varepsilon_w = 0.075$, n=39000 rpm, ps=0.276 MPa

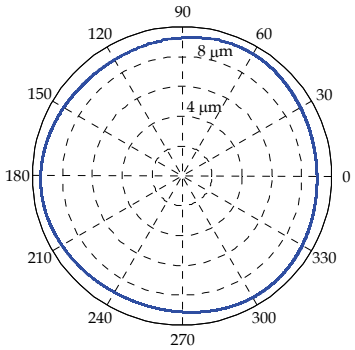


Fig. 16. Trajectory of the journal centre for $\varepsilon_w = 0.075$, n=39000 rpm, p_s=0.276 MPa , and zero run-out

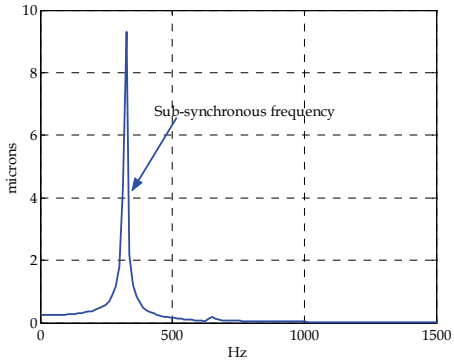


Fig. 17. FFT analysis of the motion in the horizontal direction for $\varepsilon_w = 0.075$, n=39000 rpm, p_s=0.276 MPa , and zero run-out

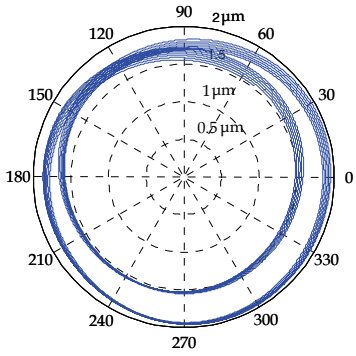


Fig. 18. Trajectory of the journal centre for $\varepsilon_w = 0.075$, n=39000 rpm, p_s=0.276 MPa , and 2 microns run-out

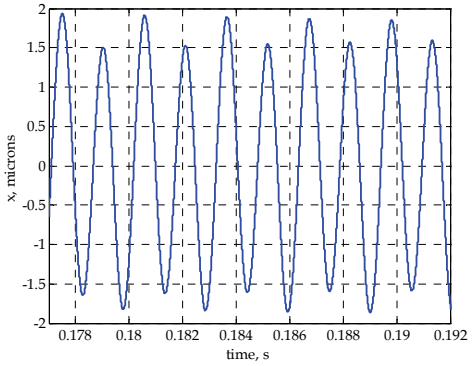


Fig. 19. The position of the shaft centre in the horizontal direction for $\varepsilon_w = 0.075$, n=39000 rpm, p_s=0.276 MPa , and 2 microns run-out

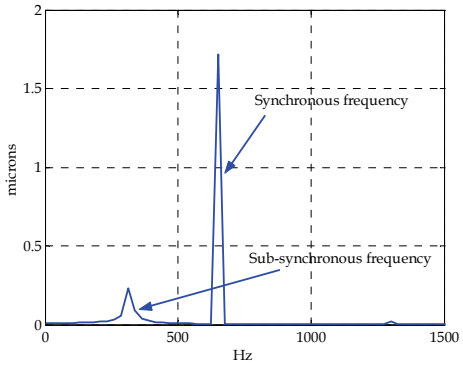


Fig. 20. FFT analysis of the motion in the horizontal direction for $\varepsilon_w = 0.075$, n=39000 rpm, p_s=0.276 MPa , and 2 microns run-out

The experiments show that the sub-synchronous frequency becomes dominant for rotational speeds greater than 44000 rpm (Fig. 21). However, the synchronous frequency is still present. The numerically predicted journal centre trajectories for a rotational speed of 44000 rpm without and with run-out are presented in Figs. (22) and (23). In both cases, the journal centre moves on closed orbits. Again, the simulation corresponding to the motion without run-out predicts only the sub-synchronous frequency (Fig. 24). The synchronous frequency is predicted only by the simulation that takes into account the small run-out (Fig. 25). In this case, the wave shape of the motion in the horizontal direction (Fig. 26) is also very similar to that predicted by experiments (Fig. 21). An increase of the oil supply pressure to 0.414 MPa stabilizes the fluid film of the bearing (Fig. 27). The theoretical analysis shows that the bearing can run stable up to 60000 rpm. The critical mass becomes greater than the rotor mass (Fig. 28). The FFT analyze of the numerical simulated motion for a rotational speed of 60000 rpm and a run-out of 2 microns is presented in Fig. (29). It can be seen that the sub-synchronous frequency disappeared. The corresponding trajectory and the motion in the horizontal direction are also presented in Figs. (30) and (31). If the run-out is zero, then the

limit cycles disappear and the journal approaches to the bearing centre and orbits around it with a very small radius (Fig. 32).
It can be noticed from all the above simulations that the wave bearing journal centre maintains its trajectory inside the bearing clearance, even for unstable motions.

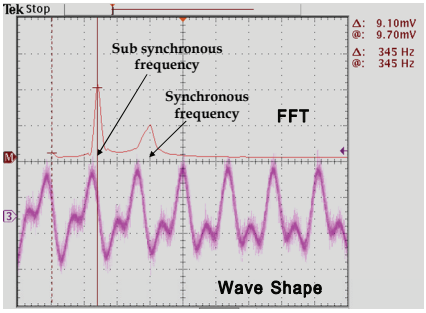


Fig. 21. FFT analysis and wave shape of the experimental signal for $\varepsilon_w = 0.075$, n=44000 rpm, p_s=0.276 MPa

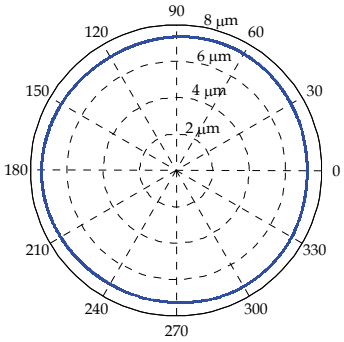


Fig. 22. Trajectory of the journal centre for $\varepsilon_w = 0.075$, n=44000 rpm, p_s=0.276 MPa , and zero run-out

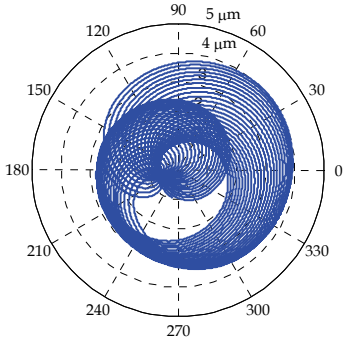


Fig. 23. Trajectory of the journal centre for $\varepsilon_w = 0.075$, n=44000 rpm, p_s=0.276 MPa , and 2 microns run-out

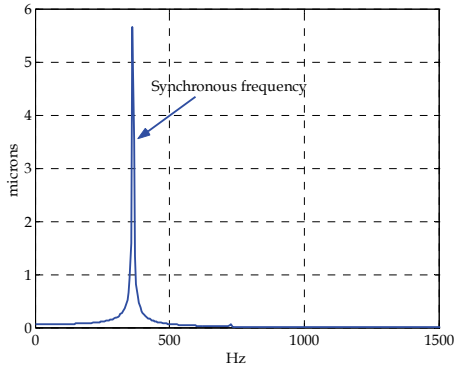


Fig. 24. FFT analysis of the motion in the horizontal direction for $\varepsilon_w = 0.075$, n=44000 rpm, ps=0.276 MPa , and zero run-out

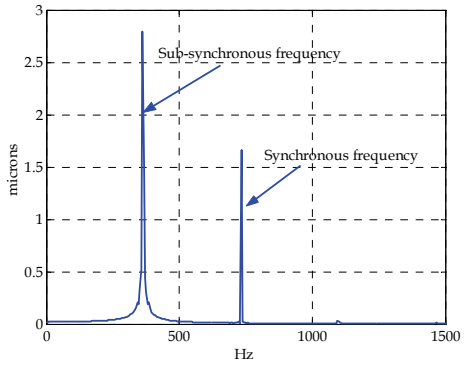


Fig. 25. FFT analysis of the motion in the horizontal direction for $\varepsilon_w = 0.075$, n=44000 rpm, p_s=0.276 MPa , and 2 microns run-out

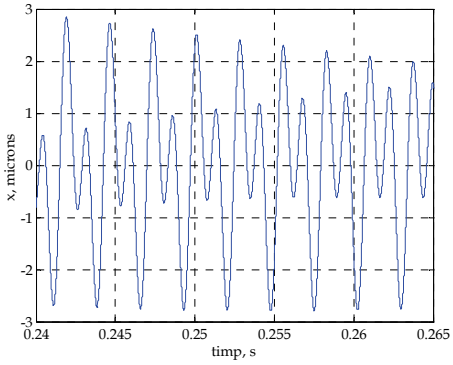


Fig. 26. The position of the shaft centre in the horizontal direction for $\varepsilon_w = 0.075$, n=44000 rpm, p_s=0.276 MPa , and 2 microns run-out

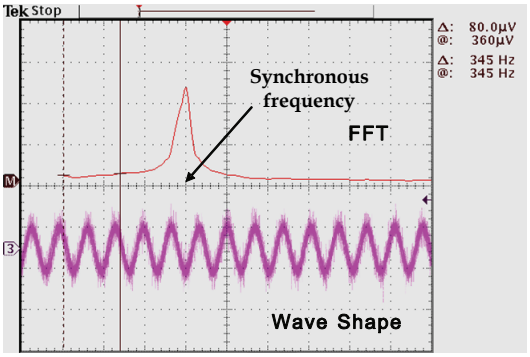


Fig. 27. FFT analysis and wave shape of the experimental signal for $\varepsilon_w = 0.075$, n=44000 rpm, p_s=0.414 MPa

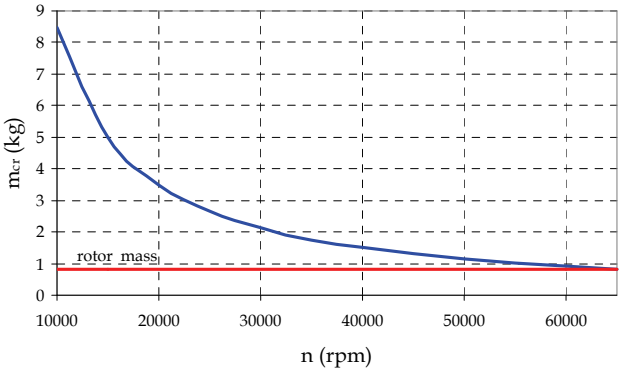


Fig. 28. Critical mass as function of running speed for $\varepsilon_w = 0.075$ and p_s=0.414 MPa

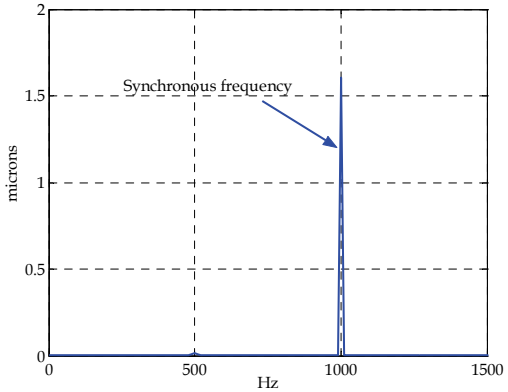


Fig. 29. FFT analysis and wave shape of the experimental signal for $\varepsilon_w = 0.075$, n=60000 rpm, p_s=0.414 MPa

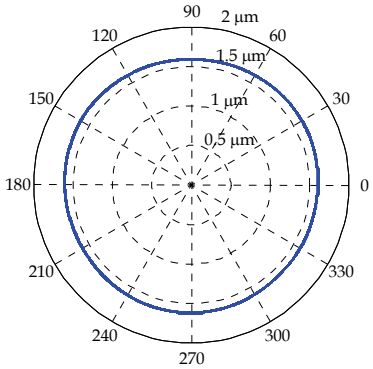


Fig. 30. Trajectory of the journal centre for $\varepsilon_w = 0.075$, n=60000 rpm, p_s=0.440 MPa, and 2 microns run-out

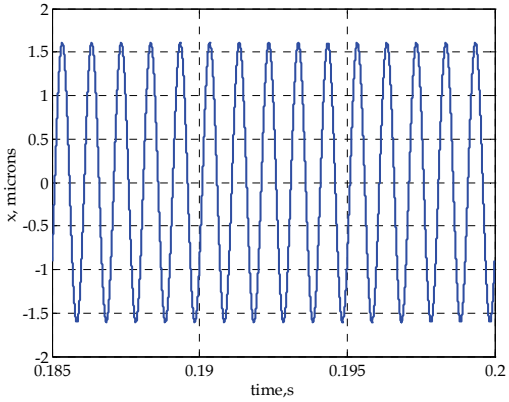


Fig. 31. The position of the shaft centre in the horizontal direction for $\varepsilon_w = 0.075$, n=60000 rpm, p_s=0.414 MPa , and 2 microns run-out

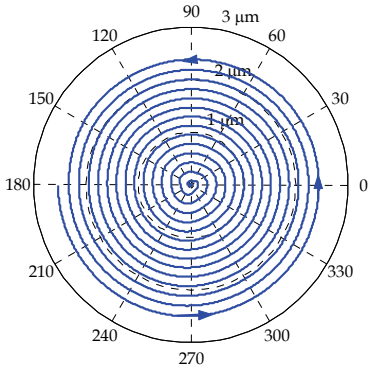


Fig. 32. Trajectory of the journal centre for $\varepsilon_w = 0.075$, n=60000 rpm, p_s=0.440 MPa, and 0 microns run-out

7. Conclusions

Both a critical mass and a transient method were developed to model and simulate the dynamic behavior of a fluid film wave journal bearing. The methods were validated by comparing the theoretical results obtained for a three-wave bearing having a diameter of 30 mm, a length of 27.5 mm and a clearance of 35 microns with experimental data. It was concluded that:

- The dynamic behavior of the bearing after the appearance of the sub-synchronous frequency could be numerically predicted only by using a transient approach.
- The experimental studies demonstrated that even when the bearing fluid film is unstable, the synchronous frequency is still present.
- The numerical simulations showed that the presence of the synchronous frequency in the unstable motions can be theoretically predicted only if the inherent unbalance of the rotor is taken into consideration.
- The theoretical and experimental investigations also proved that even if the fluid film is unstable, the wave bearing maintains the whirl orbit inside the bearing clearance.

8. References

Constantinescu, V. N., Nica, A., Pascovici, M. D., Ceptureanu, G. & Nedelcu S., (1985) *Sliding Bearings*, Allerton Press, ISBN 0-89864-011-3, New York

Dimofte, F. (1995). Wave journal bearing with compressible lubricant – Part I : The wave bearing concept and a comparison to the plain circular bearing, *Tribology Transactions*, Vol. 38, No. 1, pp. 153-160

Dimofte, F. (1995). Wave journal bearing with compressible lubricant – Part II : A comparison of the wave bearing with a groove bearing and a lobe bearing, *Tribology Transactions*, Vol. 38, No. 2, pp. 364-372

Dimofte, F. (1995). Wave journal bearing Part 1 : Analysis, NASA Report NASA-CR-19543-PT-1

Dimofte, F., Proctor, M.P., Fleming, D.P. & Keith, T. G. (2004). Experimental investigations on the influence of oil inlet pressure on the stability of wave journal bearing, *Proceedings of the 10th International Symposium on Transportation Phenomena and Dynamics of Rotating Machinery*, Honolulu, Hawaii, ISROMAC 10-2004-146

Ene, N. M., Dimofte, D. & Keith Jr., T. G. (2008). A dynamic analysis of hydrodynamic wave journal bearings, *Tribology Transactions*, Vol. 51, pp. 82-91

Frêne, J. & Constantinescu, V. N. (1975). Operating characteristics of journal bearings in the transition region, *Proceedings of the 2nd Leeds – Lyon Symposium on Tribology*, pp. 121-124

Frêne, J., Nicolas, D., Degueurce, B., Berthe, D. & Godet, M. (1997). Hydrodynamic lubrication : Bearings and thrust bearings, Elsevier, ISBN 0 444 82366 2, Amsterdam

Hashimoto, H. & Wada, S. (1990) Dynamic behavior of unbalanced rigid shaft supported on turbulent journal bearings – Theory and experiment, *ASME Journal of Tribology*, Vol. 112, pp. 404-408

Kakoty, S. K. & Majumdar, B. C., (1999). Effects of fluid inertia on stability of flexibly supported oil journal bearings: Linear perturbation analysis, *Tribology International*, Vol. 32, pp. 217-228

Kakoty, S. K. & Majumdar, B. C., (2000), Effects of fluid inertia on the dynamic coefficients and stability of journal bearings, *Proceedings IMechE, Part J*, Vol. 214, pp. 229 - 242

Kilt, P. & Lund J.W., (1986). Calculation of the dynamic coefficients of a journal bearing using a variational approach, *Journal of Tribology*, Vol. 108, pp. 421-425

Kirk, R. G. & Gunter, E. J. (1976). Short bearing analysis applied to rotordynamics I: Theory, *ASME Journal of Lubrication Technology*, Vol. 98, pp. 47-56

Kirk, R. G. & Gunter, E. J. (1976), Short bearing analysis applied to rotordynamics II: Results of journal bearing response, *ASME Journal of Lubrication Technology*, Vol. 98, pp. 319-329

Kostrzewsky G.J., Taylor D., Flack R.D. & Barrett L., (1998). Theoretical and experimental dynamic characteristics of highly preloaded three-lobe journal bearings, *Tribology Transactions*, Vol. 41, pp. 392-398

Lund, J. W. & Thomsen, K. K. (1978). A calculation method and data for the dynamic coefficients of oil–lubricated journal bearings, *Topics in Fluid Bearing and Rotor Bearing System Design and Optimization*, ASME, New York, pp. 1-28

Lund, J. W. (1987). Review of the concept of dynamic coefficients for fluid film journal bearings, *ASME Journal of Tribology*, Vol. 109, No. 37, pp. 37-41

Monmousseau, P. & Fillon, M. (1999). Frequency effects on the TEHD behavior of a tilting-pad journal bearing under dynamic loading, *ASME Journal of Tribology*, Vol. 121, pp. 321-326

Newkirk, B.L. & Taylor, H. D. (1925). Shaft whipping due to oil action in journal bearings, *General Electric Review*, Vol. 28, No. 8, pp. 985-988

Orcutt F.K. & Arwas E.B., (1967). The steady-state and dynamic characteristics of a full circular bearing and a partial arc bearing in the laminar and turbulent flow regimes, *Journal of Lubrication Technology*, pp. 143-153

Parkins D.W., (1979). Theoretical and experimental determination of the dynamic characteristics of a hydrodynamic journal bearing, *ASME Journal of Lubrication Technology*, Vol. 101, pp. 129-139

Rao T.V.V.L.N. & Sawicki J.T., (2002). Linear stability analysis for a hydrodynamic journal bearing considering cavitation effects, *STLE Tribology Transactions*, Vol. 45, No. 4, pp. 450-456

Rao T.V.V.L.N. & Sawicki J.T., (2004). Stability characteristics of herringbone grooved journal bearings incorporating cavitation effects, *ASME Journal of Tribology*, Vol. 126, pp. 281-287

San Andres, L. (1997). Transient response of externally pressurized fluid film bearings, *STLE Tribology Transactions.*, Vol. 40, No. 1, pp. 147-155

Tichy, J. & Bou-Said. B. (1991). Hydrodynamic lubrication and bearing behavior with impulsive loads, *STLE Tribology Transactions*, Vol. 34, No. 4, pp. 505-512

Tieu, A. K. & Qiu, Z. L. (1995). Stability of finite journal bearings from linear and nonlinear bearing forces, *STLE Tribology Transactions*, Vol. 38, No. 3, pp. 627-635

Vijayaraghavan, D. & Brewe & D. E., (1992). Frequency Effects on the Stability of a Journal Bearing for Periodic Loading, *ASME Journal of Tribology*, Vol. 114, pp. 107-115

White, M. F. & Chan, S. H., (1992). The sub-synchronous dynamic behavior of tilting-pad journal bearings, *Journal of Tribology*, Vol. 114, pp. 167-173

17

Workforce capacity planning using zero-one-integer programming

Said El-Quliti and Ibrahim Al-Darrab
King Abdulaziz University
Saudi Arabia

1. Introduction

Planning for human resource needs is one of the greatest challenges facing managers and leaders. In order to meet this challenge, a uniform process that provides a disciplined approach for matching human resources with the anticipated needs of an organization is essential. Workforce planning is a fundamental planning tool, critical to quality performance that will contribute to the achievement of program objectives by providing a basis for justifying budget allocation and workload staffing levels. As an organization develops strategies to support the achievement of performance goals in the strategic plans, workforce planning should be included as a key management activity.

An important problem in workforce planning arises when management needs to employ the right number of people to perform daily tasks such as in a reception department. Each employee is weekly required to work a certain number of days (not necessarily consecutive) and the number of required employees varies from day to day according to demand. In particular, if every employee is required to work five days a week and takes any two days off, then scheduling employees to meet daily requirements becomes a formidable task for management. Many practical applications of workforce planning can be cited where such problems arise such as: planning of the number of employees in receptions in hospitals, service stations, call centers, and hotels; planning the number of doctors and nurses in hospitals; planning the number of workers in restaurants, etc. Assigning tasks to employees is a difficult task. Errors committed in such assignments can have far-reaching consequences, such as reduced efficiency due to absenteeism, lack of job satisfaction, formal grievances, and generally deteriorating labor relations.

The first part of the paper is this introduction, the second part is the literature review, the third part describes the statement of the problem considering an organization where each employee works six days per week and takes one day off, or he works five days per week, and takes two days off (consecutive, or not necessarily consecutive). The number of employees needed on each day of the week differs according to the different workload on each day. It is needed to minimize the total number of workers (workforce capacity), while maintaining the performance of keeping the minimum required number of employees on each day of the week. The fourth part introduces the proposed mathematical model formulation for each case. The model comprises the objective function and the problem

constraints. The fifth part presents 15 real application examples, the examples are 12 hotels and 3 hospitals in Jeddah city, the data are taken for a period of 3 months. The six part presents the obtained results for the case studies, while the remaining parts are the conclusions and points for future research.

2. Literature Review

Workforce Planning (WFP) ensures that "the right people with the right skills are in the right place at the right time." This definition covers a methodical process that provides managers with a framework for making human resource decisions based on the organization's mission, strategic plan, budgetary resources, and a set of desired workforce competencies.
Various papers approach the problem from a spatial point of view. (Eiselt & Marianov, 2008) mapped the employees and the relevant tasks in a skill space, task assignments are determined, tasks are assigned to employees so as to minimize employee—task distances in order to avoid boredom, and minimize inequity between the individual employees' workloads, and minimize costs.
Constructing schedules is not also an easy task to accomplish in settings where work must be performed 24 hours per day and 7 days a week, such as in police and fire departments, or in emergency rooms of hospitals. (Knanth, 1996) studied the problem that one is faced when aiming to generate "good schedules" that satisfy many complicated rules, including ergonomic rules. Manpower scheduling in emergency rooms in hospitals is a famous and very critical problem in workforce capacity planning. (Carter & Lapierre, 2001) concluded that ergonomic constraints are very important in order to manage the circadian rhythm of the staff and it is critical to take them into account when building schedules. (Gendreau et al., 2007) discussed also this problem in five different hospitals of the Montreal, Canada area, the authors first propose generic forms for the constraints encountered in this context, then review several possible solution techniques that can be applied to physician scheduling problems, namely tabu search, column generation, mathematical programming and constraint programming, and examine their suitability for application depending on the specifics of the situation at hand.
(El-Quliti & Al-Darrab, 2009) address the problem of finding the optimal number of employees to be assigned each day of the week and determining the weekly schedule of each employee given that on each day at least a certain number of employees must be used to meet job or project requirements. The approach presented is to solve the problem in two stages. The first stage solves the problem with two consecutive off-days using a linear integer programming model. The second stage uses a zero-one integer programming model utilizing results of the first stage. Both mathematical formulation and solution to the problem are developed, and the LINDO computer package was used to solve an illustrative example. The optimum daily workforce size and schedule of every employee are thus obtained.
(El-Quliti & Al-Darrab, 2010) present some real world applications for the problem of finding the optimal number of employees to be assigned each day of the week, and determining the weekly schedule of each employee given that on each day at least a certain number of employees must be used to meet job requirements. 15 real case studies are presented in this research, 12 cases for receptions in hotels and 3 for emergency in hospitals

in Jeddah city. LINGO computer package has been used to solve these case studies. The optimum daily workforce size and schedule of every employee have been obtained.

3. Statement of the Problem

Consider an organization where each employee works five (or six) days per week, and takes one (or two) day(s) off. Suppose that the number of employees needed on each day of the week differs according to the different workload on each day, and suppose that the required numbers are as follows:

Day:	Mon	Tue	Wed	Thu	Fri	Sat	Sun
Number required	n_1	n_2	n_3	n_4	n_5	n_6	n_7
Number assigned	x^1	x^2	x^3	x^4	x^5	x^6	x^7

It is needed to minimize the total number of workers (workforce capacity), while maintaining the performance of keeping the minimum required number of employees on each day of the week, the questions will be:

1. How many total employees should be assigned?
2. How many employees should work on each day?
3. What is the week schedule for each employee (working days)?

4. Mathematical Model Formulation

4.1. Six-working days per week

The mathematical model is formulated by considering the number of employees that start working on Monday as x_1, and start on Tuesday as x_2, …, and start on Sunday as x_7, Fig. 1.
The objective function is clearly:

$$\text{Min. } z_1 = x_1 + x_2 + x_3 + x_4 + x_5 + x_6 + x_7$$

The Problem Constraints:
For the number of employees working on Monday:

$$x^1 = x1 + x_3 + x_4 + x_5 + x_6 + x_7 \geq n_1$$

Similar constraints can be constructed for the other six days.
The complete Integer Program will have the following form:

$$\text{Min } z = x_1 + x_2 + x_3 + x_4 + x_5 + x_6 + x_7 \tag{1}$$

Subject to:
1) The Workload Constraints:

$$\begin{aligned}
x^1 &= x_1 + x_3 + x_4 + x_5 + x_6 + x_7 \geq n_1 \\
x^2 &= x_1 + x_2 + x_4 + x_5 + x_6 + x_7 \geq n_2 \\
x^3 &= x_1 + x_2 + x_3 + x_5 + x_6 + x_7 \geq n_3 \\
x^4 &= x_1 + x_2 + x_3 + x_4 + x_5 + x_7 \geq n_4 \\
x^5 &= x_1 + x_2 + x_3 + x_4 + x_5 + x_7 \geq n_5 \\
x^6 &= x_1 + x_2 + x_3 + x_4 + x_5 + x_6 \geq n_6 \\
x^7 &= x_2 + x_3 + x_4 + x_5 + x_6 + x_7 \geq n_7
\end{aligned} \tag{2}$$

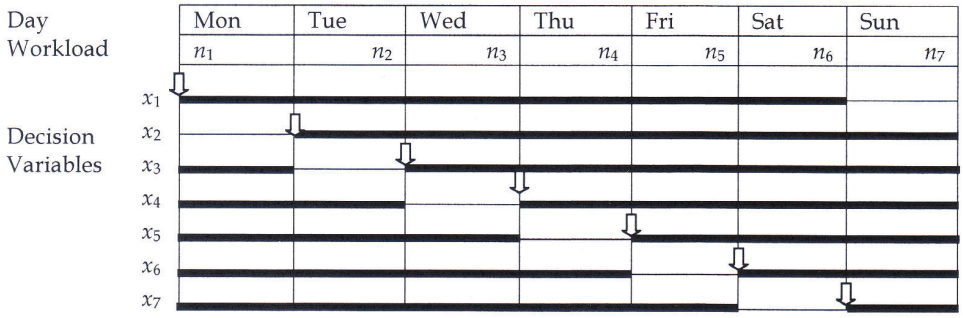


Fig. 1. Schematic diagram for 6-working days per week

2) The Minimum Workforce Capacity Constraints
Suppose that the minimum load including administrative and other related loads should not be less than a certain number $n_{\min}$, in such a case, the workload constraints will have the following modified form:

$$x^i \geq \max\{n_i, n_{\min}\}, \; i = 1, 2, \ldots, 7 \tag{3}$$

3) Non-negativity and Integrality Constraints

$$x_1, x_2, x_3, x_4, x_5, x_6, \text{ and } x_7 \geq 0, \text{ integers} \tag{4}$$

The obtained optimum solution will state the optimum number of employees planned for each day of the week. The minimum required number of employees for the organization is then:

$$z^* = x_1^* + x_2^* + x_3^* + x_4^* + x_5^* + x_6^* + x_7^*$$

4.2. Five-consecutive working days per week

The mathematical model is formulated in a similar manner like in the case of 6-working days per week, see Fig. 2.
The objective function is clearly:

$$\text{Min. } z = x_1 + x_2 + x_3 + x_4 + x_5 + x_6 + x_7$$

The Problem Constraints:
For the number of employees working on Monday:

$$x_1 + x_3 + x_4 + x_5 + x_6 + x_7 \geq n_1$$

Similar constraints can be constructed for the other six days.

The complete Integer Program will have the following form:

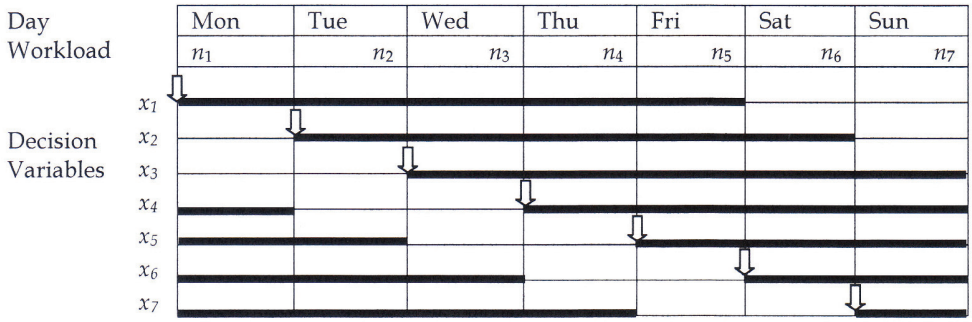


Fig. 2. Schematic diagram for 5-consecutive working days per week

$$\text{Min } z = x_1 + x_2 + x_3 + x_4 + x_5 + x_6 + x_7 \tag{5}$$

Subject to:
1) The Workload Constraints:

$$\begin{aligned} x^1 &= x_1 + x_4 + x_5 + x_6 + x_7 \geq n_1 \\ x^2 &= x_1 + x_2 + x_5 + x_6 + x_7 \geq n_2 \\ x^3 &= x_1 + x_2 + x_3 + x_6 + x_7 \geq n_3 \\ x^4 &= x_1 + x_2 + x_3 + x_4 + x_7 \geq n_4 \\ x^5 &= x_1 + x_2 + x_3 + x_4 + x_5 \geq n_5 \\ x^6 &= x_2 + x_3 + x_4 + x_5 + x_6 \geq n_6 \\ x^7 &= x_3 + x_4 + x_5 + x_6 + x_7 \geq n_7 \end{aligned} \tag{6}$$

2) The Minimum Workforce Capacity Constraints
Suppose that the minimum load including administrative and other related loads should not be less than a certain number $n_{\min}$, in such a case, the workload constraints will have the following modified form:

$$x^i \geq \max\{n_i, n_{\min}\}, \; i = 1, 2, \ldots, 7 \tag{7}$$

3) Non-negativity and Integrality Constraints

$$x_1, x_2, x_3, x_4, x_5, x_6, \text{ and } x_7 \geq 0, \text{ integers} \tag{8}$$

The obtained optimum solution will state the optimum number of employees planned for each day of the week. The minimum required number of employees for the organization is then:

$$z^* = x_1^* + x_2^* + x_3^* + x_4^* + x_5^* + x_6^* + x_7^*$$

4.3. Five-working days-not necessarily consecutive

The approach presented here is to solve the problem in two stages. The first stage solves the problem with two consecutive off-days using a linear integer programming model. As the obtained solution from the first stage constitutes a feasible solution to the original problem, the second stage uses a zero-one integer programming model utilizing results obtained from the first stage.

4.3.1 Stage I of the Algorithm

Consider the situation where each employee works five consecutive days discussed before in section 4.2. The mathematical model is formulated by considering the number of employees that start working on Monday as x_1, and start on Tuesday as x_2, …, and start on Sunday as x_7, see Fig. 2.

The complete mathematical model was formulated in equations (5-8)

The obtained optimum solution will state the optimum number of employees planned for each day of the week. The minimum required number of employees for the organization is then:

$$z_1^* = x_1^* + x_2^* + x_3^* + x_4^* + x_5^* + x_6^* + x_7^*$$

To simplify notation in the following discussion, this minimum value will, henceforth, be denoted by z_1.

4.3.2 Stage II of the Algorithm

In this stage we will consider the case where each employee works five days per week (not necessarily consecutive), and takes two days off (any two days in the week).

As the optimum solution obtained from stage I is a feasible solution to stage II, then the number of employees z_1 obtained from the first stage I can be considered as an upper bound for the total number of employees required for stage II. In this stage, we will begin with z_1 employees in the mathematical model, and then we will delete those employees who will be idle in the final solution.

Decision Variables:

Let x_i^j denote binary decision variables, where:

i = The day number such that:
1 = Monday, 2 = Tuesday, ..., and 7 = Sunday,

j = The ID Number for an employee, j = 1, 2, …, z_1, where z_1 is the minimum number of employees obtained from Stage I:

Functions with N Possible Values

(Hillier & Lieberman, 2005) considered the situation where a given function is required to take on any one of N given values. Denote this requirement by:

$$f(x_1, x_2, \ldots, x_n) = d_1, d_2, \ldots \text{ or } d_N.$$

The equivalent Binary Programming formulation of this requirement is the following:

$$f(x_1, x_2, \ldots, x_n) = \sum_{i=1}^{N} d_i y_i ,$$

$$\sum_{i=1}^{N} y_i = 1, \text{ and}$$

y_i is a binary variable, for $i = 1, 2, \ldots, N$.

This new set of constraints would replace the N possible values requirement in the statement of the overall problem. This set of constraints provides an equivalent formulation because exactly one y_i must equal 1 and the others must equal 0, so exactly one d_i is being chosen as the value of the function. In this case, there are N yes or no questions being asked, namely, should d_i be the value chosen ($i = 1, 2, \ldots, N$)? Because the y_i's respectively represent these yes-or-no decisions, the second constraint makes them mutually exclusive alternatives.

Functions with Zero – Integer Values

Consider the special case where N given functions are required to take on any one of only 2 given values, one of which is zero. Denote this requirement by:

$$f_i(x_1, x_2, \ldots, x_n) = d_i \text{ or } 0 \quad \text{for } i = 1, 2, \ldots, N.$$

The equivalent Binary Programming formulation of this requirement is the following:

$$f_i(x_1, x_2, \ldots, x_n) = d_i y_i \quad \text{for } i = 1, 2, \ldots, N, \text{ and}$$

y_i is a binary variable for $i = 1, 2, \ldots, N$.

These constraints would replace the 2 possible values requirement in the statement of the problem. It provides an equivalent formulation because exactly the respective auxiliary binary variable y_i must equal 0 or 1. In this case, there are 2N yes or no questions being asked, namely, for each one of the N functions: should d_i be the value chosen? And should 0 be the value chosen? Because the variables y_i represents these yes-or-no decisions, the binary constraints make them mutually exclusive alternatives so that each function $f_i(x_1, x_2, \ldots, x_n)$ will be equal to either d_i or 0.

Working Days for the Employees

Some of the considered z_1 employees may not be needed in the final optimal solution, while the others will be needed to satisfy the required working load. The needed employees will work 5 days per week, and the extra ones (if any) will not work at all. To model this situation, we will introduce z_1 auxiliary binary variables y_j, each of which corresponds to an employee j, and we will consider these constraints:

$$\sum_{i=1}^{7} x_i^j = 5y_j \quad , j = 1, 2, \ldots\ldots, z_1 \text{ ; and } y_j \text{ is a binary variable} \quad \text{for } j = 1, 2, \ldots, z_1.$$

For any y_j = 1, j = 1, 2, …, z_1; the corresponding employee j will work 5 days, and for any y_j=0, j = 1, 2, …, z_1; the corresponding employee j will not work and he/she is not needed. So, the total number of needed employees, z_2, will be equal to:

$$z_2 = \sum_{j=1}^{z_1} y_j$$

The objective Function
As it is required to minimize the total number of needed employees, then the objective function will take the form:

$$Min \ z_2 = \sum_{j=1}^{z_1} y_j \tag{9}$$

Problem Constraints
1) The Workload Constraints
For the number of employees working on Monday:

$$\sum_{j=1}^{z_1} x_1^j \geq n_1$$

Similar constraints can be formulated for other days, so we will have:

$$\sum_{j=1}^{z_1} x_i^j \geq n_i \ , i = 1, 2, \ldots, 7 \tag{10}$$

2) The Minimum Workforce Capacity Constraints
Suppose that the minimum load including administrative and other related loads should not be less than a certain number $n_{\min}$, in such a case, the workload constraints will have the following modified form:

$$\sum_{j=1}^{z_1} x_i^j \geq \max\{n_i, n_{\min}\}, i = 1, 2, \ldots, 7 \tag{11}$$

3) Working Days Constraints
Each employee will work 5 days or he will not work at all:

$$\sum_{i=1}^{7} x_i^j = 5y_j \ , j = 1, 2, \ldots\ldots, z_1 \tag{12}$$

4) Binary Constraints
All the decision variables and the auxiliary variables are binary ones, so we have:

$$x_i^j \text{ , } i = 1, 2, \ldots, 7, \text{ and } j = 1, 2, \ldots, z_1, \text{ and}$$

$$y_j, \; j = 1, 2, \ldots, z_1 \text{ are binary variables} \tag{13}$$

The Working Schedule for Employees
The working schedule for each employee j will be known from the optimum solution of the model. When $x_i^j = 1$, then employee j will work for day i, (i = 1, 2, …., or 7), and for $x_i^j = 0$, employee j will not work for day i, (i = 1, 2, …., or 7).

4.4. Five-working days-not necessarily consecutive (direct approach)

The mathematical model is formulated in a similar way as in the case of 5-consecutive working days per week, but for all possible combinations of the two days off, see Fig. 3.

For the number of employees working on Monday:

$$x^1 = x_7 + x_8 + x_9 + x_{10} + x_{11} + x_{12} + x_{13} + x_{14} + x_{15} + x_{16} + x_{17} + x_{18} + x_{19} + x_{20} + x_{21} \geq n_1$$

Similar constraints can be constructed for the other six days.
The complete Integer Program will have the following form:

$$\text{Min } z = x_1 + x_2 + x_3 + x_4 + x_5 + x_6 + x_7 \tag{14}$$

Subject to:
1) The Workload Constraints:

$$\begin{aligned}
x^1 &= x_7+x_8+ x_9 + x_{10} + x_{11} + x_{12} + x_{13} + x_{14} + x_{15} + x_{16} + x_{17} + x_{18} + x_{19} + x_{20} + x_{21} \geq n_1 \\
x^2 &= x_2+x_3+ x_4 + x_5 + x_6 + x_{12} + x_{13} + x_{14} + x_{15} + x_{16} + x_{17} + x_{18} + x_{19} + x_{20} + x_{21} \geq n_2 \\
x^3 &= x_1+x_3+ x_4 + x_5 + x_6 + x_8 + x_9 + x_{10} + x_{11} + x_{16} + x_{17} + x_{18} + x_{19} + x_{20} + x_{21} \geq n_3 \\
x^4 &= x_1+x_2+ x_4 + x_5 + x_6 + x_7 + x_9 + x_{10} + x_{11} + x_{13} + x_{14} + x_{15} + x_{19} + x_{20} + x_{21} \geq n_4 \\
x^5 &= x_1+x_2+ x_3 + x_5 + x_6 + x_7 + x_8 + x_{10} + x_{11} + x_{12} + x_{14} + x_{15} + x_{17} + x_{18} + x_{21} \geq n_5 \\
x^6 &= x_1+x_2+ x_3 + x_4 + x_6 + x_7 + x_8 + x_9 + x_{11} + x_{12} + x_{13} + x_{15} + x_{16} + x_{18} + x_{20} \geq n_6 \\
x^7 &= x_1+x_2+ x_3 + x_4 + x_5 + x_7 + x_8 + x_9 + x_{10} + x_{12} + x_{13} + x_{14} + x_{16} + x_{17} + x_{19} \geq n_7
\end{aligned} \tag{15}$$

$$x^i \geq \max\{n_i, n_{\min}\}, \; i = 1, 2, \ldots, 7 \tag{16}$$

$$x_1, x_2, \ldots\ldots\ldots\ldots, \text{ and } x_{21} \geq 0, \text{ integers} \tag{17}$$

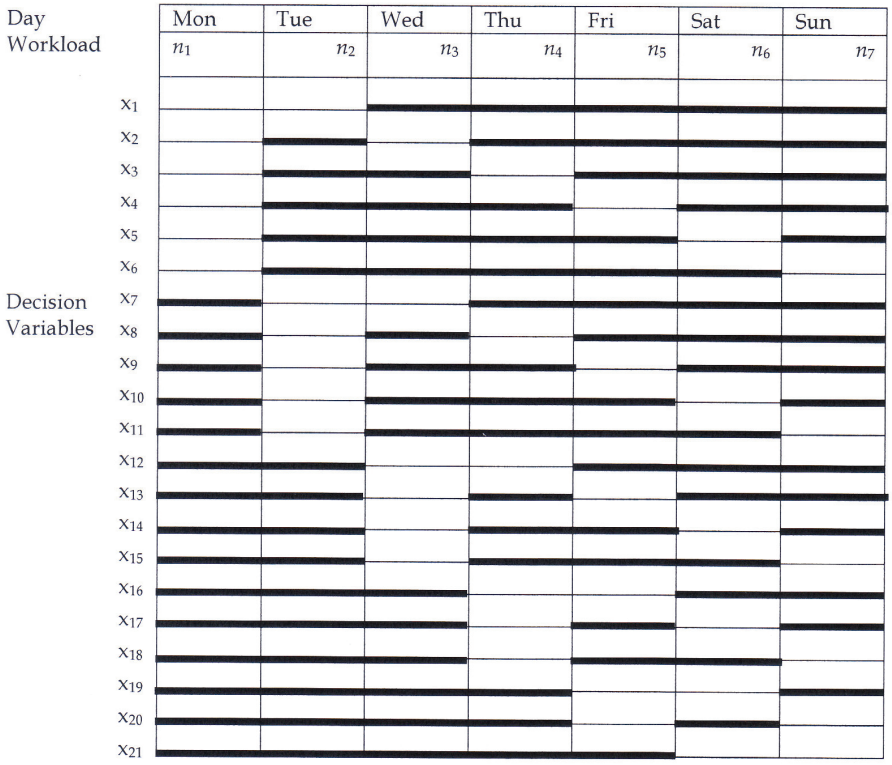


Fig. 3 Schematic diagram for 5-working days-not necessarily consecutive

5. Real Application Examples

Saturday	Sunday	Monday	Tuesday	Wednesday	Thursday	Friday
194	159	176	156	193	196	213
178	140	189	174	194	173	231
183	166	178	142	199	184	183
196	162	183	168	200	194	215
197	144	203	179	211	202	254
212	182	243	187	220	228	242
217	169	209	194	222	223	237
224	189	207	198	229	204	245
215	173	225	192	223	178	182
161	138	158	153	162	181	184
175	144	173	142	166	172	211
171	146	174	152	180	178	196
162	147	162	145	173	180	183
183						

Table 1. Number of guest in Jeddah Radisson Blu Hotel

15 real case studies are presented in this research, 12 cases for receptions in hotels and 3 for emergency in hospitals in Jeddah city, Kingdom of Saudi Arabia. Collecting data is the first and important step, the data here is the number of guests in the reception of a hotel, or the number of patients in the emergency section of a hospital. Data are collected for the hotels and hospitals during a period of 3 months. The number of guests or patients (working load) is different for different days of the week. For example, the number of guests for Jeddah Radisson Blu Hotel in the prescribed period is shown in Table 1.
The average number of guests in Radisson Blu Hotel Jeddah is shown in Figure 4.

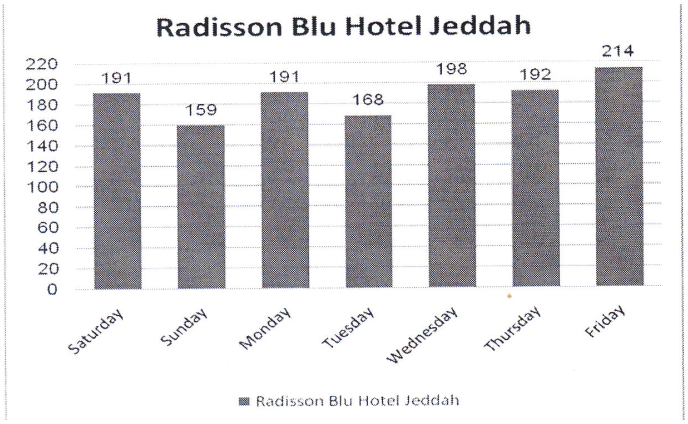


Fig. 4. Average number of guests in Radisson Blu Hotel, Jeddah

As an example of the data for hospitals, the number of patients in King Abdulaziz University Hospital in the prescribed period is shown in Table 2.
All the average numbers of guests in hotels or patients in hospitals are shown in Table 3 and in Figure 5.

Saturday	Sunday	Monday	Tuesday	Wednesday	Thursday	Friday
108	80	86	75	67	38	34
99	61	59	73	76	46	33
84	68	63	75	69	48	34
57	79	103	70	64	37	37
89	51	92	74	69	62	41
99	75	83	72	71	58	38
88	79	96	84	62	53	59
90	64	84	86	74	30	42
67	54	40	57	35	12	19
63	56	26	36	51	12	2
37	43	44	58	33	8	6
55	42	49	39	40	17	4
42	35	50	43		11	7
42						5

Table 2. Number of patients in KAU Hospital

Name	Sat	Sun	Mon	Tue	Wed	Thu	Fri
Radison Blu Hotel	191	159	191	168	198	192	214
Le Meridien Hotel	197	256	251	287	233	229	250
Sunset Hotel Hotel	81	114	140	83	113	83	111
Al-Hamra Hotel	110	85	112	119	129	113	100
Golden Tulip Hotel	50	51	47	48	55	39	44
Moevenpick Hotel	180	161	197	146	194	183	202
Trident Hotel Hotel	135	119	124	107	112	123	115
Hilton Hotel Hotel	293	232	351	327	342	352	321
Crowne Plaza Hotel	225	213	232	235	248	243	236
Ramada Hotel	131	114	105	114	95	92	105
Marriott Hotel	199	189	166	170	196	181	190
Al-Salam Hospital	215	219	191	198	191	206	198
King Abdulaziz Univ. Hospital	73	59	62	60	51	31	26
Al-Rafeea Hospital	41	31	37	34	28	18	11
Al-Salam Hospital	33	29	31	32	32	16	19

Table 3. Average Number of Guests and Patients

In all the case studies, each reception employee in a hotel or a nurse in a hospital is working 6 days per week. Each hotel employee can serve about 40 guests in his shift, and each nurse in a hospital can serve 5 patients. The mathematical model is formulated for each case according to equations 1-4 stated before.

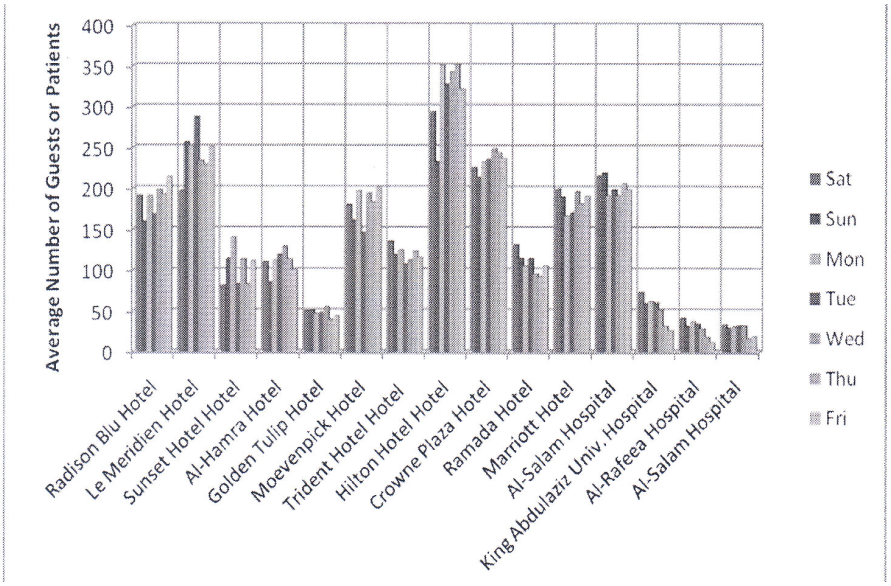


Fig. 5. Average Number of Guests and Patients

For example, for Radisson Blu Hotel Jeddah, Table 4 shows the required number of employees in each day of the week calculated according to the working load (average number of guests) and the number of guests that one employee can serve per one shift (40).

Day	Sat	Sun	Mon	Tue	Wed	Thu	Fri
Required Number of Employees	5	4	5	5	5	5	6

Table 4. Required Number of Employees for Radisson Blu Hotel Jeddah

The mathematical model for this case is as follows:

$$\text{Min } Z = x_1 + x_2 + x_3 + x_4 + x_5 + x_6 + x_7$$

Subject to:

$$x_1 + x_3 + x_4 + x_5 + x_6 + x_7 \geq 5,$$
$$x_1 + x_2 + x_4 + x_5 + x_6 + x_7 \geq 4,$$
$$x_1 + x_2 + x_3 + x_5 + x_6 + x_7 \geq 5,$$
$$x_1 + x_2 + x_3 + x_4 + x_6 + x_7 \geq 5,$$
$$x_1 + x_2 + x_3 + x_4 + x_5 + x_7 \geq 5,$$
$$x_1 + x_2 + x_3 + x_4 + x_5 + x_6 \geq 5,$$
$$x_2 + x_3 + x_4 + x_5 + x_6 + x_7 \geq 6,$$
$$x_1 , x_2 , x_3 , x_4 , x_5 , x_6 , x_7 \geq 0, \text{ integers.}$$

Similar mathematical models are formulated for the other case studies in the same manner according to the data for each case.

6. Results for the Application Examples

LINGO software is used to solve the obtained mathematical models for the case studies. An example of the obtained results are diagrammatically shown in Figure 6 for Radisson Blu Hotel Jeddah.

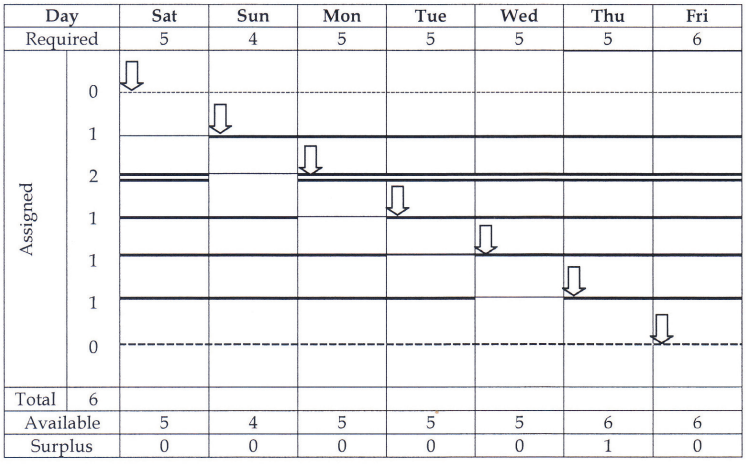


Fig. 6. Optimum solution for Radisson Blu Hotel Jeddah

The solution for all hotels and hospitals cases (needed number of employees) and the surplus (difference between assigned and the required number of employees) are shown in Table 5.

No.	Place Name	Needed Employees	Surplus (Days)	Load/employee
1	Radison Blu Hotel	6	0(6), 1(1)	40 guests
2	Le Meridien Hotel	8	0(6), 1(1)	
3	Sunset Hotel	4	0(2), 1(5)	
4	Sofitel Jeddah Al Hamra	4	0(6), 1(1)	
5	Golden Tulip Hotel	3	0(3), 1(4)	
6	Moevenpick Hotel	6	0(7)	
7	Trident Hotel	4	0(7)	
8	Hilton Hotel	10	0(6), 1(1)	
9	Crowne Plaza Hotel	8	0(7)	
10	Ramada Continental Hotel	4	0(7)	
11	Marriott Hotel	6	0(6), 1(1)	
12	Holiday Inn Jeddah Al Salam	7	0(7)	
13	KAU Hospital	15	0(4), 2(1), 4(1), 8(1),	5 patients
14	Al-Rafeea Hospital	9	0(4), 2(1), 3(1), 5(1)	
15	Al-Salam Hospital	7	0-7	

Table 5. Solution for all hotels and hospitals

7. Conclusions

1. Many real world application examples in hotels, hospitals, call centers, and many other organizations address the problem of workforce planning. The problem facing the management is to find the minimum number of employees to be assigned each day of the week given that a certain number must be assigned to meet job requirements on each day, and each employee should work a certain number of days per week (5 or 6).
2. The problem constraints are the workload in different days of the week, the minimum workforce capacity that should exist each day, the number of working days per week (consecutive or not), and the integrality constraints. The objective is to minimize the total number of needed employees, and
3. A mathematical formulation for the problem is illustrated for three cases: 6 working days per week, 5 consecutive working days per week, and 5 working days but not necessarily consecutive. All pattern possibilities for working different days are investigated for each case, and the number of employees assigned accordingly is considered as the decision variables. Another approach is proposed to solve such a problem in two stages. The first stage solves the problem with two consecutive off-days using a linear integer programming model. The second stage uses a zero-one integer programming model utilizing results of the first stage as it represents a feasible solution for the second stage.
4. 15 real case studies are presented, 12 for reception in hotels, and 3 for emergency in hospitals. The employees in hotels and the nurses in hospitals are working 6 days per

week and they have one day off. Data are collected for each case for a time period of 3 months, the data represents the working load in each day of the week.

5. The obtained mathematical model is an Integer Linear Programming and the solution to the problem are developed, and the LINDO computer package is used to solve the case studies. The obtained solutions represent the required number of employees and the scheduling for each group.

8. Points for Future Research

1. To categorize the employees into two categories: senior and junior, where some of the tasks can be done only by seniors.
2. To enhance the mathematical model formulation by including supervisors' constraints whom should be distributed allover days of the week.
3. To perform sensitivity analysis for the problem.
4. To consider the case involving stochastic constraints.
5. To perform a complete decision support system in order to help decision makers finding the optimum solutions for such practical applications.

9. References

Carter, M. & Lapierre, S. (2001). Scheduling emergency room physicians. *Health Care, Management Science, Vol.* 4, (2001) 347–360.

Eiselt, H., & Marianov, V. (2008). Employee positioning and workload allocation. *Computers & Operations Research*, Vol. 35, (Issue 2, February 2008) 513-524.

El-Quliti, S. & Al-Darrab, I. (2009). A Zero-One Integer Programming model for the optimum workforce capacity planning with workload constraints. *Proceedings of The* 20th. *IASTED International Conference on Modelling and Simulation*,"*MS 2009*", Banff, Alberta, Canada, July 2009, IASTED.

El-Quliti, S. & Al-Darrab, I. (2010). Optimum Workforce Capacity Planning With Real World Applications Using Integer Programming. *Proceedings of The First International Symposium on Computing in Science & Engineering,* Kuşadası- İzmi, June 2010.

Gendreau, M., Ferland, J. , Gendron,B. , Hail ,N., Jaumard, B. , Lapierre, S., Pesant, G. & Soriano, P. (2007). *Physician Scheduling in Emergency Rooms,* Lecture Notes in Computer Science, Springer Berlin / Heidelberg.

Hillier, F. & Lieberman, G. (2005).Introduction to Operations Research, 8th edition, McGraw-Hill International Edition, Industrial & Plant Engineering Series, ISBN 0-07-232169-5, Singapore.

Knaunth, P. (1996). Design better shift systems. *Applied Ergonomics*, Vol. 27 (1996) 39–44.

LINDO Systems Corporation (2010). LINGO 11.0, *Optimization Modeling Software for Linear, Nonlinear, and Integer Programming*,

http://www.lindo.com/index.php?option=com_content&view=article&id=2&Itemid=10, visited on 03 April (2010).

Taha, H. (2003). *Operations Research: An Introduction*, 8th. Edition, Pearson Education, Inc.,ISBN-10: 0-13-136014-0, New Jersey.