Haeng Kon Kim · Sio-Iong Ao
Mahyar A. Amouzegar *Editors*

Transactions on Engineering Technologies

Special Issue of the World Congress on Engineering and Computer Science 2013

Springer

Transactions on Engineering Technologies

Haeng Kon Kim · Sio-Iong Ao
Mahyar A. Amouzegar
Editors

Transactions on Engineering Technologies

Special Issue of the World Congress on Engineering and Computer Science 2013

Springer

Editors
Haeng Kon Kim
School of Information Technology
Catholic University of DaeGu
DaeGu
Korea, Republic of (South Korea)

Mahyar A. Amouzegar
College of Engineering
California State Polytechnic University
Pomona, CA
USA

Sio-Iong Ao
IAENG Secretariat
International Association of Engineers
Hong Kong
Hong Kong SAR

ISBN 978-94-017-9114-4 ISBN 978-94-017-9115-1 (eBook)
DOI 10.1007/978-94-017-9115-1
Springer Dordrecht Heidelberg New York London

Library of Congress Control Number: 2013953195

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

A large international conference on Advances in Engineering Technologies and Physical Science was held in San Francisco, California, USA, in October 23–25, 2013, under the World Congress on Engineering and Computer Science (WCES 2013). The WCECS 2013 was organized by the International Association of Engineers (IAENG). IAENG is a non-profit international association for the engineers and the computer scientists, which was founded originally in 1968 and has been undergoing rapid expansions in the recent years. The WCECS Congress serves as an excellent platform for the engineering community to meet with each other and to exchange ideas. The Congress has also struck a balance between theoretical and application development. The conference committees have been formed with over 200 members who are mainly research center heads, deans, department heads/chairs, professors, and research scientists from over 30 countries. The full committee list is available at the congress' website: www.iaeng.org/WCECS2013/committee.html. The Congress is truly an international meeting with a high level of participation from many countries. The response to the conference call for papers was excellent with more than six hundred manuscript submissions for the WCECS 2013. All submitted papers went through the peer review process and the overall acceptance rate was 50.8 %.

This volume contains 56 revised and extended research articles, written by prominent researchers participating in the congress. Topics covered include electrical engineering, chemical engineering, circuits, computer science, communications systems, engineering mathematics, systems engineering, manufacture engineering, and industrial applications. This book offers the state of art of tremendous advances in engineering technologies and physical science and applications, and also serves as an excellent source of reference for researchers and graduate students working with/on engineering technologies and physical science and applications.

Haeng Kon Kim
Sio-Iong Ao
Mahyar A. Amouzegar

Contents

Chapter 1
Tidal Current Turbine, Wind Turbine and Their Dynamic Modelling

Hamed H. H. Aly and M. E. El-Hawary

Abstract Wind and tidal current are of the most common energy resources for generating electricity in the near future because of the oil problems (crises and pollution). Within that frame, wind and tidal current energies are surging to the fore. The dynamic model of the offshore wind and tidal current is very important topic for dealing with these renewable energies. This chapter describes the overall dynamic models of wind and tidal current turbine using three different types of generators (doubly fed induction generator (DFIG), squirrel cage induction generator (SCIG) and direct drive permanent magnet synchronous generator (DDPMSG)). The state space for all types of the generators is concluded.

Keywords Controller · Direct drive permanent magnet synchronous generator (DDPMSG) · Doubly fed induction generator (DFIG) · Modelling · Offshore wind · Tidal current

1 Introduction

In recent years, conventional non-renewable electrical energy production has become an increasing concern due to its high costs, limited resources, and negative influence on global warming from CO2 emissions. In response to these challenges,

H. H. H. Aly (✉)
Ivan Curry School of Engineering, Acadia University, Wolfville, NS B4P 2R6, Canada
e-mail: hamed.aly@dal.ca; hamed.aly@acadiau.ca

M. E. El-Hawary
Department of Electrical and Computer Engineering, Dalhousie University, Halifax
NS B3H 4R2, Canada
e-mail: elhawary@dal.ca

H. K. Kim et al. (eds.), *Transactions on Engineering Technologies*,
DOI: 10.1007/978-94-017-9115-1_1,

scientists have begun to focus their research on renewable energy sources. Renewable energy is generally a clean source of supplying electrical loads, especially in remote and rural areas. Wind energy is one of the most common and rapidly growing renewable energy sources. Wind energy is produced from air motion caused by the uneven heating of the earth's surface by the sun. While wind turbines are associated with negative issues such as noise, visual impacts, erosion, birds and bats being killed, and radio interference, it is still an extremely useful form of energy for rural areas where access to utility transmission facilities is limited. Moreover, the use of wind energy reduces greenhouse gas emissions and positively impacts climate change due to fossil fuel replacement. Worldwide wind capacity is growing fast and may reach up to 1 million MW by 2050. This means that wind energy integration will become an important factor in the stability of the electric grid. Thus, there is a need for a 'smart' grid that is able to work through any disturbance and supply high quality electric energy to consumers. To date, however, wind power as an energy source is intermittent, challenging to predict, and requires using some form of storage to integrate it into the electric grid. New control techniques and improved forecasting methods are helpful in establishing operating practices that will increase the reliability of wind energy supply [1, 2].

Tidal currents are fluctuating, intermittent but a predictable source of energy compared to wind. Its use is very effective as it relies on the same technologies used in wind turbines. The electrical-side layout and modeling approaches used in tidal in-stream systems are similar to those used for wind and offshore wind systems. The speed of water currents is lower than wind speed, while the water density is higher than the air density and as a result wind turbines operate at higher rotational speeds and lower torque than tidal in-stream turbines which operate at lower rotational speed and high torque [1–5].

The easier predictability of the tidal in-stream energy resource makes it easier to integrate in an electric power grid. Recognizing that future ocean energy resources are available far from load centers and in areas with limited grid capacity will result in challenges and technical limitations. With the growing penetration of tidal current energy into the electric power grid system, it is very important to study the impact of tidal current turbines on the stability of the power system grid and to do that we should model the overall system. The model of the ocean energy system consists of three stages. The first stage contains the fluid mechanical process. The second stage consists of the mechanical conversion and depends on the relative motion between bodies. This motion may be mechanical transmission and then using mechanical gears or may be depending on the hydraulic pumps and hydraulic motors. The third stage consists of the electromechanical conversion to the electrical grid [4–9].

2 Wind and Tidal Current Model Using IG

2.1 The Speed Signal Resource Model for the Tidal Current and Offshore Wind

The tidal current speed may be expressed as a function of the spring tide speed, neap tide speed and tides coefficient. Paper (10) proposed different algorithm models for the tidal current speed. The Wind Speed Signals Model (v_w) consists of four components related to the mean wind speed (v_{mw}), the wind speed ramp (v_{rw}) (which is considered as the steady increase in the mean wind speed), the wind speed gust (v_{gw}), and the turbulence (v_{tw}).

$$v_w = v_{mw} + v_{gw} + v_{rw} + v_{tw}$$

The mean wind speed is a constant; a simple ramp function will be used for ramp component (characterized by the amplitude of the wind speed ramp ($\mathrm{A_r}$ (m/s)), the starting time ($\mathrm{T_{sr}}$), and the ending time ($\mathrm{T_{er}}$)). The wind speed gust component is characterized by the amplitude of the wind speed gust ($\mathrm{A_g}$ (m/s)), the starting time ($\mathrm{T_{sg}}$), and the ending time ($\mathrm{T_{eg}}$). The wind speed gust may be expressed as a sinusoidal function. The most used models are given by:

$$\begin{aligned} v_{gw} &= \mathrm{A_g}(1 - \cos(2\Pi(\mathrm{t/D_g - T_{sg}/D_g}))) && \mathrm{T_{sg} \le t \le T_{eg}} \\ v_{gw} &= 0 && \mathrm{t < T_{sg}\ or\ t > T_{eg}} \\ \mathrm{D_g} &= \mathrm{T_{eg} - T_{sg}} \end{aligned}$$

A triangular wave is used to represent the turbulence function which has adjustable frequency and amplitude [1, 10].

2.2 The Rotor Model

For the wind, the rotor model represents the conversion of kinetic energy to mechanical energy. The wind turbine is characterized by C_p (wind power coefficient), λ (tip speed ratio), and β (pitch angle). $\lambda = \omega_t R/v_{w,}$ where R is the blade length in m, v_w is the wind speed in m/s, and ω_t is the wind turbine rotational speed in rad/sec. C_P-λ-β curves are manufacturer-dependent but there is an approximate relation expressed as:

$$\mathrm{C_p} = 1/2[(\mathrm{RC_r})/\lambda - 0.026\beta - 2]\mathrm{e}^{-0.295(\mathrm{RCf})/\lambda}$$

C_f is the wind turbine blade design constant. The rotor model may be represented by using the equation of the power extracted from the wind ($P_w = 0.5\rho \Pi R^2 C_p v_w^3$) [1, 10].

For the tidal current turbines the rotor model may be represented by using the equation of the power extracted from the wind ($P_w = 0.5\rho \Pi R^2 C_p v_w^3$), The power ($P_{ts}$) may be found using: $P_{ts} = ½\ \rho A(V_{tide})^3$. The turbine harnesses a fraction of this power, hence the power output may be expressed as: $P_t = ½\ \rho C_p A(V_{tide})^3$. The power output is proportional to the cube of the velocity. The velocity at the bottom of the channel is lower than at the water column above seabed. The mechanical torque applied to the turbine (T_m) can be expressed as [4–9]:

$$\mathrm{Tm} = \left(0.5\pi\rho R^2 C_p V_{tide}^3\right)/\omega_t \tag{1}$$

The shaft system for tidal current and offshore turbines may be represented by a two mass system one for the turbine and the other for the generator as shown:

$$2H_t \frac{d\omega_t}{dt} = T_t - K_s(\Theta_r - \Theta_t) - D_s(\omega_r - \omega_t) \tag{2}$$

$$2H_g \frac{d\omega_r}{dt} = T_e - K_s(\Theta_r - \Theta_t) - D_s(\omega_r - \omega_t) \tag{3}$$

$$\Theta_{tr} = \Theta_r - \Theta_t \tag{4}$$

$$\frac{d\theta \mathrm{tr}}{dt} = \omega_r - \omega_t \tag{5}$$

There is a ratio for the torsion angles, damping and stiffness that need to be considered when one adds a gear box as all above calculations must be referred to the generator side. The same model used for the offshore wind is used for tidal in-stream turbines; however, there is a number of differences in the design and operation of marine turbines due to the changes in force loadings, immersion depth, and different stall characteristics.

2.3 Dynamic Model of DFIG

The DFIG model is developed using a synchronously rotating d-q reference frame with the direct-axis oriented along the stator flux position. The reference frame rotates at the same speed as the stator voltage. The stator and rotor active and reactive power are given by [4–9, 11]:

$$P_s = 3/2\left(v_{ds} i_{ds} + v_{qs} i_{qs}\right), \quad P_r = 3/2\left(v_{dr} i_{dr} + v_{qr} i_{qr}\right) \tag{6}$$

$$\mathrm{P_g = P_s + P_r} \tag{7}$$

$$\mathrm{Q_s = 3/2\left(v_{qs}i_{ds} - v_{ds}i_{qs}\right), \quad Q_r = 3/2\left(v_{qr}i_{dr} - v_{dr}i_{qr}\right)} \tag{8}$$

The model of the DFIG can be described as:

$$\mathrm{v_{ds}} = -\mathrm{R_s i_{ds}} - \omega_\mathrm{s}\psi_\mathrm{qs} + \frac{\mathrm{d}}{\mathrm{dt}}\Psi_\mathrm{ds} \tag{9}$$

$$\mathrm{v_{qs}} = -\mathrm{R_s i_{qs}} + \omega_\mathrm{s}\psi_\mathrm{ds} + \frac{\mathrm{d}}{\mathrm{dt}}\Psi_\mathrm{qs} \tag{10}$$

$$\mathrm{v_{dr}} = -\mathrm{R_r i_{dr}} - \mathrm{s}\omega_\mathrm{s}\psi_\mathrm{qr} + \frac{\mathrm{d}}{\mathrm{dt}}\Psi_\mathrm{dr} \tag{11}$$

$$\mathrm{v_{qr}} = -\mathrm{R_r i_{qr}} + \mathrm{s}\omega_\mathrm{s}\psi_\mathrm{dr} + \frac{\mathrm{d}}{\mathrm{dt}}\Psi_\mathrm{qr} \tag{12}$$

$$\Psi_{ds} = -L_{ss}\, i_{ds} - L_m\, i_{dr}, \quad \Psi_{qs} = -L_{ss}\, i_{qs} - L_m\, i_{qr} \tag{13}$$

$$\Psi_{dr} = -L_{rr}\, i_{dr} - L_m\, i_{ds}, \quad \Psi_{qr} = -L_{rr}\, i_{qr} - L_m\, i_{qs} \tag{14}$$

$$\mathrm{s} = (\omega_\mathrm{s} - \omega_\mathrm{r})/\omega_\mathrm{s} \tag{15}$$

$$\frac{d\omega_r}{dt} = -\omega_\mathrm{s}\frac{ds}{dt} \tag{16}$$

where, $L_{ss} = L_s + L_m$, $L_{rr} = L_r + L_m$, L_s, L_r and L_m are the stator leakage, rotor leakage and mutual inductances, respectively. The previous model may be reduced by neglecting stator transients and is described as follows:

$$\mathrm{v_{ds} = -R_s\, i_{ds} + X'\, i_{qs} + e_d} \tag{17}$$

$$\mathrm{v_{qs} = -R_s\, i_{qs} + X'\, i_{ds} + e_q} \tag{18}$$

$$\frac{de_d}{dt} = -\frac{1}{T_0}\left(e_d + (X - X')i_{qs}\right) + s\omega_s e_q - \omega_s\frac{L_m}{L_{rr}}v_{qr} \tag{19}$$

$$\frac{de_q}{dt} = -\frac{1}{T_0}\left(e_q - (X - X')i_{ds}\right) - s\omega_s e_d + \omega_s\frac{L_m}{L_{rr}}v_{dr} \tag{20}$$

The components of the voltage behind the transient are $e_d = -\frac{\omega_s L_m}{L_{rr}}\psi_{qr}$ and $e_q = \frac{\omega_s L_m}{L_{rr}}\psi_{dr}$. The stator reactance $X = \omega_s\, L_{ss} = X_s + X_m$, and the stator transient reactance $X' = \omega_s\left(L_{ss} - L_m^2/L_{rr}\right) = X_s + (X_r X_m)/(X_r + X_m)$. The transient open circuit time constant is $T_o = L_{rr}/R_r = (L_r + L_m)/R_r$, and the electrical torque is $T_e = \left(i_{ds}i_{qr} - i_{qs}i_{dr}\right)X_m/\omega_s$.

2.4 Dynamic Model of SCIG

In the SCIG, the rotor is short circuited and so, the stator voltages are the same as the DFIG. In the SCIG there are capacitors to provide the induction generator magnetizing current and for compensation.

2.5 The Pitch Controller Model

The pitch controller is used to adjust the tidal current turbine to achieve a high speed magnitude. This may be represented by a PI controller as shown in Fig. 1.

$$\mathrm{\beta = \left(K_{pt} + K_{it}/S\right)\omega_t} \tag{21}$$

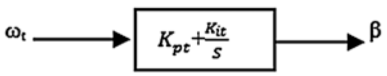
$$\mathrm{d\beta/dt = K_{pt} d\omega_t/dt + K_{it}\omega_t} \tag{22}$$

2.6 The Converter Model

A converter feeds or takes power from the rotor circuit and gives a variable speed (a partial scale power converter used). The rotor side of the DFIG is connected to the grid via a back to back converter. The converter at the side connected to the grid is called supply side converter (SSC) or grid side converter (GSC) while the converter connected to the rotor is the rotor side converter (RSC). The RSC operates in the stator flux reference frame. The direct axis component of the rotor current acts in the same way as the field current as in the synchronous generator and thus controls the reactive power change. The quadrature component of the rotor current is used to control the speed by controlling the torque and the active power change. Thus the RSC governs both the stator-side active and reactive powers independently. The GSC operates in the stator voltage reference frame. The *d*-axis current of the GSC controls the DC link voltage to a constant level, and the *q*-axis current is used for reactive power control. The GSC is used to supply or draw power from the grid according to the speed of the machine. If the speed is higher than synchronous speed it supplies power, otherwise it draws power from the grid but its main objective is to keep dc-link voltage constant regardless of the magnitude and direction of rotor power. The balanced power equation is given by [1, 12, 13]:

$$P_r = P_g + P_{DC} \tag{23}$$

Fig. 1 Pitch angle control block



$$P_{DC} = v_{DC}i_{DC} = -Cv_{DC}dv_{dc}/dt, \quad P_g = v_{Dg}i_{Dg} + v_{Qg}i_{Qg} \tag{24}$$

$$Cv_{DC}dv_{DC}/dt = v_{Dg}i_{Dg} + v_{Qg}i_{Qg-}v_{dr}i_{dr} - v_{qr}i_{qr} \tag{25}$$

2.7 Rotor Side Converter Controller Model

The rotor side converter controller used here is represented by four states ($\dot{x}_1, \dot{x}_2, \dot{x}_3$ and $\dot{x}_4$), $\dot{x}_1$ is related to the difference between the generated power of the stator and the reference power that is required at a certain time, $\dot{x}_2$ is related to the difference between the quadrature axis generator rotor current and the reference current that is required at a certain time, $\dot{x}_3$ is related to the difference between the stator terminal voltage and the reference voltage that is required at a certain time, and $\dot{x}_4$ is related to the difference between the direct axis generator rotor current and the reference current that is required at a certain time [12, 13]. Figure 2 shows the rotor side converter controller. This is described by Eqs. (26–34).

$$\dot{x}_1 = P_{ref} - P_s \tag{26}$$

$$\dot{x}_1 = -K_{i1}/K_{p1}x_1 + 1/\,K_{p1}i_{qr\_ref} \tag{27}$$

$$\dot{x}_2 = i_{qr\_ref} - i_{qr} \tag{28}$$

$$\dot{x}_2 = K_{p1}\dot{x}_1 + K_{i1}x_1 - i_{qr} \tag{29}$$

$$\begin{aligned}\dot{x}_2 = {} & -K_{i2}/K_{p2}x_2 + 1/K_{p2}v_{qr} - \omega_s L_m/K_{p2}i_{ds} - \omega_s L_{rr}/K_{p2}i_{dr} \\ & + (L_m/K_{p2})i_{ds}\omega_r + (L_{rr}/K_{p2})i_{dr}\omega_r\end{aligned} \tag{30}$$

$$\dot{x}_3 = v_{s\_ref} - v_s \tag{31}$$

$$\dot{x}_3 = -K_{i3}/K_{p3}x_3 + 1/K_{p3}i_{dr\_ref} \tag{32}$$

$$\dot{x}_4 = i_{dr\_ref} - i_{dr} \tag{33}$$

$$\begin{aligned}\dot{x}_4 = {} & -K_{i2}/K_{p2}x_4 + 1/K_{p2}v_{dr} - \omega_s L_m/K_{p2}i_{qs} - \omega_s L_{rr}/K_{p2}i_{qr} \\ & + (L_m/K_{p2})i_{qs}\omega_r + (L_{rr}/K_{p2})i_{qr}\omega_r\end{aligned} \tag{34}$$

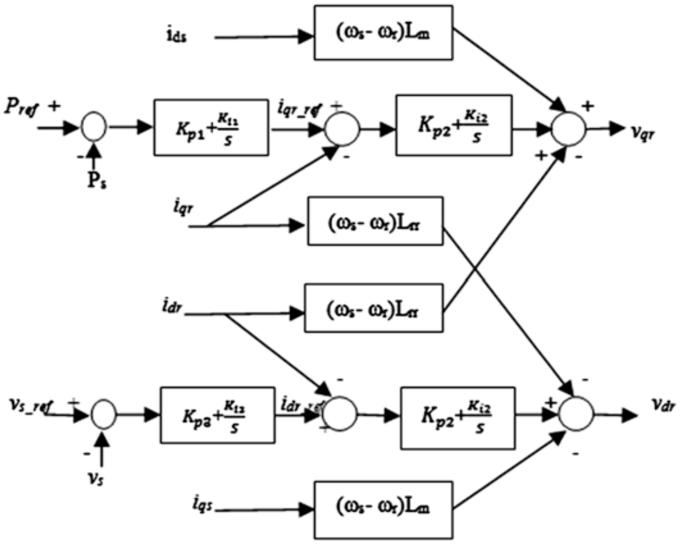


Fig. 2 Generator side converter controller

2.8 Grid Side Converter Controller Model

The grid side converter controller used here is represented by four states ($\dot{x}_5$, $\dot{x}_6$, $\dot{x}_7$ and $\dot{x}_8$), x_5 is related to the difference between the DC voltage and the reference DC voltage required at a certain time, x_6 related to the difference between the grid terminal voltage and the reference terminal voltage required at a certain time, x_6 is a combination of x_5 and direct axis grid current and x_8 is a combination of x_6 and quadrature axis grid current as shown in Eqs. (35–40). Figure 3 shows the grid side converter controller.

$$\dot{x}_5 = v_{DC\_ref} - v_{DC} \tag{35}$$

$$\dot{x}_6 = K_{p4}\dot{x}_5 + K_{i4}x_5 - i_{Dg} \tag{36}$$

$$\dot{x}_7 = v_{t\_ref} - v_t \tag{37}$$

$$\dot{x}_8 = K_{p6}\dot{x}_7 + K_{i6}x_7 - i_{Qg} \tag{38}$$

$$v_{Dg} = K_{p5}\dot{x}_6 + K_{i5}x_6 + X_c i_{Qg} \tag{39}$$

$$v_{Qg} = K_{p5}\dot{x}_8 + K_{i5}x_8 - X_c i_{Dg} \tag{40}$$

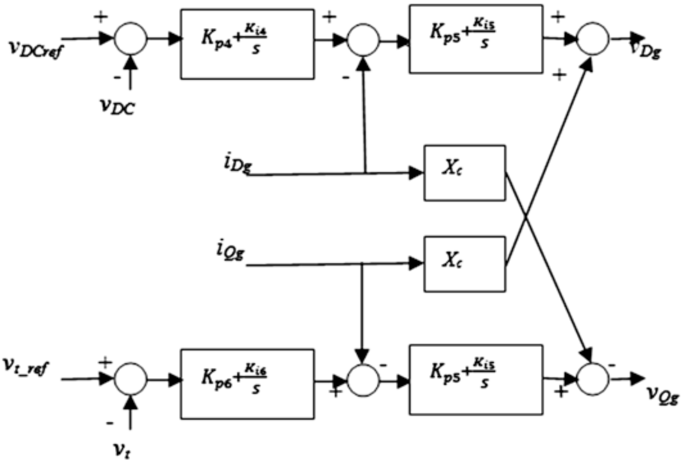

+
v_{DCref}
-
v_{DC}
$K_{p4}+\frac{K_{i4}}{s}$
+
-
$K_{p5}+\frac{K_{i5}}{s}$
+
+
v_{Dg}
i_{Dg}
X_c
i_{Qg}
X_c
v_{t_ref}
+
-
v_t
$K_{p6}+\frac{K_{i6}}{s}$
+
-
$K_{p5}+\frac{K_{i5}}{s}$
-
+
v_{Qg}

Fig. 3 Grid side converter controller

3 Tidal Current Turbine Model Using DDPMSG

3.1 The Dynamic Modeling of the DDPMSG

The DDPMSG can be modeled as the following [3, 12]:

$$\mathrm{v_{ds}} = -\mathrm{R_s} \times \mathrm{i_{ds}} - \omega_{\mathrm{s}} \times \psi_{\mathrm{qs}} + \frac{\mathrm{d}}{\mathrm{dt}}\Psi_{\mathrm{ds}} \tag{41}$$

$$\mathrm{v_{qs}} = -\mathrm{R_s} \times \mathrm{i_{qs}} + \omega_{\mathrm{s}} \times \psi_{\mathrm{ds}} + \frac{\mathrm{d}}{\mathrm{dt}}\Psi_{\mathrm{qs}} \tag{42}$$

The flux linkages and the torque can be expressed as:

$$\Psi_{ds} = -L_d \times i_{ds} + \psi_f \tag{43}$$

$$\Psi_{qs} = -L_q \times i_{qs} \tag{44}$$

$$T_e = (3/2)p\ i_{qs}\left(\left(L_d - L_q\right) i_{ds} + \psi_f\right) \tag{45}$$

L_d, and L_q are the direct and quadrature inductances of the stator. Ψ_{f} is the excitation field linkage, and p is the number of pair poles. For simplicity we will assume that $L_d = L_q = L_{s,}$ and so the generator model can be rewritten in a state space representation as:

$$L_s \frac{d}{dt} i_{ds} = -v_{ds} - R_s \times i_{ds} + L_s \times \omega_s \times i_{qs} \tag{46}$$

$$L_s \frac{d}{dt} i_{qs} = -v_{qs} - R_s \times \ i_{qs} - \ L_s \times \omega_s \times i_{ds} + \omega \times \psi_f \tag{47}$$

The converters models used for the DFIG are the same converters that used for the DDPMSG keeping in mind that a full scale power converter is used.

3.2 The Generator Side Converter Controller Model for DDPMSG

The generator side converter controller used here is represented by two states only (x_1, and x_2), x_1 related to the difference between the generated power and the reference power that required at a certain time and x_2 related to the difference between the direct axis generator current and the reference current that required at a certain time [1, 12–14]. Figure 4 shows the generator side converter controller described by:

$$\dot{x}_1 = P_s - P_{ref} \tag{48}$$

$$\dot{x}_2 = i_{ds} - i_{ds\_ref} \tag{49}$$

$$v_{qs} = K_{p1}\dot{x}_1 + K_{i1}x_1 - L_s w i_{ds} \tag{50}$$

$$v_{ds} = K_{p2}\dot{x}_2 + K_{i2}x_2 + L_s w i_{qs} \tag{51}$$

Where: K_{p1}, K_{p2}, represent the proportional controller constants and K_{i1}, K_{i2} represent the integral controller constants for the generator side converter controller.

3.3 The Grid Side Converter Controller Model for DDPMSG

The grid side converter controller used here is represented by four states (x_3, x_4, x_5, x_6), x_3 related to the difference between the DC voltage and the reference DC voltage that required at a certain time, x_5 related to the difference between the terminal voltage and the reference terminal voltage that required at a certain time, x_4 is a combination of x_3 and direct axis grid current and x_6 is a combination of x_4 and quadrature axis grid current. Figure 5 shows the grid side converter controller.

$$\dot{x}_3 = v_{sDC\_ref} - v_{DC} \tag{52}$$

$$\dot{x}_4 = K_{p3}\dot{x}_3 + K_{i3}x_3 - i_{Dg} \tag{53}$$

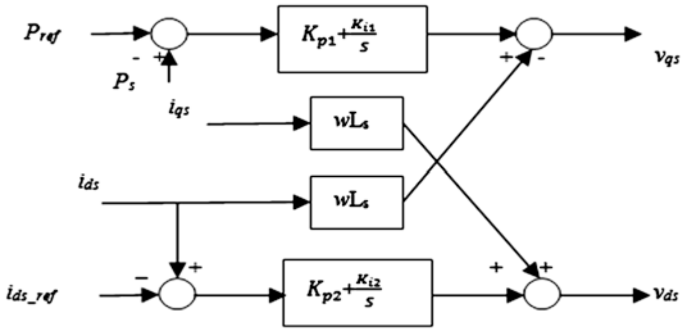


Fig. 4 Generator side converter controller for DDPMSG

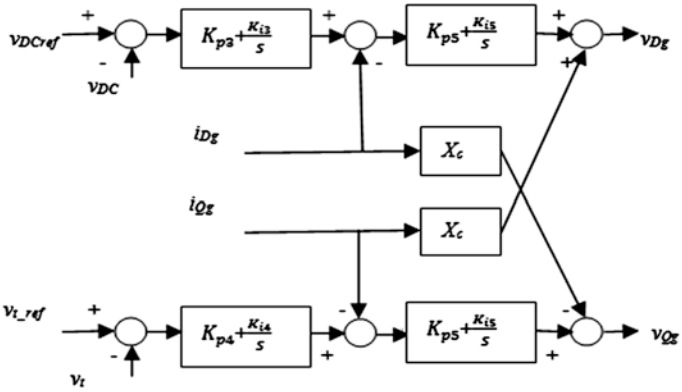

v_{DCref}
v_{DC}
$K_{p3}+\frac{K_{i3}}{s}$
$K_{p5}+\frac{K_{i5}}{s}$
v_{Dg}
i_{Dg}
X_c
i_{Qg}
X_c
v_{t_ref}
v_t
$K_{p4}+\frac{K_{i4}}{s}$
$K_{p5}+\frac{K_{i5}}{s}$
v_{Qg}

Fig. 5 Grid side converter controller for DDPMSG

$$\dot{x}_5 = v_{t\_ref} - v_t \tag{54}$$

$$\dot{x}_6 = K_{p4}\dot{x}_4 + K_{i4}x_5 - i_{Qg} \tag{55}$$

$$v_{Dg} = K_{p5}\dot{x}_4 + K_{i5}x_4 + X_c i_{Qg} \tag{56}$$

$$v_{Qg} = K_{p5}\dot{x}_6 + K_{i5}x_6 - X_c i_{Dg} \tag{57}$$

where: V_{DC} is the DC link voltage, i_{Dg}, i_{Qg} are the D and Q axis grid currents, v_{Dg}, v_{Qg} are the D and Q axis grid voltages, K_{p3}, K_{p4}, K_{p5} represent the proportional controller constants, X_c is the grid side smoothing reactance, and K_{i3}, K_{i4}, K_{i5} represent the integral controller constants for the grid side converter.

Table 1 Eigenvalues of the system using PI controllers for DFIG

Eigen value	Real part	Imaginary part	Freq.
λ_1	−10	–	–
λ_2	−1542.9	–	–
λ_3	−1	+j50	7.96
λ_4	−1	−j50	7.96
λ_5	−2	+j3	0.48
λ_6	−2	−j3	0.48
λ_7	−1444	–	–
λ_8	−100	–	–
λ_9	−27	–	–
λ_{10}	−175	–	–
λ_{11}	−27	–	–
λ_{12}	−10	–	–
λ_{13}	−8	–	–
λ_{14}	−10	–	–
λ_{15}	−27	–	–

Table 2 Eigenvalues of the system with PI controllers for DDPMSG

Eigen value	Real part	Imaginary part	Freq.
λ_1	−15	–	–
λ_2	−1	J200	31.8
λ_3	−1	−j200	31.8
λ_4	−48537	–	–
λ_5	−333	–	–
λ_6	−16	–	–
λ_7	−25	–	–
λ_8	−100	–	–
λ_9	−240	–	–
λ_{10}	−100	–	–

4 Tidal Current Turbine with the Proposed PI Controllers

In this section we find the eigenvalues for the overall system using the two types of generators. Tables 1 and 2 give the eigenvalues of the model for the system with the grid side and the rotor side controllers for DFIG and DDPMSG. From that we concluded that the system is asymptotically stable with the controllers.

5 Conclusion

Among renewable sources, tidal current energy shows great promise for satisfying future energy needs. However, as technology related to the energy sources is still in its infancy, vast improvements need to be made before it can truly become a commercially viable alternative. For instance, it is important to accurately forecast tidal current energy and to modify the control system depending on the forecasted model to adjust the output power. Also, during the faulty behavior, generators used for converting renewable energy will be different compared to generators used for the nonrenewable energy conversion.

The use of wind and tidal current as a renewable source of energy is very effective as it relies on similar technologies. The overall dynamic system of the offshore and tidal current turbine based on three different types of generators for a single machine infinite bus system has been modeled. The converter used has been modelled. Different controller models are also discussed for different machines. The state space representation of the overall system is concluded. The tidal current energy is more predictable compared to offshore wind energy. It is beneficial to use a hybrid of offshore wind and tidal current turbine.

References

1. H.H. Aly, M.E. El-Hawary, Modelling of Offshore Wind and Tidal Current Turbine for Stability Analysis, Lecture Notes in Engineering and Computer Science: in *Proceedings of The World Congress on Engineering and Computer Science 2013, WCECS 2013*, San Francisco, USA, pp 294–299, 23–25 October 2013
2. Tidal Stream, http://www.tidalstream.co.uk/html/background.html. Accessed Jan 2011
3. H.H.H. Aly, M.E. El-Hawary, State of the art for tidal currents electrical energy resources, in *24th Annual Canadian IEEE Conference on Electrical and Computer Engineering*, Niagara Falls, Ontario, Canada, 2011
4. H.H. Aly, M.E. El-Hawary, An overview of offshore wind electrical energy systems, in *23rd Annual Canadian IEEE Conference on Electrical and Computer Engineering*, Calgary, Alberta, Canada, 2–5 May 2010
5. M.V.A. Nunes, J.A.P. Lopes, H. Helmut, U.H. Bezerra, G. Rogério, Influence of the variable-speed wind generators in transient stability margin of the conventional generators integrated in electrical grids. IEEE Trans. Energ. Convers. **19**, 692–701 (2004)
6. J.G. Slootweg, H. Polinder, W.L. Kling, Dynamic modeling of a wind turbine with doubly fed induction generator, in *IEEE Power Engineering Society Summer Meeting*, 2001
7. J.B. Ekanayake, L. Holdsworth, X.G. Wu, N. Jenkins, Dynamic modeling of doubly fed induction generator wind turbines. IEEE Trans. Power Syst. **18**(2), 803–809 (2003)
8. M.J. Khan, G. Bhuyan, A. Moshref, K. Morison, An assessment of variable characteristics of the pacific northwest regions wave and tidal current power resources, and their interaction with electricity demand & implications for large scale development scenarios for the region. Technical Report 17485-21-00 (Rep 3), Jan 2008
9. L. Mihet-Popa, F. Blaabjerg, I. Boldea, wind turbine generator modeling and simulation where rotational speed is the controlled variable. IEEE Trans. Ind. Appl. **58**(1), 37 (2004)
10. H.H.H. Aly, M.E. El-Hawary, The current status of wind and tidal in-stream electric energy resources. Am. J. Electr. Power Energy Syst. **2**(2), 23–40 (2013)

11. H.H.H. Aly, M.E. El-Hawary, A proposed algorithms for tidal in-stream speed model. Am. J. Energ. Eng. **1**(1), 1–10 (2013)
12. F. Wu, X.-P. Zhang, P. Ju, Small signal stability analysis and control of the wind turbine with the direct-drive permanent magnet generator integrated to the grid. J. Electr. Power Eng. Res. **12**, 1661–1667 (2009)
13. F. Wu, X.-P. Zhang, P. Ju, Small signal stability analysis and optimal control of a wind turbine with doubly fed induction generator. IET J. Gener. Transm. Distrib, **5**, 751–760 (2007)
14. H.H. Aly, Forecasting, modeling, and control of tidal currents electrical energy systems. PhD thesis, Dalhousie University, Halifax, Canada, 2012

Chapter 2
Practical Approach in Design of HVAC Systems Utilizing Geothermal Energy

M. Fathizadeh and D. Seims

Abstract Geothermal is the Earth's thermal energy. In recent years geothermal energy has been utilized for generation of electricity, heating and air conditioning (HVAC). Geothermal HVAC systems are cost effective, energy efficient, and environmentally friendly way of heating and cooling buildings. The Department of Energy (DOE) and the Environmental Protection Agency (EPA) have both endorsed geothermal HVAC systems. Their flexible design requirements make them a good choice for schools, high-rises, government's buildings, commercial and residential properties. Lower operating and maintenance costs, durability, and energy conservation make geothermal a great alternative to conventional HVAC systems. This chapter gives the step-by-step for estimate, design, calculation, procurement, installation and commissioning of geothermal heating and air conditioning system for residential or small businesses. A comparison between the conventional HVAC and the geothermal counterpart is provided to demonstrate their advantages.

Keywords Air conditioning · Energy · Efficiency · Geothermal · Green energy · Heating · Renewable · Ventilation

1 Introduction

A Geothermal Heat Pump (GHP); also known as a ground source heat pump is a central heating and/or cooling system that pumps heat to or from the ground. To transfer energy to or from the ground, geothermal systems typically have a "loop" which consists of a long length of pipes. As a thermal fluid is pumped through the

M. Fathizadeh (✉) · D. Seims
Purdue University Calumet, 2200 169th Street, Hammond, IN 46323, USA
e-mail: fathizad@purduecal.edu

D. Seims
e-mail: siems@purduecal.edu

H. K. Kim et al. (eds.), *Transactions on Engineering Technologies*,
DOI: 10.1007/978-94-017-9115-1_2,

loop, conductive heat transfer occurs between the fluid, piping, and the ground. It uses the earth as a heat source (in the winter) or a heat sink (in the summer). This design takes advantage of the stable temperatures in the ground to boost efficiency and reduce the utility costs of heating and cooling systems. Like an air conditioner, these systems use a heat pump to transfer heat from the ground [1].

A heat pump is basically a loop of refrigerant pumped through a vapor-compression refrigeration cycle that moves heat1. Although many parts of the country experience seasonal temperature extremes from scorching heat in the summer to sub-zero cold in the winter—a few feet below the earth's surface the ground remains at a relatively constant temperature. Depending on latitude, ground temperatures range from 45 °F (7 °C) to 75 °F (21 °C). Like a cave, this ground temperature is warmer than the air above it during the winter and cooler than the air in the summer. The GHP takes advantage of this by exchanging heat with the earth through a ground heat exchanger [2]. Using the earth is very energy-efficient because underground temperatures are far more stable than air temperatures through the year. Especially during the hot and humid summer, an earth loop heat exchanger provides a greater temperature differential than conventional Air Conditioning (A.C.) condensers or cooling towers. There is insignificant seasonal temperature variation below 30 ft. Much like a cave, the stable ground temperature is warmer than the air above during the winter and cooler than the air in the summer.

The initial costs of geothermal HVAC systems are generally higher than conventional systems, but the difference is usually returned in energy savings in as little as 3 years. System life is estimated at 25 years for inside components and 50 years or more for the loop piping. There are approximately 50,000 geothermal heat pumps installed in the United States each year3.

Ground source heat pumps are categorized as having closed or open loops, and those closed loops can be installed in three ways: horizontally, vertically, or in a lake/river. The type chosen depends on the available land areas, the soil and rock type at the installation site. These factors will help determine the most economical choice for installation of the ground loop. For closed loop systems, water or an antifreeze solution is circulated through plastic pipes buried beneath the earth's surface. During the winter, the fluid collects heat from the earth and carries it into the building. During the summer, the system reverses itself to cool the building by removing heat from the building and placing it in the ground.

Geothermal systems are very similar to conventional boiler/tower systems. Both boiler/tower and geothermal systems use basically the same heat pump equipment and the Coefficient of Performance (COP) are similar when rated at the same conditions. Geothermal heat pumps can operate at a greater range in temperature allowing a larger temperature differential. The major difference in efficiency is in the external heat exchanger. Geothermal systems use the constant reliable temperature of the earth as a sink or a source for energy while boiler/tower systems use the seasonally affected ambient air [1–4].

One advantage of the closed loop system design is that it eliminates large ductwork runs. The concept allows energy that is not required in some areas of the building (cooling load) to be moved and used in areas that do require energy (heating load). Closed building loop design would be applicable to hotels where each room has its own independent control. When a portion of the rooms are heating and a portion is cooling, the building loop allows the zones to simply "trade" energy. Applications that could benefit would be facilities that have both heating and cooling process running simultaneously.

Water source heat pump systems (chilled water systems) generally require smaller mechanical rooms than other HVAC systems. Geothermal mechanical rooms are even smaller, requiring space for only the circulating pumps and the main header piping. This frees up valuable building space [5].

2 Cumulative Annual Energy Load on Ground Loop

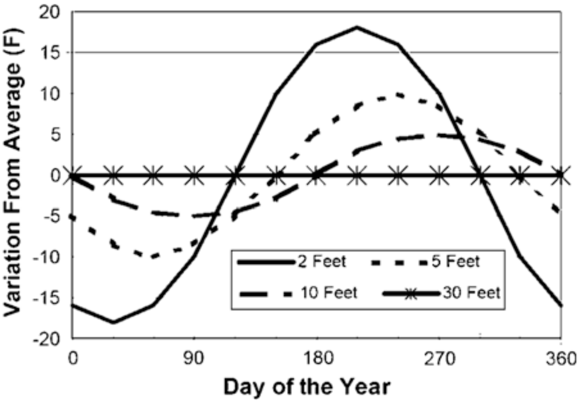
Cumulative annual energy load is the change in the ground temperature over many years. If the building has a net heat gain or a net heat loss, the ground temperature will change. This is referred to as a thermal-flywheel effect. The more closely placed the boreholes or trenches, the larger the effect. Ground water moving through the borehole field can remove substantial energy and limit the long-term temperature changes. Most commercial buildings have a very high net cooling load; this can cause the earth temperature of the loop to rise. It is not uncommon for skyscrapers in Chicago to never turn on their boilers throughout the winter. Long term effects must be considered when designing a ground loop. Long term temperature rise is the most common problem in large-scale geothermal systems. Typically the loop return fluid temperature will rise in the first few years but it should settle to consistent annual variations [6].

The graph of Fig. 1 shows the variation from ground temperature for various depths. Most horizontal systems usually stay within 5 ft. of the surface, which can swing as much as 20 °F from summer to winter. Evaporation can also cool the surface soil and improve horizontal loop performance.

2.1 Thermal Properties of Soil

Soil or rock composition greatly affects the thermal properties of a ground loop. A sieve test can be performed to determine the composition of sandy and clay soils. A soil map can usually be attained from the county surveyor's office or the Department of Natural Resources. A soil study may only be necessary for larger projects; most designers should already be aware of the composition of the soil commonly found on the project site [7–9].

Fig. 1 Seasonal ground temperature depth variation



2.2 Ground Water and Ground Temperature

Ground temperature is best obtained from local water well logs and geological surveys. Ground water movement through the bore-hole field can have a large impact on its performance [10].

2.3 Pressurized Versus Non-pressurized Flow Centers

Geothermal pumping configurations (also referred to as flow-centers) consist of the "indoor" portion of the piping, the pumps, and the heat pumps. Flow-centers designated as either pressurized or non-pressurized. Pressurized flow-centers typically require a contractor to pressurize the system to 40 psi with flushing pump. The loop should maintain the hydraulic pressure indefinitely. However, debris and dissolved air in the system often require the system to be flushed and recharged.

A non-pressurized configuration is vented to atmospheric pressure typically at the highest point on the system to help eliminate air. Advantages of non-pressurized flow-centers include ability to easily check the fluid chemistry and fluid level (Fig. 2).

2.4 Net Positive Suction Head Required (NPSHR)

The major disadvantage of non-pressurized flow-centers is the lack of static pressure in the system required to operate the pumps at higher fluid temperatures. Pumps have a specified Net Positive Suction Head Required (NPSHR). This is the pressure required on the suction side of the pump so the pumps will not cavitate.

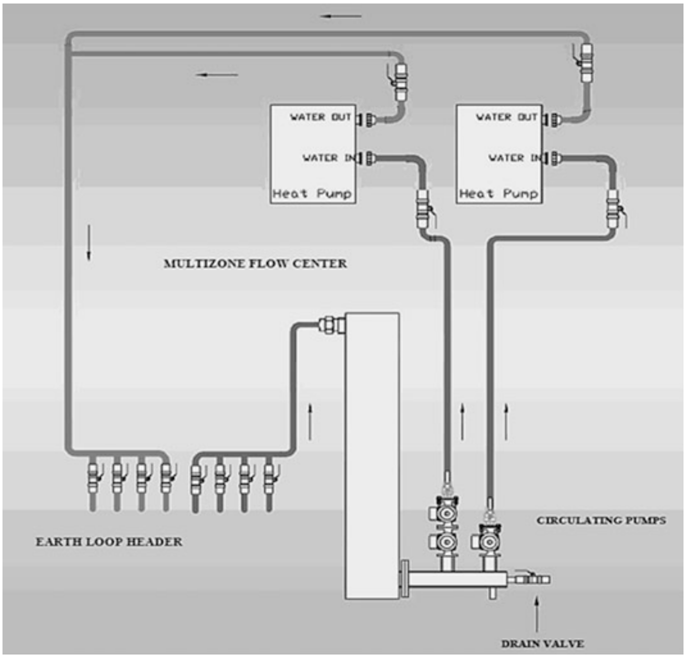


Fig. 2 Non-pressurized flow center configuration

This information is available from pump manufacturers. With a non-pressurized flow-center, the fluid does not have static pressure when vented to atmosphere. A tall vertical tank just before the pump is recommended to maintain a small hydrostatic pressure on the pump. Pump cavitation is a function of fluid temperature, fluid properties, hydrostatic pressure, and NPSHR of the pump.

2.5 NPSHA Calculation-(Cavitation Prevention)

To prevent pump cavitation, the Net Positive Suction Head Available (NPSHA) must be greater than the Net Positive Suction Head Required (NPSHR).

$$\text{NPSHA} > \text{NPSHR} \tag{1}$$

The objective is to determine adequate height for the water column on the flow center while checking the margin of safety for pump cavitation. NPSHR is given by the pump manufacturer. This is the pressure and temperature conditions at which the pump will “pull the water apart” causing cavitation. Cavitation will immediately destroy the pumps which in turn will cause system failure.

$$\text{NPSHA} = H_a + H_s + H_{vpa} + H_f \tag{2}$$

where:

H_a the atmospheric pressure on surface of liquid entering the pump

H_s the Static elevation above pump

H_{vpa} the absolute vapor pressure at max liquid temperature

H_f the friction head losses on suction side of pump

2.6 Practical NPSHA Calculation

H_a (at 710 ft. above sea level) = 32 ft. of head

H_s what we are solving for. This is the required height of the flow center water column

H_{vpa} at max temp of 140 °F = 6.6 ft. of head (140 °F is the maximum design temp for the system. This is a rule of thumb for open systems. This temperature restriction is also caused by the use of Schedule 40 PVC piping in the system). Table 1 shows the minimum pressure for different water temperatures.

H_f is small enough to be negligible for the flow rate of our system. The water column is 6 in. diameter pipe with a short (<2 ft. long) section of horizontal 3 in. piping. Therefore:

$$H_f = 0.$$
$$\text{NPSHA} = 32 + H_s + (-6.6) - 0$$
$$\text{NPSHA} = 25.4 + H_s$$

For the family of pumps selected for our system, the minimum inlet pressure is given by the following: Grundfus UP-26 family of pumps (From Table 2). Therefore, at 140 °F maximum system temperature:

$$\text{NPSHR} = 3 \text{ ft. of } H_2O \text{ column}$$

To create the equation to determine our Hs, the previous values and set NPSHA equal to NPSHR.

$$\begin{aligned} \text{NPSHA} &= 25.4 + H_s \\ \text{NPSHR} &= 3 \\ 3 &= 25.4 + H_s \\ H_s &= (-22.4) \end{aligned}$$

Table 1 Minimum inlet pressure for different water temperature

Water temperature in °F	Min inlet pressure in feet of head H_20
30	36
190	14
140	3

Therefore, with the UP family of pumps (Table 2), it should be not expected to have any cavitation. This leaves a safety margin of 22.4 ft. (9.7 PSI). The system will still have a water column to help balance volumetric thermal expansion and to provide a visual water level indicator. The water column also will act as a means of air bubble elimination.

If the system temperature was above 200 °F, or a lower quality pump was selected, the water column would have to be taller to accommodate the NPSHR System.

3 Configuration and Flow Calculations

The curve shown in Fig. 3 represents the calculation methods used to determine and plot the piping network's performance curve. The performance curve is the system's loss due to friction plotted against its flow rate. Since the recommended flow rates were specified by the heat pump manufacturer, design flow rates were simply divided by the number of pipes in parallel.

3.1 Performance Curve

The system's friction head is calculated at each node in the piping network for three different corresponding flow rates. Please note that friction losses due to pipe fitting were simply added to the length using a corresponding length of pipe given in manufacturer's datasheets. Using the systems cumulative friction head the systems performance curve can be plotted.

3.2 Pump Selection

The next step is to select a pumping configuration that can achieve the proper flow to match the systems curve. Pump curves are available in datasheets from the pump manufacturer. The selected pumps had to be specified with bronze casings to prevent corrosion. The data from three similar pumps was plotted in excel to create

Table 2 Pump selection for different flow rate

UPS26-99F-LS				UPS26-99F-MS				UPS 26-99F-HS			
GPM	HD	GPM X 2	HD X 2	GPM	HD	GPM X 2	HD X 2	GPM	HD	GPM X 2	HD X 2
0	24.25	0	48.5	0	28	0	56	0	29.25	0	58.5
2	20.75	4	41.5	2	26.25	4	52.5	2	28	4	56
4	17.25	8	34.5	4	24.25	8	48.5	4	26.75	8	53.5
6	14	12	28	6	22.5	12	45	6	25.5	12	51
8	11	16	22	8	20.5	16	41	8	24.25	16	48.5
10	8.5	20	17	10	18.5	20	37	10	22.75	20	45.5
12	6.25	24	12.5	12	16.5	24	33	12	21.25	24	42.5
14	4.25	28	8.5	14	14.5	28	29	14	19.5	28	39
16	2.5	32	5	16	12.25	32	24.5	16	18	32	36
18	1	36	2	18	10.25	36	20.5	18	16.25	36	32.5
20	0	40	0	20	8	40	16	20	14.25	40	28.5
22		44		22	6	44	12	22	12.25	44	24.5
24		48		24	3.75	48	7.5	24	10	48	20
26		52		26	1.5	52	3	26	7.5	52	15
28		56		28	0	56	0	28	5	56	10
30		60		30		60		30	2.25	60	4.5
32		64		32		64		32	0	64	0
34		68		34		68		34		68	
36		72		36		72		36		72	
38		76		38		76		38		76	
40		80		40		80		40		80	
42		84		42		84		42		84	
44		88		44		88		44		88	

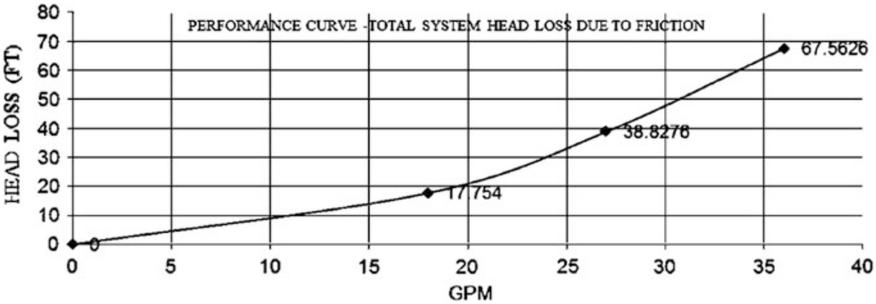


Fig. 3 System performance curve showing head loss due to friction

pump curves and are shown in Fig. 4. Since two duplex pumps are running in parallel, the head must be doubled for duplex and doubled the flow for parallel. Figure 4 shows the pump curves against the system's performance values. A chart for the pump selection for different values of flow rates is shown in Table 2.

By using the above chart, a pump selection can be made. The total system flow rate is based on the cumulative nominal flow recommended by the heat pump manufacturer. For this case is 38.8 gallons per minute. The pump was specified for

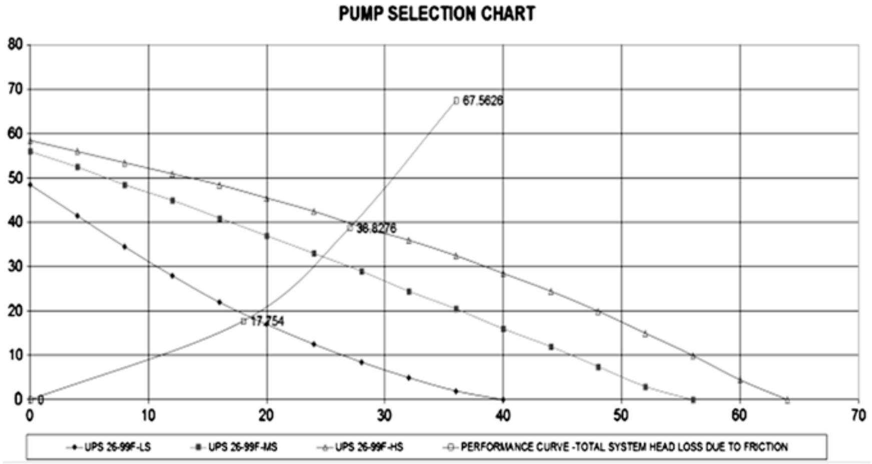


Fig. 4 Pump curves plotted against the system's performance

the system is the Grundfos UPS 26-99F-HS in parallel duplex configuration. This was the correct choice of pump and was later verified by measuring the differential pressure across the units and referencing it with their given head loss. The primaries were extended to the rear of the building to feed the secondary loop field.

4 Project Cost and Comparison to Applicable High Efficiency Alternatives

Figure 5 shows the spread sheet for the project budget. The project budget is given in the light blue column below. The other columns are alternative systems that were quoted for the project.

4.1 Natural Gas Fired Boiler and Hydronic Heat System

An 110,000 BTU natural gas boiler was installed to provide hydronic heating to the plant.

The heated water is pumped through 10,000 ft. of ½″ PEX tubing suspended in the concrete floor. The boiler is an HTP Elite-110 modulating and condensing boiler that currently the most efficient boiler on the market. The 32′ tall PVC flue stack allows a larger condensing area to reclaim heat that would normally be lost with the exhaust gas. With the longer flue, the boiler may very well be more efficient than the advertised 98.1 % AFUE.

Cost Comparison of Applicable High EfficiencyAlternatives			
MFG	Diakin Macquary	Diakin Macquary	Trane
Type	GSHP(Temp Range)	Air Cooled Curt/Rot Scoll Compressor	Water Unitr Cooled Roof Top Unitry
HeatPump Unit 2	$ 8,296.00	11,513.00	11,628.00
Installed Duckwork	$ 14,550.00	16650.00	14550.00
Underground PE Tubing	$ 3,743.00		
Building Piping	$ 1,200.00	300.00	1000.00
Circulation Pumps	$ 1,430.00		
Special Tools and Equipment Rental	$ 4,500.00	1900.00	1900.00
Central System	$ 1,250.00	1250.00	1250.00
Total Labor	$ 15,000.00	10000.00	11000.00
Total Installed Cost	$ 49,969.00	41513.00	42438.00
Subtracted Fedral Investment Tax Credit 10%	$ 44,972.10		
Subtracted State Property Tax Credit 1%	$ 44,522.38		
Initial Differential Cost	$ 2,496.88		
Unin EER (5 ton used for comparison)	19.2	13.9	13.1
Total System EER	19	13.9	12.9
System Capacity (BTU)	108000	108000	108000
Muliply by Annual Runtime (Hrs)	1500	1500	1500
BTU-Hours Produced Per Year	162000000	162000000	162000000
Divide by System EER	19	13.9	12.9
Watt-Hours Consumed Per Year	8526315.789	11654676.26	12558139.53
Muliply by W to KW Conversion	0.001	0.001	0.001
Annual kWh Consumed	8526.315789	11654.67626	12558.13953
Multiply by Cost of Electricity	$ 0.12	$ 0.12	$ 0.12
Annual Cost of Electricity for Cooling	$ 1,023.16	$ 1,398.56	$ 1,506.98
Annual Electicity Savings	$ 429.61		
Energy Reduction Utility Rebate (0.045/kW)	$ (140.78)		
Total Annual Electrical Savings	$ 570.39		
Divide by Differential Cost			
Payback Period (Years)	4.40		

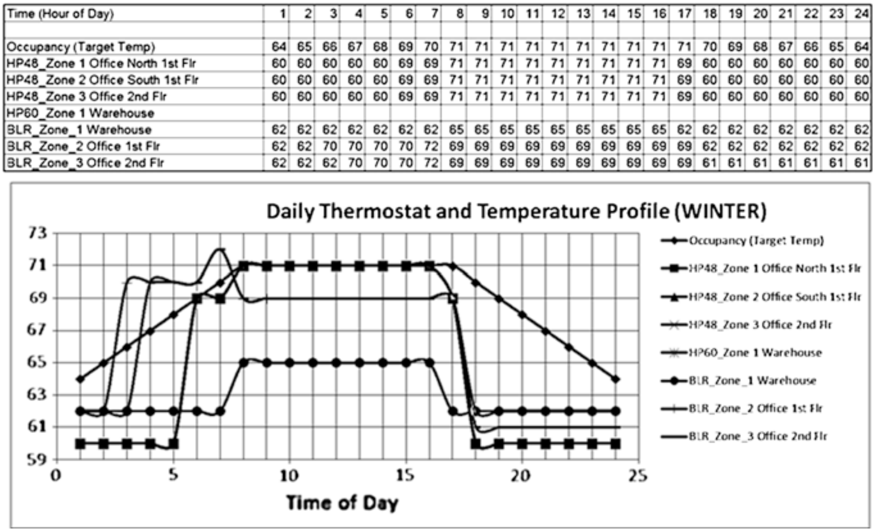
Fig. 5 Spreadsheet showing cost breakdown and payback period of applicable project alternatives

The boiler is the primary heat source. The large thermal mass of the concrete floor will also allow temperature to be easier to control. Once the entire system is tuned properly, the boiler should only fire one time each day per zone. (See thermostat profiles shown in Fig. 6). Using the concrete floor to radiate heat will also allow the plant to operate at a lower temperature. This is due to the fact that the human comfort temperature level is lower when the floor is warm.

The insulating foam on the tubing is installed. This is to prevent the boiler room from overheating. The tubing is temporarily terminated in series for pressure testing. The system was held at 150 PSI prior to and during the concrete pour so any punctures caused by the concrete contractors would be immediately apparent.

4.2 Thermostat Setting

The heating and cooling system zones were controlled by seven programmable thermostats. We wanted the boiler to preheat the building every morning to minimize the load on the heat pumps throughout the day because they are not as



Time (Hour of Day)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
Occupancy (Target Temp)	64	65	66	67	68	69	70	71	71	71	71	71	71	71	71	71	71	70	69	68	67	66	65	64
HP48_Zone 1 Office North 1st Flr	60	60	60	60	60	69	69	71	71	71	71	71	71	71	71	71	69	60	60	60	60	60	60	60
HP48_Zone 2 Office South 1st Flr	60	60	60	60	60	69	69	71	71	71	71	71	71	71	71	71	69	60	60	60	60	60	60	60
HP48_Zone 3 Office 2nd Flr	60	60	60	60	60	69	69	71	71	71	71	71	71	71	71	71	69	60	60	60	60	60	60	60
HP60_Zone 1 Warehouse																								
BLR_Zone_1 Warehouse	62	62	62	62	62	62	62	65	65	65	65	65	65	65	65	65	62	62	62	62	62	62	62	62
BLR_Zone_2 Office 1st Flr	62	62	70	70	70	70	72	69	69	69	69	69	69	69	69	69	69	62	62	62	62	62	62	62
BLR_Zone_3 Office 2nd Flr	62	62	62	70	70	70	72	69	69	69	69	69	69	69	69	69	69	61	61	61	61	61	61	61

Daily Thermostat and Temperature Profile (WINTER)
73
71
69
67
65
63
61
59
0
5
10
15
20
25
Time of Day
Occupancy (Target Temp)
HP48_Zone 1 Office North 1st Flr
HP48_Zone 2 Office South 1st Flr
HP48_Zone 3 Office 2nd Flr
HP60_Zone 1 Warehouse
BLR_Zone_1 Warehouse
BLR_Zone_2 Office 1st Flr
BLR_Zone_3 Office 2nd Flr

Fig. 6 Programmable thermostat profiles

efficient for heating as the boiler. The boiler is programmed to overshoot the desired temperature for one hour prior to occupancy because the hydronic system's large thermal mass would keep the heat for most of the pre-noon hours of occupancy. The warehouse portion of the building needed only to be heated to 65° and care was taken to make sure the warehouse and office zones did not fire at the same time. Please see thermostat profiles below.

5 Government Subsidies

There are many federal, state or local government subsidies in the form of incentives, lower utility cost or tax credit available for potential customers to use. Several conditions must be met to qualify for subsides or incentives. In the following a few of these conditions are listed

- There is no maximum credit for systems placed in service after 2008.
- Systems must be placed in service on or after January 1, 2008, and on or before December 31, 2016.
- The geothermal heat pump must meet federal Energy Star criteria.

The home served by the system does not have to be the taxpayer's principal residence.

6 Summary

A practical step by step design procedure for the design and installation of a geothermal HAVC system was given. Calculation and specific consideration for the design was mentioned. Equipment selection and performance criteria presented. The cost evaluation and performance criteria were determined and a comparison between the conventional HVAC system and the new geothermal one is shown. A brief version of this publication was presented [10].

References

1. McQuay International: Application guide engineering system solutions newsletter. Optimizing geothermal heat pump systems for higher efficiency, maximum LEED points and lower installed costs (2005). Available at www.go.mcquay.com
2. Cost Containment for Ground Source Heat Pumps. Final Report Submitted to Alabama. Available at http://www.geoheat.oit.edu/pdf
3. K. Harsh Gupta, R. Sukanta, *Geothermal Energy: An Alternative Resource for the 21st Century*, 1st edn. (Elsevier Science Publishing, Amsterdam, 2006)
4. Indiana Department of Natural Resources. Available at http://www.in.gov/dnr/water/
5. Daikin Press Releases. Available at www.daikinpress.com/pressrelease.php
6. J. Poppei, R. Schwarz, Innovative Improvements of Thermal Response Test Intermediate Report, Aug 2006
7. ASTM International-Standards Worldwide: *ASTMC136-06* (2006). Available at http://www.astm.org
8. Grundfos Industrial Solutions: Pump handbook (2004). Available at www.grundfos.com
9. Water Source Heat Pump Design Manual: A design manual for the professional engineer C330-1 (1999), AAF-McQuay Incorporated. Available at www.mcquay.ru/downloads
10. M. Fathizadeh, D. Seims, Design and implementation of geothermal systems for heating and air conditioning, lecture notes in engineering and computer science, in *Proceedings of the World Congress on Engineering and Computer Science* 2013 vol. I, WCECS 2013, San Francisco, pp. 326–331, 23–25 Oct 2013

Chapter 3
Efficiency Analysis of Submersible Induction Motor with Broken Rotor Bar

Hayri Arabaci and Osman Bilgin

Abstract This study analyzes effects of squirrel cage faults on submersible induction motors efficiency at steady-state condition. There are a lot of studies about effects of the cage faults on motor performance. Especially, the effects of the cage faults on the motor parameters such as current, torque and speed are well known. Unlike the literature, cage fault effects on efficiency are analyzed in this study. Furthermore, fluctuations and mean value changes resulting from the rotor faults are ranked according to size of these faults. Healthy and five different faults were investigated by using 10, 25, 30 and 50 HP submersible induction motors in both simulations and experiments. Time stepping finite element method solution was used to compute motor quantities in the simulation. Good agreement was achieved between simulation and experimental results. The effects of rotor faults on motor efficiency were clearly ranked according to size of faults.

Keywords Efficiency analysis · Energy efficiency · Finite element method · Induction motors · Rotor faults · Squirrel cage

1 Introduction

Induction motors are widely used in industry because of their durability and low production costs. They are generally used to drive moving mechanical loads. Unlike traditional induction motors, submersible induction motors (SIMs) are used

H. Arabaci (✉)
Department of Electrical and Electronics Engineering, Technology Faculty,
Selcuk University, 42075 Konya, Turkey
e-mail: hayriarabaci@selcuk.edu.tr

O. Bilgin
Department of Electrical and Electronics Engineering, Engineering Faculty,
Selcuk University, 42075 Konya, Turkey
e-mail: obilgin@selcuk.edu.tr

H. K. Kim et al. (eds.), *Transactions on Engineering Technologies*,
DOI: 10.1007/978-94-017-9115-1_3,

to drive pump to carry underground water to the surface, seawater lift service on offshore platforms and floaters or other storage facilities. Although working principles are same with traditional induction motors, it has some small differences in their structures. Diameter of the motor is limited because of size of wells drilled. Most SIMs are designed to fit in four inch, six inch wells or larger. To increase motor power, rotor length is extended because of the limited diameter. Therefore, lengths of SIMs are longer and their diameters are shorter from traditional induction motors.

Despite their reliability and stiffness of SIMs, some faults may still occur. Majority of faults occur at stator side but some of they occur in rotor. Rotor faults may result from production faults or from mechanical, environmental, electromagnetic or thermal pressure exerted on the unit [1].

Faults other then production errors will occur as small-scale faults. Because of above listed pressures, initial faults may grow and may result in increasing the scale of potential subsequent faults [2]. The faults negatively affect motor performance [3]. The effects of faulted rotor are recorded as unbalanced motor currents, torque fluctuations, increasing in loss, larger thermal pressures and degradation of transient performance. Symptoms of the faults are small-scale and could not be detected by simple methods. Even though the rotor fault is at an early stage or it reaches a certain size, the motor does not stop and the fault can not be understood. It requires deeply analysis. Losses of energy in this process have not been taken into account. Especially considering irrigation systems, a lot of energy loss arises [4].

Most of studies about the rotor faults have been conducted to detection and diagnosis of the faults [5–19]. In the most of such studies, a current analysis method has been adopted [13–24] and a rotor with one broken bar has been studied as the rotor fault. Vibration analysis [25] and flux analysis [26–28] are also available in the diagnosis of rotor faults. While some of these studies have examined two or three broken bars or a broken short circuit ring, they have focused on the available indicators of the motor current and concentrated on methods which are for diagnosis rather than effects on performance.

A number of studies have proposed to investigate the effects of broken rotor bars and on their performances in many decades. Initially, motor performance analyses were made by using equivalent circuit of induction motor [29]. In later years, magnetic field [30] and differential equations were used for dynamic behavior of induction motor. Solutions of these equations were done by computer [31]. The development of computers and software has enabled the use of finite element methods (FEM) to solve the equations of mathematical model of induction motors. [32] presents detail induction motor performance computation and prediction using combined finite element-state space model of induction motor which includes rotor bars. Motor current, torque and rotor-stator losses are analyzed within context of the performance but the efficiency is unavailable. In [33], effects of broken bar faults on motor torque and current are investigated by using time-stepping coupled finite element-state space modeling approach. So, it is shown that the broken bar fault simulated and analyzed by using FEM. Faiz and

Ebrahimi [34] analyzes current spectrum, torque, speed, magnetic flux distribution, magnetic potential vector and magnetic flux density of motor by time-stepping finite element method (TSFE) in a healthy and with broken rotor bar. The field, current, torque, speed, and their relationship are analyzed and used for diagnosis. Li et al. [35] investigates the effects of rotor bar faults by TSFE approach on motor performance which contains stator current waveforms, current density and magnetic force distribution. TSFE method is used for modeling of induction motor with broken rotor bars in [36]. It investigated and ranked the effects of broken bar faults on amplitudes of harmonics components of current and torque. In this modeling, geometrical and physical characteristics of all parts of the motor, spatial distribution of stator windings, slots on both sides of the air gap and non-linear characteristic of the core materials are included. That is while the current of the broken bar was taken to be non-zero, instead resistance of the broken bar was considered large enough. Kurihara et al. [37] analyzes performance of single phase induction motor by using 2D TSFE approach. Simulation results are verified by experimental result. The current, torque and efficiency of motor were investigated for performance analysis.

The current, torque, speed and magnetic flux of motor have been included into performance content at the studies about motor performance analysis so far. Furthermore the size of rotor faults effect on the motor efficiency is not known in detail. So in the presented study, efficiency of motor are included to the motor performance and the effects of rotor faults on motor efficiency under steady-state condition is ranked according to size of the rotor faults.

Three-phase, 50 Hz, 380 V, two poles, 10, 25, 30 and 50 HP squirrel cage submersible induction motors were used in experiments and in simulations conducted on the five different types of faults and on a healthy rotor. Within the scope of the present study, five different rotor faults were analyzed. The efficiency of motor was analyzed in the scope of the motor performance. The effects on the efficiency were detected via comparison between the values of a healthy rotor and faulted rotors.

The simulations results have been obtained by using the FEM solution in order to investigate the effects of the broken rotor bars. FEM is able to compute the magnetic field distribution within the motor using geometry and magnetic parameters of the motor. Having the magnetic field distribution, other quantities of the motor can be obtained [38, 39]. Modeling of the rotor cage is the first step in the design of the SIMs. Because the field picture totally changes, the situation is more complex in simulation of rotor cage faults [40]. The finite-element analysis (FEA) can be used to model the induction motor with the rotor cage modeling [41]. The FEA has been coupled to circuit simulation. This external circuit coupling allows to simulate the operating conditions of the induction motor with the real power-supply connections [42]. Fixed load was used loading of the motor in simulation.

Unlike the traditional induction motor, submersible motor' length is longer and it's diameter is shorter. So, material of rotor shaft will be important. If the material is not magnetic, leakage flux in motor will be early reached to saturation. Therefore, the shaft material should be magnetic material in simulation model. This material in other induction motors is generally chosen nonmagnetic [43].

The following assumptions have been used for simplification in the FEM solution procedure for SIMs, similar to that in classical induction motor [35].

- The rotor bars are insulated from the rotor core, and there is no direct electrical contact between the rotor bars and the rotor core.
- The leakage on the outer surface of the stator and the inner surface of the rotor is neglected.
- The 2-D domain is considered, the magnetic vector potential and the current density have only the axial z component.
- Displacement current is neglected because the frequency of the source is very low.

Geometrical and physical characteristics of all parts of the motor, spatial distribution of stator windings, slots on both sides of the air gap and non-linear characteristic of the core materials are included in the present modeling as in [36].

2 Efficiency Analysis

Stator current has been generally used in the investigating of rotor faults. The faults affect the current according to motor slip as in Eq. (1).

$$f_b = (1 \pm 2ks)f, \quad k = 1, 2, 3, \ldots \tag{1}$$

where f is main frequency and s is motor slip. The effects cause fluctuations on the current. Because the current is directly related to other motor' outputs, the motor performance is indirectly affected by the fluctuations. In this study, the motor efficiency is evaluated. The following two criteria are used for investigating:

- The ratio of fluctuations,
- The change in the mean values.

$$\text{The fluctuations ratio } = 2 * \frac{\text{Max. value} - \text{Min value}}{\text{Max. value} + \text{Min. value}} \tag{2}$$

The fluctuation ratio is calculated according to Eq. (2). These outputs are analyzed for each fault size and they are compared with other fault sizes and healthy rotor.

Broken rotor bar faults and end ring fault are mainly caused by manufacturing defects. Especially, rotor bars of low and medium power motors are generally made of casting. Small defects may occur during casting process and it grows and causes important faults. Copper bar are generally used in rotor of the SIMs, while welding bars to end rings some defect (i.e. bad welding and small crack in end rings) may occur. The analyzed rotor faults are shown as follows:

- A rotor with one broken rotor bar,
- A rotor with two adjacent broken rotor bars,
- A rotor with three adjacent broken rotor bars,
- A rotor with high resistance rotor bar,
- A rotor with broken end ring (only in experimental study).

The simulations and experiments were made by using four different squirrel cage SIMs. The specifications of motors were;

Motor1	50 HP, 8″, with 18 bars, 380 V, 2 poles and 50 Hz
Motor2	30 HP, 8″, with 18 bars, 380 V, 2 poles and 50 Hz
Motor3	25 HP, 6″, with 22 bars, 380 V, 2 poles and 50 Hz
Motor4	10 HP, 6″, with 22 bars, 380 V, 2 poles and 50 Hz

3 Simulation and Results

A mathematical model, in which rotor parameters can be changed, is needed to simulate the rotor cage faults. Therefore, the model of squirrel cage is included within the mathematical model of SIMs. To solve such a model by circuit solutions requires a lot of assumptions (very small air gap, infinite magnetic permeability, magnetic saturation etc.). But in the FEM solution fewer assumptions are needed. So in this study simulation of SIMs were made by using FEM. FEM analysis has been coupled to the circuit simulation. This external circuit coupling allows to simulate the operating with the real power-supply connections. Time stepping transient magnetic 2D FEM were used in the solutions.

In the FEM modeling, benefiting from symmetry of motor structure reduces to the number of calculations. Because the used motors in this study have two poles, half of the motors can be modeled for simulations, but we used whole of motor which is need to simulate the one broken bar fault.

The efficiency is one of the important outputs of motor. The fluctuation in the wave forms of current affects the mechanical power and input power of the motor. Therefore, the efficiency changes in time. Figures 1 and 2 shows these changes for steady state operations for Motor1 and Motor4 respectively. There are considerable fluctuations according to size of fault. The ratios of the size of these fluctuations are listed in Table 1. The fluctuations corresponding to each rotor cases are clearly shown in this table. The decrease in mean value of the efficiency is shown more clearly in the Table 2 (Fig. 2).

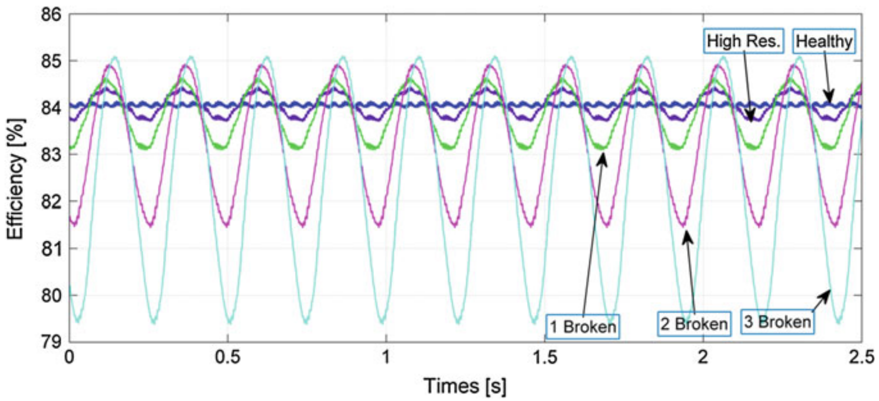


Fig. 1 The time variations of efficiency of Motor1 obtaining by simulation

Table 1 Fluctuation rate of efficiencies

	Healthy rotor (%)	High res. (%)	One broken bar (%)	Two broken bars (%)	Three broken bars (%)
Motor1	0.3	1.1	2.2	4.9	8.1
Motor2	0.2	1.5	1.9	4.4	7.3
Motor3	0.4	1.0	1.7	3.6	5.9
Motor4	0.7	1.3	2.0	3.9	6.1

Table 2 Mean value of efficiencies

	Healthy rotor (%)	High res. (%)	One broken bar (%)	Two broken bars (%)	Three broken bars (%)
Motor1	86.09	85.98	85.70	84.93	83.86
Motor2	81.84	81.80	81.69	81.14	80.29
Motor3	82.10	81.99	81.73	81.08	80.21
Motor4	81.12	81.07	80.85	80.30	79.54

4 Experimental Study and Test Results

The motors (Motor1, Motor2, Motor3 and Motor4) were tested in the motor factory by using experiment system. The purposed five different rotor faults were created in the factory at the production phase for each motor. In order to ensure accuracy of measurements, each rotor fault was created separately and passed through each assembly phase. Data were obtained in steady state operation under nominal loaded condition for each fault.

To obtain broken rotor bar faults, a small part (5 mm length) is cut from the mid side of the rotor bar and the two parts of bar are stacked from both sides. So

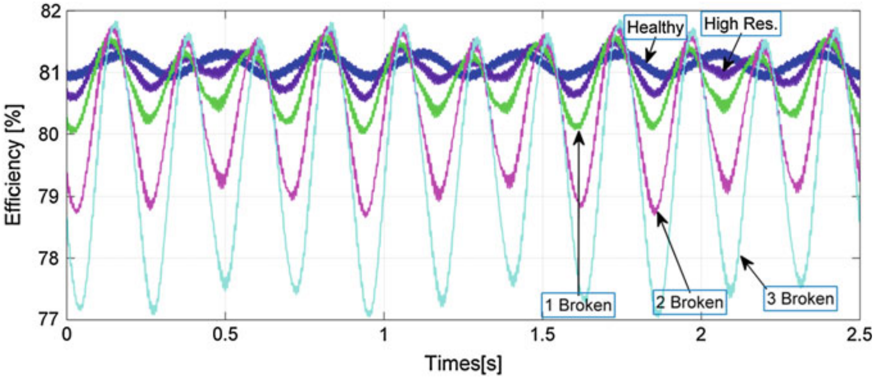


Fig. 2 The time variations of efficiency of Motor4 obtaining by simulation

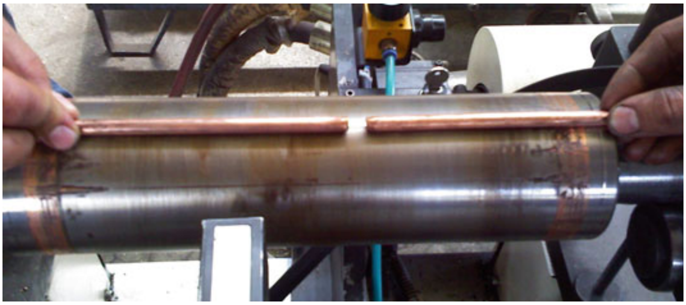
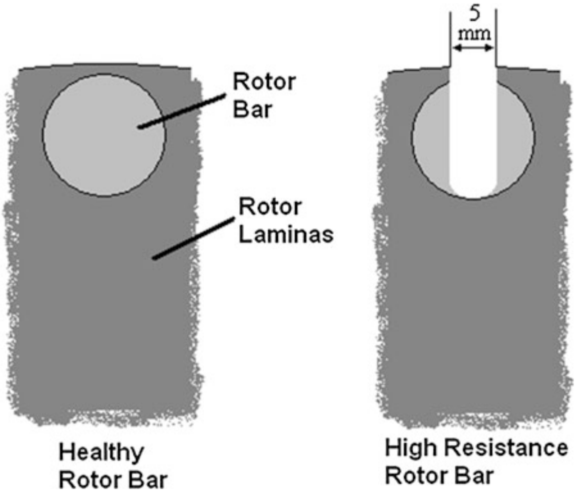


Fig. 3 Obtaining of the broken bar

Fig. 4 Obtaining of a bar with high resistance



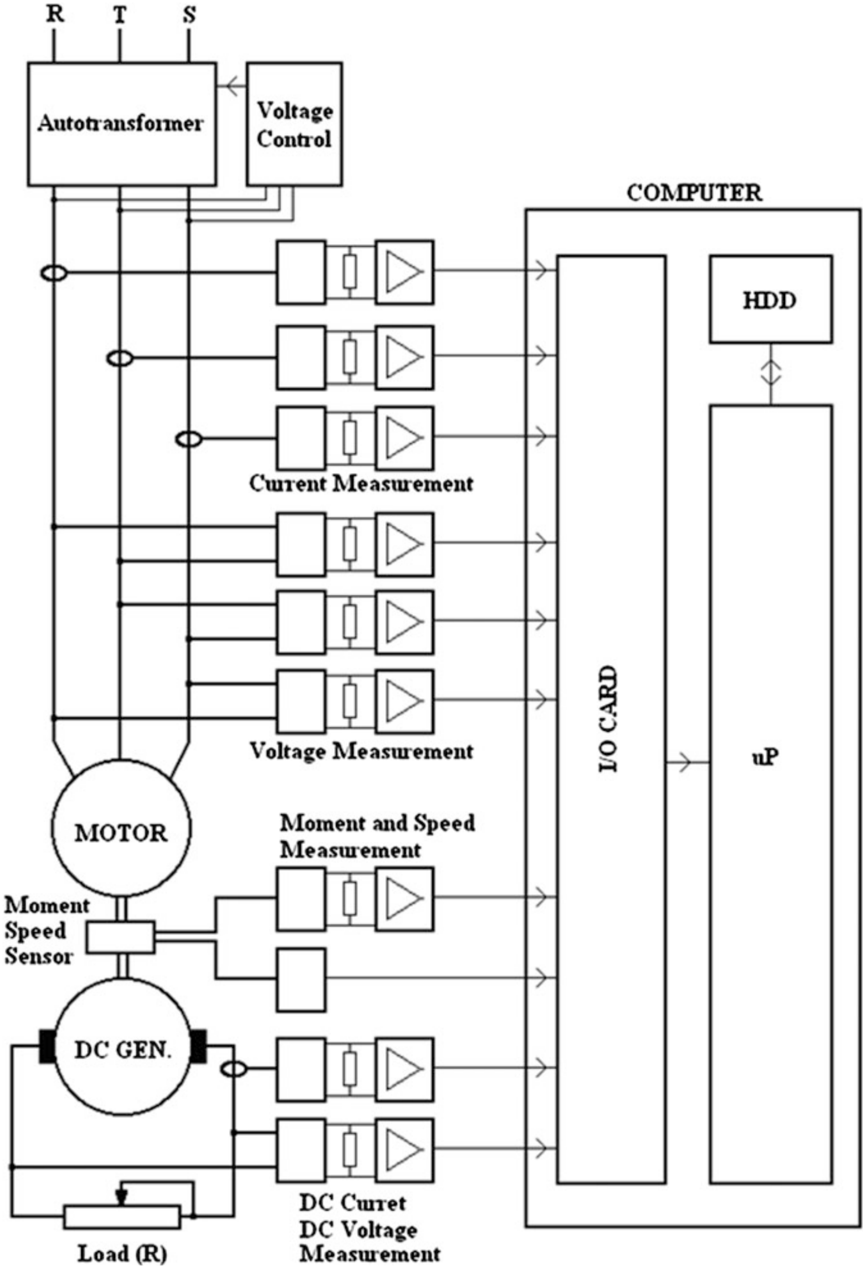


Fig. 5 Block diagram of experimental system

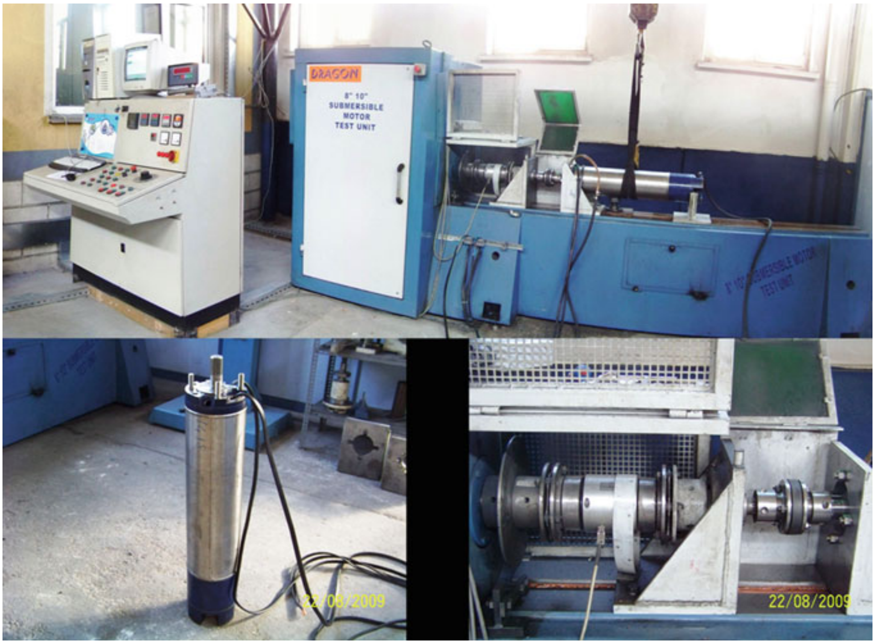


Fig. 6 Photographs of experimental system, SIMs and torque–speed sensor

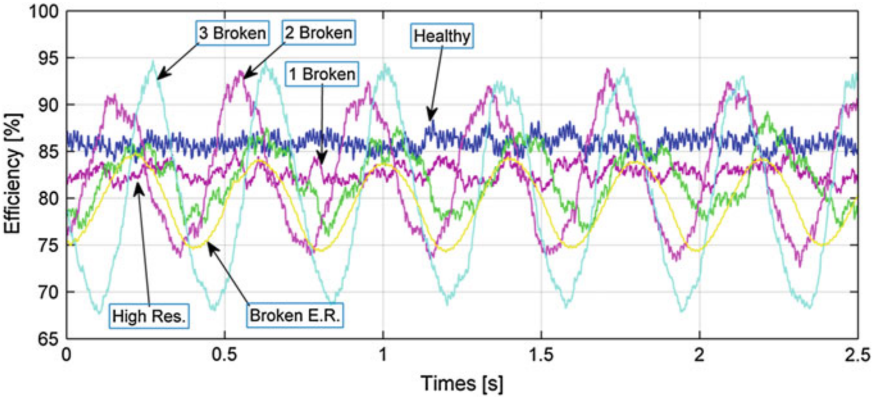


Fig. 7 The time variations of motor efficiency for Motor1

conductivity of the bar decreased to zero. Broken rotor bar photograph is given in Fig. 3. The broken end-ring fault is obtained in a similar way. A bar with a highly resistance is obtained by drilling the bar. So the conductivity of the bar decreases to 96 % from 100 % as shown Fig. 4.

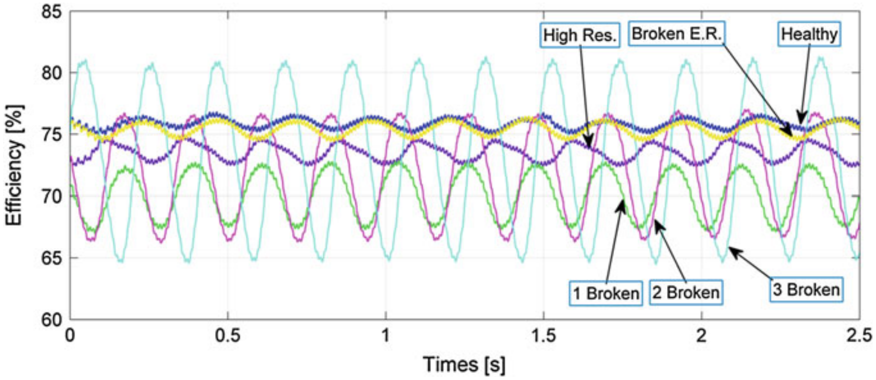


Fig. 8 The time variations of motor efficiency for Motor4

Table 3 Fluctuation rate of efficiencies

	Healthy rotor (%)	High res. (%)	1 Broken bar (%)	2 Broken bars (%)	3 Broken bars (%)	Broken end-ring (%)
Motor1	5.1	5.6	15.4	24.4	32.8	12.7
Motor2	3.0	2.2	8.1	18.8	26.3	3.9
Motor3	3.2	5.4	8.2	13.0	19.4	3.5
Motor4	5.4	4.0	9.4	17.0	24.5	4.6

Table 4 Mean value of efficiencies

	Healthy rotor (%)	High Res (%)	1 Broken bar (%)	2 Broken bars (%)	3 Broken bars (%)	Broken end-ring (%)
Motor1	85.07	82.45	82.16	82.00	78.32	80.77
Motor2	78.94	78.38	77.95	75.04	75.41	78.58
Motor3	76.37	76.08	76.48	75.78	72.48	75.48
Motor4	74.81	73.39	72.34	71.19	70.51	74.87

The motors were tested in the submersible motor factory by using test unit. The tested motor was loaded by DC generator. Loading of motor is leveled by using resistors which conducted to the generator. The block diagram of the system is given in Fig. 5 and the photographs of the used experiment system are given in Fig. 6.

The mechanical power of the motor is calculated from generator data and motor torque which are obtained from tested motors. The efficiency is estimated by using the mechanical power and input power of the motor. Figures 7 and 8 shows the efficiency changes in time under steady state operations for Motor1 and Motor4 respectively. There are considerable fluctuations in proportion of the size of the fault. The ratios of the size of these fluctuations are listed in Table 3 for every tested motor. The decrease in mean value of the efficiencies is shown in Table 4.

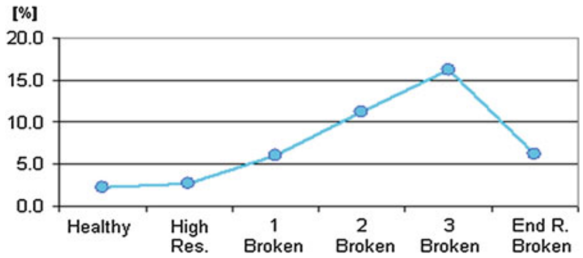


Fig. 9 The generalized fluctuation ratios in mean value of efficiencies

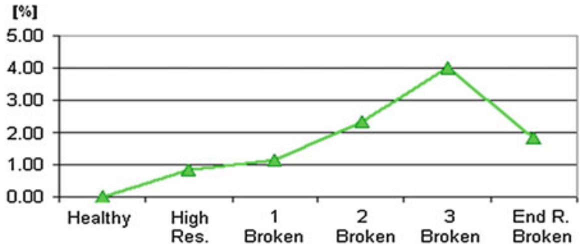


Fig. 10 The generalized decrease ratios in mean value of efficiencies

5 Conclusion and Future Work

This study presents an analysis of the effects of rotor cage faults on motor efficiency under steady state operation. For the analysis, four different SIMs are used. Experiments and simulations were performed for five different rotor faults and healthy motor conditions. Experiment and simulation results were used to analyze the efficiencies of motor. The effects on the efficiency of the rotor faults were seen as fluctuations and decrease in the mean values and compared according to size of the faults. The generalized form of these data for each motor condition is given for fluctuations in Fig. 9 and for decrease in mean value of the efficiencies in Fig. 10. It is clearly shown from these figures that the fluctuation rates increase and the mean values decrease according to the size of the rotor faults. The fluctuation of efficiency reaches up to 32 %. And its mean value is reduced by 4 %. All of these show that the effects of the rotor faults are significantly effective on motor performance. It is considered that rotor faults generally occur as small faults and grow over time. Moreover their symptoms could not be detected by simple methods. So, the motor does not stop and the fault can not be understood until the rotor fault is at an early stage or it reaches a certain size.

Losses of energy in this process are not taken into account. Especially considering irrigation systems, a lot of energy loss arises. This study clarifies these losses.

Finally, efficiency was analyzed for motor with rotor faults under steady state operations by this study.

- A notable decrease in efficiency is shown as long as fault grows.
- These effects are ranked according to size of rotor faults and analyzed the effects of such faults on motor efficiency.

References

1. A.H. Bonnett, G.C. Houkup, Cause and analysis of stator and rotor faults in three-phase squirrel-cage induction motors. IEEE Trans. Ind. Appl. **28**(4), 921–937 (1992)
2. J. Penman, A. Stavrou, Broken rotor bars their effect on the transient performance of induction machines. IEE Proc. Electr. Power Appl. **143**(6), 449–457 (1996)
3. X. Ying, Characteristic performance analysis of squirrel cage induction motor with broken bars. IEEE Trans. Magn. **45**(2), 759–766 (2009)
4. H. Arabacı, O. Bilgin, Analysis of rotor faults effects on submersible induction motor' efficiency, in *Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering and Computer Science 2013, WCECS 2013*, pp. 265–270, San Francisco, USA, 23–25 Oct 2013
5. H. Arabacı, O. Bilgin, Automatic detection and classification of rotor cage faults in squirrel cage induction motor. Neural Comput. Appl. **19**(5), 713–723 (2010)
6. M. Haji, H.A. Toliyat, Pattern recognition: a technique for induction machines rotor fault detection 'Eccentricity and broken bar fault'. IEEE Ind. Appl. Conf. Thirty-Sixth IAS Annu. Meet. **3**, 1572–1578 (2001)
7. G. Didier, E. Ternisien, O. Caspary, H. Razik, Fault detection of broken rotor bars in induction motor using a global fault index. IEEE Trans. Ind. Appl. **42**(1), 79–88 (2006)
8. M. Chow, S.O. Yee, Methodology for on-line incipient fault detection in single-phase squirrel-cage induction motors using artificial neural networks. IEEE Trans. Energy Convers. **6**(3), 536–545 (1991)
9. H. Arabaci, O. Bilgin, Neural network classification and diagnosis of broken rotor bar faults by means of short time fourier transform, in *International MultiConference of Engineers and Computer Scientists*, Hong Kong, pp. 219–223 (2009)
10. R.R. Schoen, T.G. Habetler, Effects of time-varying loads on rotor fault detection in induction machines. IEEE Trans. Ind. Appl. **31**(4), 900–906 (1995)
11. R. Casimir, E. Boutleux, G. Clerc, F. Chappuis, Broken bars detection in an induction motor by pattern recognition, in *IEEE Bologna PowerTech Conference*, Bologna, Italy (2003)
12. P.J.C. Branco, J.A. Dente, R.V. Mendes, Using immunology principles for fault detection. IEEE Trans. Industr. Electron. **50**(2), 362–373 (2003)
13. B. Yazici, G.B. Kliman, An adaptive statistical time-frequency method for detection of broken bars and bearing faults in motors using stator current. IEEE Trans. Ind. Appl. **35**(2), 442–452 (1999)
14. M.E.H. Benbouzid, G.B. Klimam, What stator current processing-based technique to use for induction motor rotor faults diagnosis? IEEE Trans. Energy Convers. **18**(2), 238–244 (2003)
15. J. Cusido, L. Romeral, J.A. Ortega, J.A. Rosero, A.G. Espinosa, Fault detection in induction machines using power spectral density in wavelet decomposition. IEEE Trans. Industr. Electron. **55**(2), 633–643 (2008)
16. L. Sun, H. Li, B. Xu, A hybrid detection method of broken rotor bars in cage induction motors, in *International Conference on Power System Technology*, Singapore, pp. 177–181 (2004)
17. B. Ayhan, M. Chow, M. Song, Multiple signature processing-based fault detection schemes for broken rotor bar in induction motors. IEEE Trans. Energy Convers. **20**(2), 336–343 (2005)
18. K. Kim, A.G. Parlos, R.M. Bharadwaj, Sensorless fault diagnosis of induction motors. IEEE Trans. Industr. Electron. **50**(5), 1038–1051 (2003)
19. A. Widodo, B.S. Yang, Support vector machine in machine condition monitoring and fault diagnosis. Mech. Syst. Signal Process. **21**(6), 2560–2574 (2007)
20. S. Günal, D.G. Ece, Ö.N. Gerek, Induction machine condition monitoring using notch-filtered motor current. Mech. Syst. Signal Process. **23**(8), 2658–2670 (2009)

21. I. Aydin, M. Karakose, And Akin E., Chaotic-based hybrid negative selection algorithm and its applications in fault and anomaly detection. Expert Syst. Appl. **37**(7), 5285–5294 (2010)
22. Y. Lei, Z. He, Y. Zi, Application of the EEMD method to rotor fault diagnosis of rotating machinery. Mech. Syst. Signal Process. **23**(4), 1327–1338 (2009)
23. M. Eltabach, J. Antoni, G. Shanina, S. Sieg-Zieba, X. Carniel, Broken rotor bars detection by a new non-invasive diagnostic procedure. Mech. Syst. Signal Process. **23**(4), 1398–1412 (2009)
24. A.Y. Ben Sasi, F. Gu, Y. Li, A.D. Ball, A validated model for the prediction of rotor bar fault in squirrel-cage motors using instantaneous angular speed. Mech. Syst. Signal Process. **20**, 1572–1589 (2006)
25. H. Su, T. Chong, Induction machine condition monitoring using neural network modeling. Trans. Industr. Electron. **54**(1), 241–249 (2007)
26. S. Nandi, R.M. Bharadwaj, H.A. Toliyat, Performance analysis of a three-phase induction motor under mixed eccentricity condition. IEEE Trans. Energy Convers. **17**(3), 392–399 (2002)
27. X. Li, Q. Wu, Performance analysis of a three-phase induction machine with inclined static eccentricity. IEEE Trans. Ind. Appl. **43**(2), 531–541 (2007)
28. X. Luo, Y. Liao, H.A. Toliyat, A. El-Antably, T.A. Lipo, Multiple couple circuit modelling of induction machines. IEEE Trans. Ind. Appl. **31**(4), 311–318 (1995)
29. P.D. Agarwal, Equivalent circuits and performance calculations of canned motors, on power apparatus and systems part-III. Trans. Am. Inst. Electr. Eng. **79**(3), 635–642 (1960)
30. M. Ito, N. Fujimoto, H. Okuda, N. Takahashi, T. Miyata, Analytical model for magnetic field analysis of induction motor performance. IEEE Trans. Power Apparatus Syst. PAS **100**(11), 4582–4590 (1981)
31. P. Vas, Steady state and transient performance of induction motors with rotor asymmetry. IEEE Trans. Power Apparatus Syst. PAS **101**(9), 3246–3251 (1982)
32. P. Baldassari, N.A. Demerdash, A combined finite element-state space modeling environment for induction motors in the ABC frame of reference: the blocked-rotor and sinusoidally energized load conditions. IEEE Trans. Energy Convers. **7**(4), 710–720 (1992)
33. J.F. Bangura, N.A. Demerdash, Diagnosis and characterization of effects of broken bars and connectors in squirrel-cage induction motors by a time-stepping coupled finite element-state space modeling approach. IEEE Trans. Energy Convers. **14**(4), 1167–1176 (1999)
34. J. Faiz, B.M. Ebrahimi, Signature analysis of electrical and mechanical signals for diagnosis of broken rotor bars in an induction motor. Electromagnetics **27**, 507–526 (2007)
35. W. Li, Y. Xie, J. Shen, Y. Luo, Finite-element analysis of field distribution and characteristic performance of squirrel-cage induction motor with broken bars. IEEE Trans. Magn. **43**(4), 1537–1540 (2007)
36. J. Faiz, B.M. Ebrahimi, Locating rotor broken bars in induction motors using finite element method. Energy Convers. Manage. **50**, 125–131 (2009)
37. K. Kurihara, T. Kubota, M. Hori, Steady-state and transient performance analysis for a single-phase capacitor-run permanent-magnet motor with skewed rotor slots. IEEE Trans. Industr. Electron. **57**(1), 44–51 (2010)
38. J. Faiz, B.M. Ebrahimi, B. Akin, H.A. Toliyat, Finite-element transient analysis of induction motors under mixed eccentricity fault. IEEE Trans. Magn. **44**(1), 66–74 (2008)
39. J. Sprooten, J.C. Maun, Influence of saturation level on the effect of broken bars in induction motors using fundamental electromagnetic laws and finite element simulations. IEEE Trans. Energy Convers. **24**(3), 557–564 (2009)
40. O.A. Mohammed, N.Y. Abed, S. Ganu, Modeling and characterization of induction motor internal faults using finite-element and discrete wavelet transforms. IEEE Trans. Magn. **42**(10), 3434–3436 (2006)
41. T.W. Preston, A.B.J. Reece, P.S. Sangha, Induction motor analysis by time-stepping techniques. IEEE Trans. Magn. **24**(1), 471–474 (1988)

42. J. Faiz, B.M. Ebrahimi, A new pattern for detecting broken rotor bars in induction motors during start-up. IEEE Trans. Magn. **44**(12), 4673–4683 (2008)
43. D.G. Dorrell, P.J. Holik, P. Lombard, H.J. Thougaard, F. Jensen, A multisliced finite-element model for induction machines incorporating interbar current. IEEE Trans. Ind. Appl. **45**(1), 131–141 (2009)

Chapter 4
Analysis of a Multilevel Inverter Topology with Reduced Number of Switches

Hulusi Karaca

Abstract Multilevel inverter is energy conversion device that is generally used in medium-voltage and high-power applications. It offers lower total harmonic distortion, switching losses and voltage stress on switches than conventional inverter. In this chapter, the topologies of the conventional multilevel inverters are discussed and a novel multilevel inverter topology with the reduced number of power switches is proposed. Also, a modulation technique for the proposed multilevel inverter is introduced. This multilevel inverter structure consists of the level module units to enhance the level of the output voltage. Since a level module unit consists of only a DC voltage source and a bidirectional switch, this structure allows a reduction of the system cost and size. Effectiveness of the proposed topology has been demonstrated by analysis and simulation.

Keywords Converter · H-bridge · Inverter · Multilevel inverter · Power converter · Reduced number of switches

1 Introduction

In recent years, a great number of industrial applications have begun to request higher power instruments [1]. However, semiconductor power switches which endure the medium voltage have not been produced still. A source in medium voltage level is inconvenient to directly connect only one switching device. Consequently, the multilevel inverter topology has emerged as a different option for working with medium voltage and high power. A multilevel inverter not only produces high power ratings, but also facilitates the use of renewable energy sources [2].

H. Karaca (✉)
Technology Faculty, Department of Electrical and Electronics Engineering, Selcuk University, 42075 Konya, Turkey
e-mail: hkaraca@selcuk.edu.tr

H. K. Kim et al. (eds.), *Transactions on Engineering Technologies*,
DOI: 10.1007/978-94-017-9115-1_4,

A multilevel inverter is power conversion device that produces an output voltage in the needed levels by using DC voltage sources applied to input [3]. Multilevel inverter which performs power conversion by using the discrete DC voltage sources was firstly introduced in 1975 [4]. This multilevel inverter structure consists of the H-bridges connected in series as shown in Fig. 1a. Then, the diode-clamped multilevel inverter was emerged [5, 6]. It employs the capacitors connected in series to separate the DC bus voltage in different levels as illustrated in Fig. 1b. In 1992, the capacitor clamped multilevel inverter was introduced [7]. This structure is similar to the structure of the diode-clamped multilevel inverter, but it uses the capacitors instead of the diodes to clamp the voltage levels as shown in Fig. 1c.

Some advantages can be obtained using multilevel inverter as follows: the output voltages and input currents with low THD, the reduced switching losses due to lower switching frequency, good electromagnetic compatibility owing to lower *dv/dt*, high voltage capability due to lower voltage stress on switches. These attractive features have encouraged researchers to undertake studies on multilevel inverter.

However, multilevel inverters have required more power components. Their driver isolations become more complicated because each extra level requires the additional isolated power source. So, the cost of the driver circuit will be increased according to the traditional single-cell inverters [8]. Recently, some multilevel inverter structures with decreased number of switches have been developed to overcome this disadvantage [3, 9, 10].

In this chapter, a novel multilevel inverter topology with reduced number of switches which uses the separate DC sources has been proposed. The control technique has been discussed to control this topology. The proposed topology consists of level module units to increase the output voltage level. The feasibility of the proposed topology has been proved by analysis and simulation [11].

2 Proposed Multilevel Inverter Topology with Reduced Number of Switches

The proposed multilevel inverter topology consists of the level module units. A level module unit is constructed by a DC voltage source and a bidirectional switch capable of conducting current and blocking voltage in both directions as shown in Fig. 2. Such devices are currently available as module in markets. Also, discrete devices can be used to construct suitable switch cells. Advantage of this switch structure is that each level module unit requires only one isolated power supply instead of two for the gate driver.

As a result, the construction cost of the proposed topology is lower than other conventional and multilevel inverters with reduced number of switches. The generalized structure of the proposed multilevel inverter is given in Fig. 3.

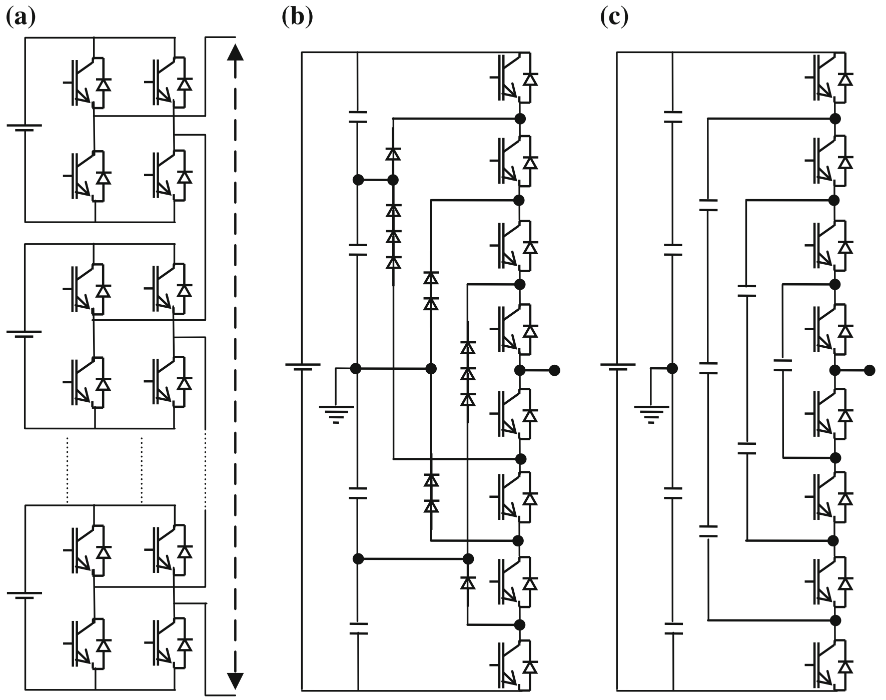
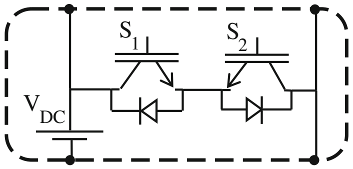


Fig. 1 Topologies of conventional multilevel inverters. **a** H-bridges. **b** Diode clamped. **c** Capacitor clamped

Fig. 2 A level module unit



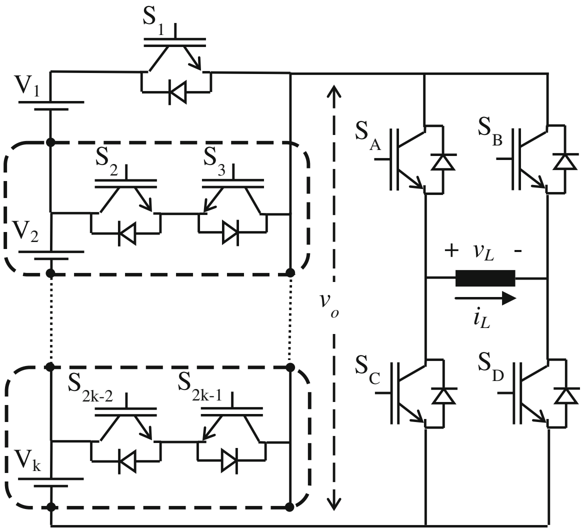
It is shown that this structure consists of two basic parts. The first is the side of level module units producing DC voltage levels.

The second is H-bridge topology, which generates both of the positive and the negative output voltages. It is clear that system can be easily expanded by adding level module units and the voltage level number in the multilevel inverter can be increased.

Considering that k is the number of discrete DC voltage sources, the maximum and minimum values of output voltage of level module units are

$$V_{o\_\max} = kV_{dc} \quad (1)$$

Fig. 3 The proposed structure of multilevel inverter



$$V_{0\_\min} = 0 \tag{2}$$

and if the power switches is turn on and turn off, $S_i(t) = 1$ and $S_i(t) = 0$ for $i = 1, 2, \ldots, (2\ k - 1)$, respectively. So, the output voltage of level module units can be expressed by

$$v_0(t) = V_1S_1(t) + [V_2S_2(t) + V_3S_4(t) + \cdots + V_kS_{2k-2}(t)] \tag{3}$$

The positive output voltage on the load is

$$v_L(t)^{+} = [V_1S_1(t) + V_2S_2(t) + \cdots + V_kS_{2k-2}(t)][S_A(t)\&S_D(t)] \tag{4}$$

and the negative output voltage on the load is

$$v_L(t)^{-} = [V_1S_1(t) + V_2S_2(t) + \cdots + V_kS_{2k-2}(t)][S_B(t)\&S_C(t)] \tag{5}$$

As a result, the overall output voltage of the proposed multilevel inverter can be given as follows:

$$v_L(t) = v_L(t)^{+} + v_L(t)^{-} \tag{6}$$

The maximum and minimum values of the generated output voltage are respectively

$$v_{L\_\max} = kV_{dc} \tag{7}$$

$$v_{L\_\min} = -kV_{dc} \tag{8}$$

If all discrete DC voltage sources V_k in Fig. 3 are equal, the number of the output voltage levels N_{level} is

$$N_{level} = 2k + 1 \tag{9}$$

The number of switches used is

$$S_{number} = 2k + 3 \tag{10}$$

This value is 4 k in the H-bridge cascaded multilevel inverter.

3 The Control Strategy of Proposed Multilevel Inverter

In this study, the number of switching angles N_α which need to be calculated has been changed according to the number of output voltage levels. This quantity is also same as the number of DC voltage sources as seen in (11). The used switching angles have been calculated by (12).

$$N_\alpha = k = \frac{N_{level} - 1}{2} \tag{11}$$

$$\alpha_i = \arcsin\left(\frac{2j - 1}{N_{level} - 1}\right) \qquad j = 1, 2, \ldots, k \tag{12}$$

In Fig. 4, the angles required are illustrated to obtain a full period of the output voltage in 5-level multilevel inverter. It is clear that the other angles can be easily derived from α_1 and α_2.

The states of switches which cannot be on simultaneously are given in Table 1 in order to avoid a short circuit as understood from Fig. 3.

There are ten different switching states to obtain a full period of the output voltage in 5-level multilevel inverter. The used switching state quantities can be expressed by $N_{level} + 5$ according to the number of output levels or by $2\,k + 6$ according to the number of DC voltage sources. All states of switches are presented for 5-level multilevel inverter between Figs. 5 and 14.

The output voltage is 0 for states in Figs. 5 and 6. While the direction of load current is positive, the current flows through D_B and S_A. While it is negative, it flows through S_C and D_D.

Fig. 4 The output waveforms of 5-level

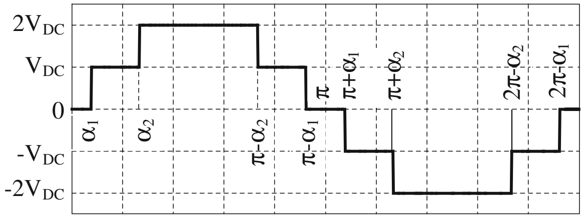


Table 1 Inconvenient switch states in the proposed multilevel inverter

Switches cannot be on simultaneously						
1	S_A	S_C				
2	S_B	S_D				
3	S_1	S_3	S_5	S_7	…	S_{2k-1}

Fig. 5 $v_L = 0$, $i_L = +$

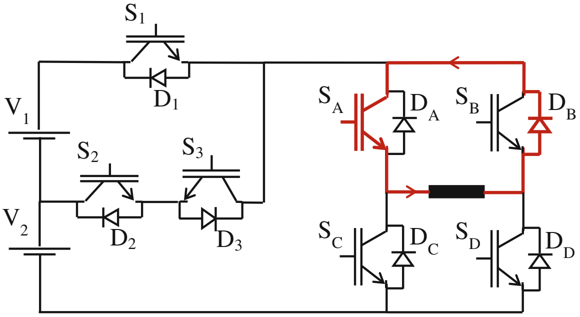
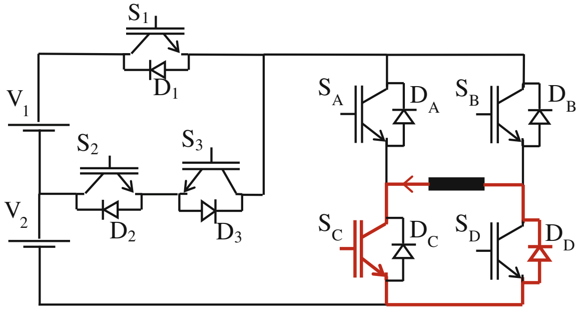


Fig. 6 $v_L = 0$, $i_L = -$



The load voltage for states in Figs. 7 and 8 is $+V_{DC}$. If the direction of load current is positive, it is used the state in Fig. 7. If it is negative, which this is the regenerative mode, it is used the state in Fig. 8.

The voltage for states in Figs. 9 and 10 is $-V_{DC}$. If the direction of current is positive, the current flows through D_B, S_3, D_2, V_2 and D_C, where the energy

Fig. 7 $v_{\mathrm{L}} = V_2,\ i_L = +$

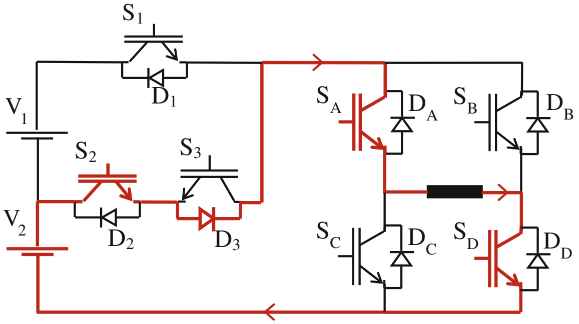


Fig. 8 $v_{\mathrm{L}} = V_2,\ i_L = -$

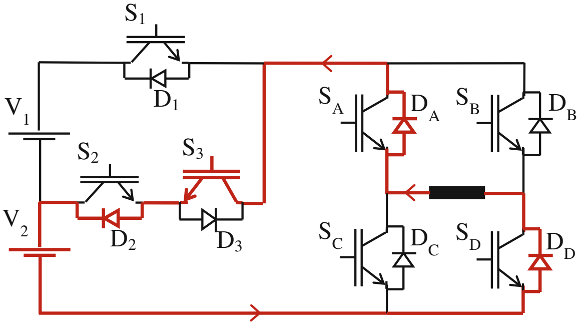
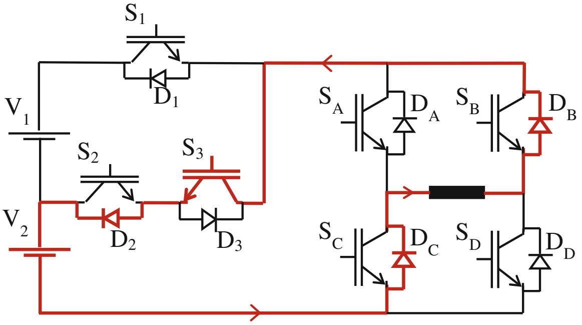


Fig. 9 $v_{\mathrm{L}} = -V_2,\ i_L = +$



regeneration has been done to source as shown in Fig. 9. If the direction of current is negative, the current flows through S_2, D_3, S_{B} and S_{C} as shown in Fig. 10.

In Figs. 11 and 12, the load voltage is +($V_1 + V_2$). The direction of current is positive in Fig. 11 and it flows from S_1, S_{A} and S_{D}. The direction of current is negative in Fig. 12 and it flows from D_{A}, D_1, V_1, V_2 and D_{D}.

The voltage for states in Figs. 13 and 14 is $-(V_1 + V_2)$ The direction of load current is positive in Fig. 13 and it flows from the load to the source (regenerative

Fig. 10 $v_{\mathrm{L}} = -V_2,\ i_L = -$

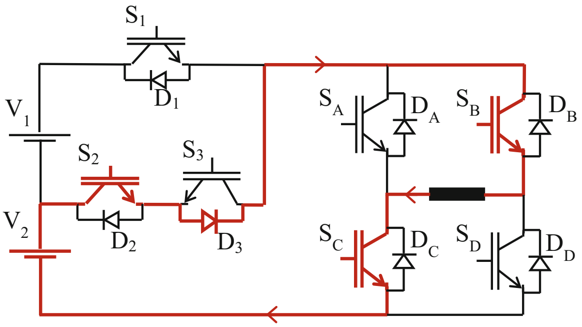


Fig. 11 $v_{\mathrm{L}} = V_1 + V_2,\ i_L = +$

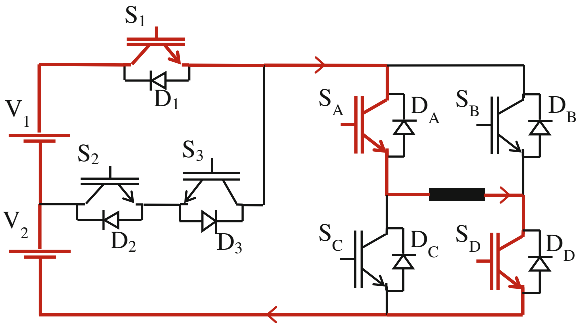
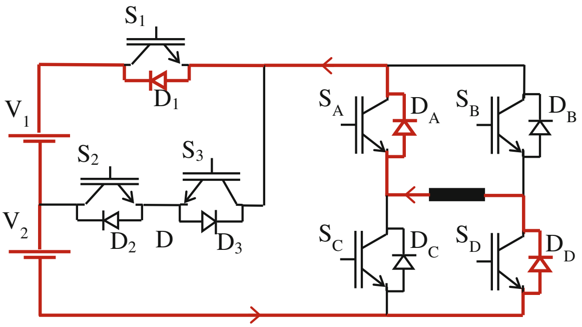


Fig. 12 $v_{\mathrm{L}} = V_1 + V_2,\ i_L = -$



mode). The direction of load current is negative in Fig. 14. All working states of power switches of the proposed 5-level multilevel inverter are summarized in Table 2.

Fig. 13 $v_{\mathrm{L}} = -(V_1 + V_2),\ i_L = +$

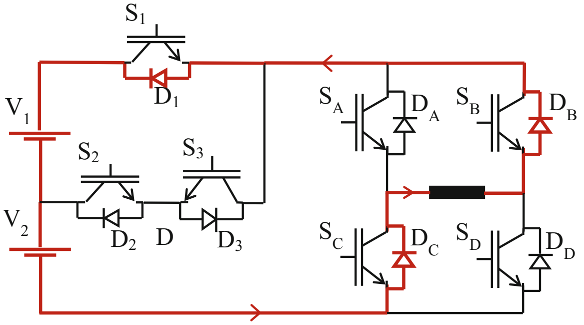
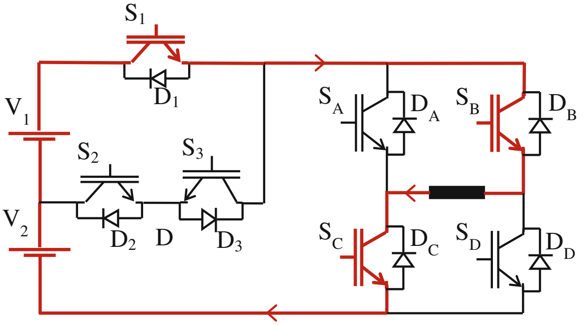


Fig. 14 $v_{\mathrm{L}} = -(V_1 + V_2),\ i_L = -$



4 Results

Since the aim of this study is to introduce a novel multilevel inverter, it has not been mentioned about the improving of total harmonic distortion. Some simulation results for 5-level and 7-level inverters have been given to evaluate the performance of the proposed novel multilevel inverter in the synthesis of a requested load voltage waveform.

In this study, a 5-level inverter and a 7-level inverter can generate staircase voltage waveform respectively with maximum 200 and 300 V, due to the values of the used DC voltage sources are 100 V. The simulated multilevel inverters have been loaded by two different loads. The previous load is a pure resistive; its value is 5 Ω. The next one is a series connected resistance and inductance; their values are 1 Ω and 15 mH. In this study, the fundamental frequency switching pattern has been utilized. This switching pattern causes less switching losses from other patterns due to its low switching frequency [11–16].

In Fig. 15, the gate signals of all switches in 5-level multilevel inverter are clearly shown. While the staircase output voltage is synthesized, the power flow in this switching structure is bidirectional, both from source to load and from inductive load to source.

Table 2 The output voltage and current values for state of power switches in the proposed 5-level multilevel inverter

	States of power switches							Voltage current	
	S_1	S_2	S_3	S_A	S_B	S_C	S_D	V_L	i_L
1	off	off	off	on	off	off	off	0	+
2	off	off	off	off	off	on	off	0	–
3	off	on	off	on	off	off	on	V_1	+
4	off	off	on	off	off	off	off	V_1	–
5	off	off	on	off	off	off	off	$-V_1$	+
6	off	on	off	off	on	on	off	$-V_1$	–
7	on	off	off	on	off	off	on	$V_1 + V_2$	+
8	off	off	off	off	off	off	off	$V_1 + V_2$	–
9	off	off	off	off	off	off	off	$-(V_1 + V_2)$	+
10	on	off	off	off	on	on	off	$-(V_1 + V_2)$	–

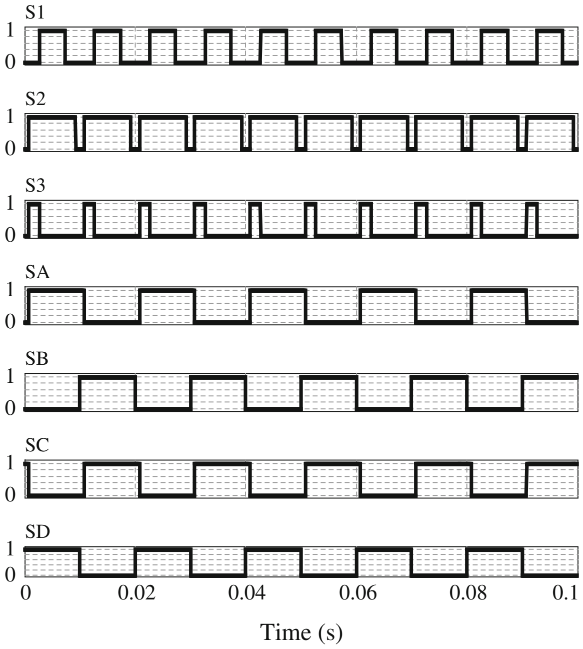


Fig. 15 Gate signals of all switches in 5-level multilevel inverter

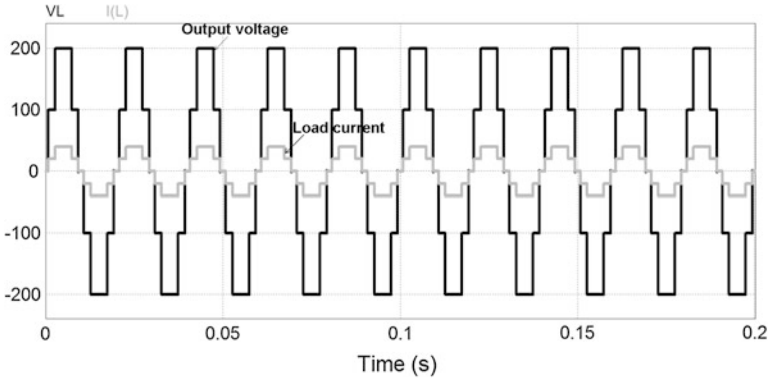


Fig. 16 The output voltage and current of 5-level multilevel inverter for pure resistive load

Figure 16 shows the voltage and current of 5-level multilevel inverter for pure resistive load. It is shown that the waveform of the load current is the stepped. Also, it has a unity power factor. The voltage and current of 5-level multilevel

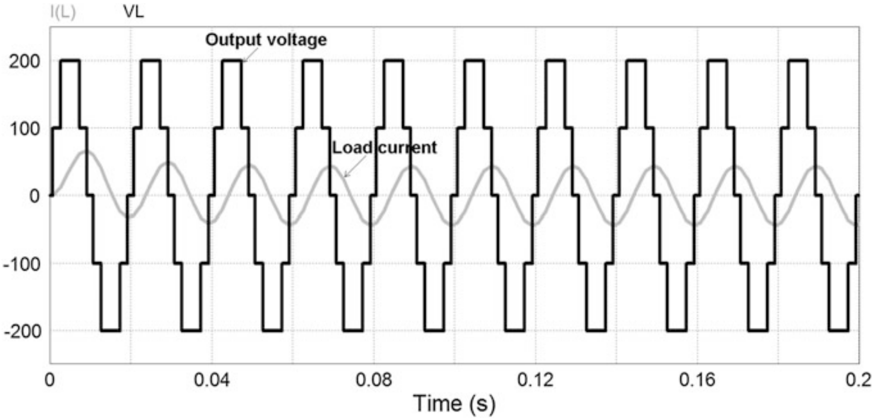


Fig. 17 The output voltage and current of 5-level multilevel inverter for series R-L load

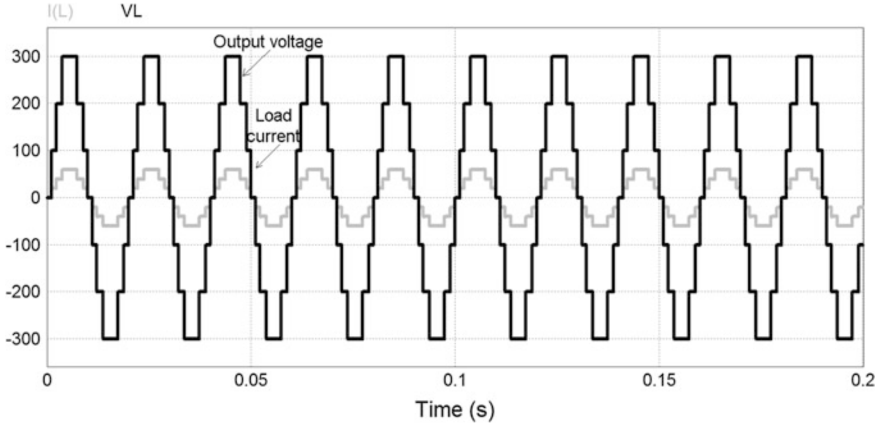


Fig. 18 The output voltage and current of 7-level multilevel inverter for pure resistive load

inverter for R-L load is illustrated in Fig. 17. The output voltage is same as in Fig. 16, but the output current has sinusoidal waveform with a fundamental frequency due to inductive load. Also, the power factor is not unity.

However, the number of the output voltage levels should be enhanced to obtain the load voltage and current with lower harmonic contents. The output voltage and current waveforms of 7-level multilevel inverter are given for pure resistive and R-L loads in Figs. 18 and 19, respectively.

It is understood that the more the number of levels are increased, the more the output waveforms are similar to the pure sinusoidal. The number of output levels in the proposed multilevel inverter can be easily increased by connecting the level module units.

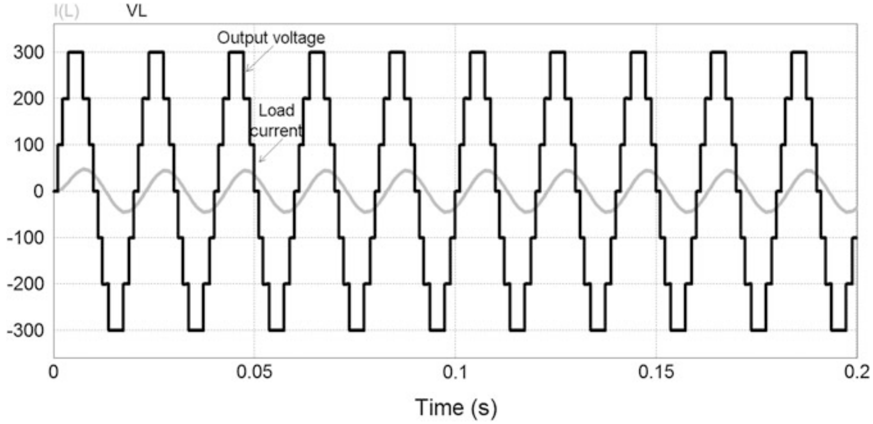


Fig. 19 The output voltage and current of 7-level multilevel inverter for series R-L load

5 Conclusion and Future Work

A novel structure of multilevel inverter which needs the minimum number of switches has been proposed. The modulation technique and calculation of conducting angles for the proposed topology has been introduced. The additional level module units have been use to increase the output levels. Since the devices in the level module units are common emitter back to back, each unit has been required only one isolated power supply for driver. Therefore, construction cost of the proposed multilevel inverter is lower and it is not bulky. The operation and performance of the proposed structure has been proved by simulations of 5-level and 7-level multilevel inverter.

Acknowledgment This work was supported by Scientific Research Projects Coordinating Office of Selcuk University (SÜBAP Project Number: 13701399).

References

1. J. Rodriguez, J.S. Lai, F.Z. Peng, Multilevel inverters: survey of topologies, controls, and applications. IEEE Trans. Ind. Appl. **49**(4), 724–738 (2002)
2. S. Khomfoi, L.M. Tolbert, Power Electronics Handbook, 2nd Edn, Chapter 31—Multilevel Power Converters (Elsevier, Amsterdam, 2007), pp. 31/1–31/50
3. E. Babaei, S.H. Hosseini, New cascaded multilevel inverter topology with minimum number of switches. Energy Convers. Manag. **50**, 2761–2767 (2009)
4. R.H. Baker, L.H. Bannister, Electric power converter. U.S. Patent 3 867 643, Feb 1975
5. R.H. Baker, High-voltage converter circuit, U.S. Patent 04-203-151, May 1980
6. A. Nabae, I. Takahashi, H. Akagi, A new neutral-point clamped PWM inverter, in *Proceedings of the Industry Application Society Conference*, pp. 761–766, 1980

7. T.A. Meynard, H. Foch, Multi-level conversion: high voltage choppers and voltage source inverters, in *Proceedings of the IEEE Power Electronics Specialist Conference*, vol. 1, pp. 397–403, 1992
8. F.J.T. Filho, Real-time selective harmonic minimization for multilevel inverters using genetic algorithm and artificial neural network angle generation. Ph.D. dissertation, University of Tennessee, 2012
9. E. Beser, B. Arifoglu, S. Camur, E.K. Beser, Design and application of a single phase multilevel inverter suitable for using as a voltage harmonic source. J. Power Electron. **10**(2), 138–145 (2010)
10. R.A. Ahmed, S. Mekhilef, H.W. Ping, New multilevel inverter topology with reduced number of switches, in *Proceedings of the 14th International Middle East Power Systems Conference*, pp. 565–570, 2010
11. H. Karaca, A novel topology for multilevel inverter with reduced number of switches, lecture notes in engineering and computer science, in *Proceedings of the World Congress on Engineering and Computer Science 2013, WCECS 2013*, San Francisco, USA, pp. 350–354, 23–25 Oct 2013
12. K.A. Corzine, M.W. Wielebski, F.Z. Peng, J. Wang, Control of cascaded multi-level inverters. IEEE Trans. Power Electron. **19**(3), 732–773 (2004)
13. Z. Du, L.M. Tolbert, J.N. Chiasson, Active harmonic elimination for multilevel converters. IEEE Trans. Power Electron. **21**(2), 459–469 (2006)
14. M.S.A. Dahidah, V.G. Agelidis, M.V.Rao, Hybrid genetic algorithm approach for selective harmonic control. Energy Convers. Manag. **49**, 131–142 (2008)
15. J.H. Seo, C.H. Choi, D.S. Hyun, A new simplified space—vector pwm method for three-level inverters. IEEE Trans. Power Electron. **16**(4), 545–550 (2001)
16. B.P. McGrath, D.G. Holmes, Multicarrier PWM strategies for multilevel inverters. IEEE Trans. Ind. Electron. **49**(4), 858–867 (2002)

Chapter 5
Comparison of PWM Techniques for Back to Back Converters in PSCAD

Agustina Hernandez, Ruben Tapia, Omar Aguilar and Abel Garcia

Abstract This article presents the simulation and comparison between the space vector pulse width modulation and sinusoidal pulse width modulation techniques for back to back converters with a decoupling control strategy, PSCAD/EMTDC for simulation purpose is used. Also, a study of steady state and transient performance characteristics of the system is carried out for both techniques. The simulation results show that the transient response is similar for both schemes, and the SVPWM technique has the advantage less harmonic content, which it is useful in applications that require a low harmonic level for avoiding overheats and malfunction in sensitive systems.

Keywords Back to back · Harmonics · Power electronics · PSCAD simulator · PWM technique · SPWM technique

A. Hernandez (✉) · R. Tapia · O. Aguilar
Department of Engineering, Polytechnic University of Tulancingo,
CO 43629 Tulancingo, Hidalgo, Mexico
e-mail: agustina.hernandez@upt.edu.mx

R. Tapia
e-mail: ruben.tapia@upt.edu.mx

O. Aguilar
e-mail: omar.aguilar@upt.edu.mx

A. Garcia
Department of Mechatronics, Polytechnic University of Pachuca,
CO 43830 Zempoala, Hidalgo, Mexico
e-mail: abel@upp.edu.mx

H. K. Kim et al. (eds.), *Transactions on Engineering Technologies*,
DOI: 10.1007/978-94-017-9115-1_5,

1 Introduction

Power many devices based on power electronics are used since finer and ever more intelligent controls are needed to improve the electrical grid performance in different voltage levels [1, 2]. They are essential for applications in emerging generating systems derived on renewable resources. It must ensure a robustness performance when are applied in isolated operation or when are connected with the conventional grid [3]. Their use is increasing in the industry for machines regulation. Furthermore, consumer market applications such as remote generating systems and quality energy applications result in a growing topic for analysis and improvement [4]. Electrical power generation with wind and solar as source of power may be considered of the most promising renewable energy sources, but the control schemes are an important issue [5, 6]. There are extensive frameworks of various types of electrical machines and controls algorithms that have been developed for wind generator applications as are presented in [7–9]. Among the schemes of wind energy conversion systems (WECS), the induction machine and permanent magnet synchronous machine are seen as promising elements for energy conversion. In all cases back to back (BTB) arrangement based on power electronics is required to take advantage of the benefits [9–11]. BTB converter is formed by two identical voltage source converters (VSC) connected by a common dc-link. This topology presents several advantages in terms of power processing and allows bidirectional power flow with quasi-sinusoidal currents. The load can be active, passive or even another network, in such case achieve a unity power factor is possible if it is required. The dc-link in the middle provides decoupling between both converters; as a result, they could be driven independently. Therefore, it is possible to have a fast and independent control of active and reactive power for both converters and improve the system operation [12]. To attain simultaneously these benefits is important to explore control strategies, which allow obtaining the desired regulation.

These devices have as their main component a voltage source converter, which uses pulse width modulation (PWM) for controlling purposes. The switching elements are power semiconductors that operate at high frequencies. There are several techniques for controlling semiconductors' converter; each one has its own advantages and disadvantages. The sinusoidal PWM (SPWM) is matured technology. In this method a triangular wave is compared to a sinusoidal wave. The space vector PWM (SVPWM) has been increasingly used in last decade. This modulation, instead of using a separate modulator for each of the three phases, the complex reference voltage vector is processed as a whole. An aptitude for easy digital implementation is the notable feature of space vector modulation, which can be easily implemented in digital signal processor (DSP) [13]. Moreover, suitable control scheme is necessary to deal with different operating conditions and uncertainties in the network due to the parameters or load variations.

This article presents the performance comparison of a back to back converter connected to a stand-alone load, where the SPWM and SVPWM techniques are applied. The two schemes characteristics and the implemented methodology in PSCAD are presented. The simulation results show some differences in the behavior of both techniques. The system under study is a back to back converter and is subject to different operating conditions and load variations. This work addresses two main issues: (a) first, it presents a detail implementation procedure of PWM techniques in industrial software and; (b) it realizes a behavior comparison for the analysis system.

2 AC-DC-AC Converter Model

The back to back converter shown in Fig. 1 is formed by two shared VSC with a common DC bus. Both converters can operate as a rectifier or inverter depending on the power flow direction and the operation is complementary. The source side converter is designated as VSC1 and the converter connected to the load side as VSC2. This topology presents several advantages in terms of power processing and allows bidirectional power flow with quasi-sinusoidal currents.

The control objectives which are set for the BTB operation depend on the application, for example: (1) AC voltage; (2) AC frequency; (3) active power; (4) reactive power; and (5) DC voltage regulation. Two control tasks are assigned to each VSC and how they allocate can be arbitrary. Figure 1 shows the BTB converter structure.

2.1 Source Side Converter Control

The source side converter objective is to keep the dc-link voltage constant. The vector control method is used with a reference frame oriented along vector position, enabling independent control of the active and reactive power flowing between the source and the converter [14]. The PWM signals are obtained by current regulated scheme, with d-axis current used to manipulate the dc-link voltage and the q-axis current component to regulate the reactive power. The reference values for the grid-side converter V_{1d}^{*} and V_{1q}^{*} are established with a PI controller, respectively. The Eqs. (1) and (2) show the development control scheme [15, 16].

$$\mathbf{V}_{1d}^{*} = -V_{1d}' + (\omega_e L_1 i_{1q} + V_{1d}) \tag{1}$$

$$\mathbf{V}_{1q}^{*} = -V_{1q}' - (\omega_e L_1 i_{1d}) \tag{2}$$

where $\acute{V}_{1d}$ y $\acute{V}_{1q}$ are reference voltage for the d-axis and q-axis, respectively; ω_e is the AC frequency in the grid; L_1 is inductance of source side; i_{1d} and i_{1q} are currents of the d-axis and q-axis; V_{1d} is voltage the source in the $dq0$ frame.

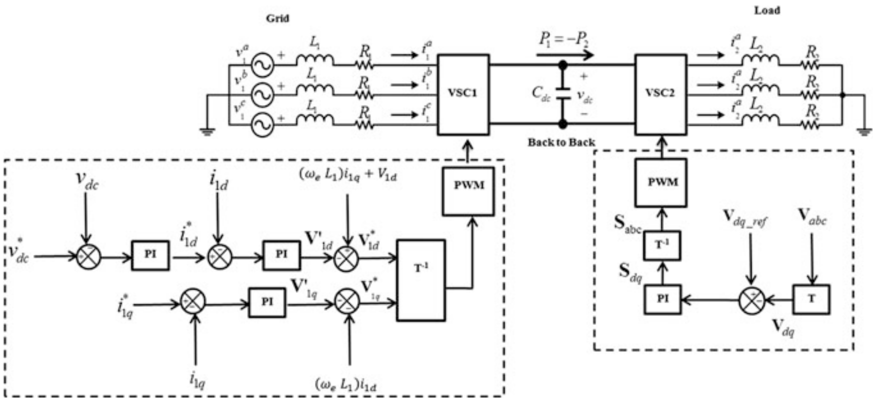


Fig. 1 Back to back converter schematic diagram

2.2 Load Side Converter Control

The control targets can be choice and drivers depending on the application. For example, in distributed generation systems it is usual that the control block includes a scheme for regulating the AC frequency and voltage. Therefore, in this article we are controlling the AC voltage and frequency values; also DC link voltage regulation is required. The VSC2 must keep constant AC frequency and voltage values for different load variables. In this converter the control problem is to determine the reference signal in the PWM control scheme to allow the load voltage tracking the reference. The carrier signal is fixed in 5 kHz for 60 Hz load. Hence, the needed measurement is the three-phase voltage at the connection point. For control purpose a *dq0* conversion is employed and its expression is the following,

$$\mathbf{V_{dq0}} = \mathbf{T}\mathbf{V_{abc}} \tag{3}$$

where **T** is the transformation matrix.

$$\mathbf{T} = \begin{bmatrix} \sin(\omega t) & \sin\left(\omega t - \frac{2\pi}{3}\right) & \sin\left(\omega t + \frac{2\pi}{3}\right) \\ \cos(\omega t) & \cos\left(\omega t - \frac{2\pi}{3}\right) & \cos\left(\omega t - \frac{2\pi}{3}\right) \\ \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \end{bmatrix} \tag{4}$$

The angular velocity is acquired by phase lock loop (PLL) strategy and the error signal employed for obtaining the PWM reference signal takes the form,

$$\mathbf{e}(\mathbf{t}) = \begin{bmatrix} V_{d\_ref} \\ V_{q\_ref} \end{bmatrix} - \begin{bmatrix} V_d \\ V_q \end{bmatrix} \tag{5}$$

Defining the reference voltages, there are two separate axes 90 electrical degrees. *d*-axis is in phase with the reference vector, whereas *q*-axis is delayed 90° respect to voltage, similarly, into a *d*-*q* component the current is decomposed. Therefore, $v_{\text{dref}} = 300$ V and $v_{\text{qref}} = 0$ V. For obtaining the PWM reference signal in *dq* frame a PI controller is included, after that, an inverse transform is applied from *dq* to *abc*. To obtain a PWM signal for VSC2 control a comparison between both techniques SPWM and SVPWM is carried out.

3 Programming PWM Techniques in PSCAD

3.1 SPWM Implementation

Within the programming environment, the SPWM technique is developed using some blocks that have already been designed in master library. The configuration of each block is based on the mathematical model. In Fig. 2 the triangular wave (carrier) is compared with each phase sinusoidal voltage, where the output variables represent the pulses to switching the power electronic elements. The carrier signal operates at 5 kHz and is compared with the modulated signals at 60 Hz [16].

3.2 SVPWM Implementation

The SVPWM technique in PSCAD/EMTDC is carried out by five user-defined sub-systems, (A)–(E). These are shown in Fig. 3. In sub-system (A): at first the magnitude (Vmag) and phase angle (Theta) are determined from the reference three-phase voltage. To find out the sector, the range of calculated angle value must be between 0 and 2π. The Eqs. (6) and (7) are programming in this block.

$$\mathbf{V}_S^* = \left|V_S^*\right| e^{j\theta_S^*} \tag{6}$$

$$\begin{aligned} V_S^* &= \sqrt{V_{S\alpha}^{*2} + V_{S\beta}^{*2}} \\ \theta_S^* &= \tan^{-1}\left(\frac{V_{S\beta}^*}{V_{S\alpha}^*}\right) \end{aligned} \tag{7}$$

where V_s^* is the magnitude and θ_s^* is the phase angle; (B): this sub-system is designed to calculate the turn-on time of two adjacent active vectors and the zero state vectors; sub-system (C): the transition time of active and zero vectors are calculated; (D): in this block the duty cycles generation procedure is established; finally, sub-system (E): the switching pulses for power electronic devices are generated. The technique was taken from [17].

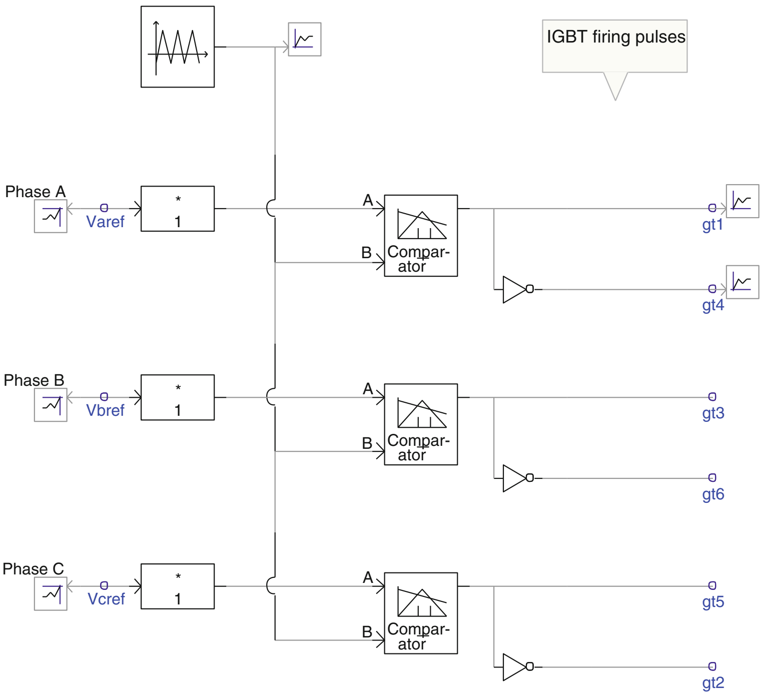


Fig. 2 SPWM pulse generator scheme in PSCAD

4 Simulation Results

In order to verify the control scheme applicability for the back to back converter the PSCAD software is used. To analyze the results two cases are presented. Case A and B are described below. The system data are: the voltage source is 380 V operating at 60 Hz and the load to 25KW, 300 V. We can see [18].

4.1 Case 1

In this case, initially the systems is inactive; all variables have a value equal to zero, after a transient period they are stabilized around of a constant value at steady state. Through the feedback control strategy the DC voltage achieves, $V_{dcref} = 700$ V. Figure 4 shows the DC voltage performances when the SPWM and SVPWM control strategies are applied. These techniques implemented by software determine the sequence and duration of the time on and off the switches. One can see that the steady state and transient responses have similar characteristics for both

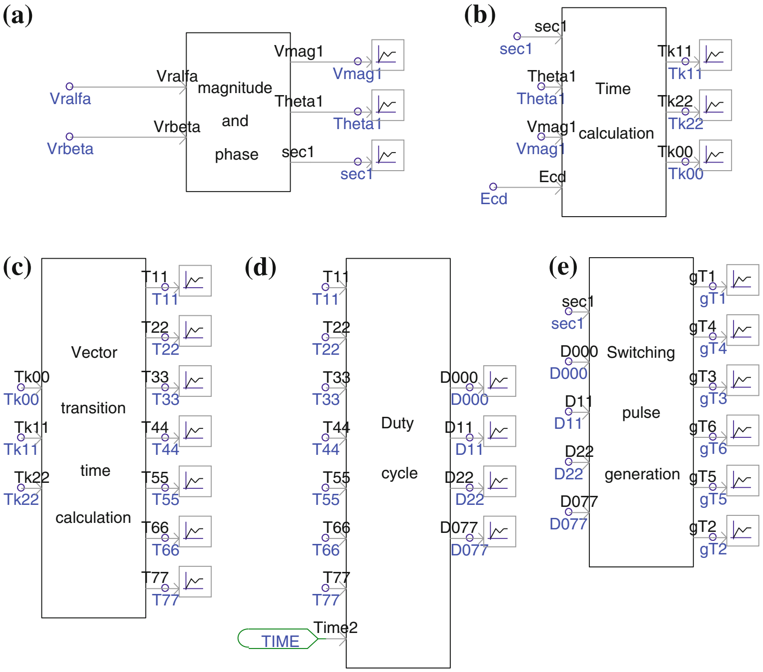


Fig. 3 SVPWM pulse generator scheme in PSCAD

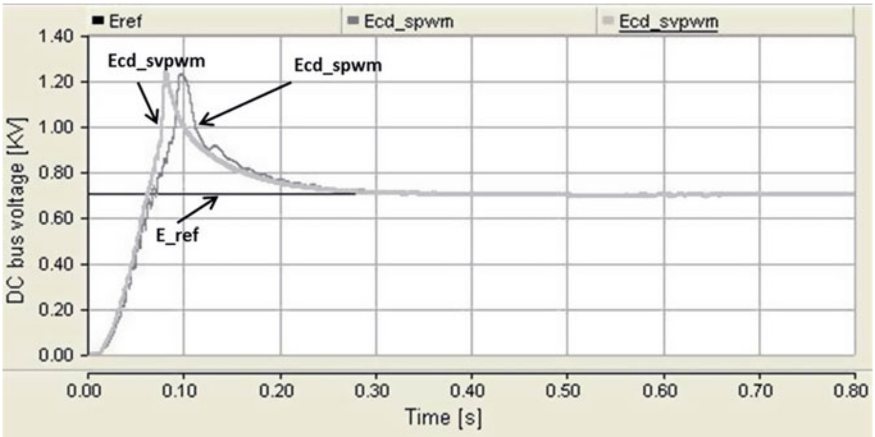


Fig. 4 DC bus voltage evolution

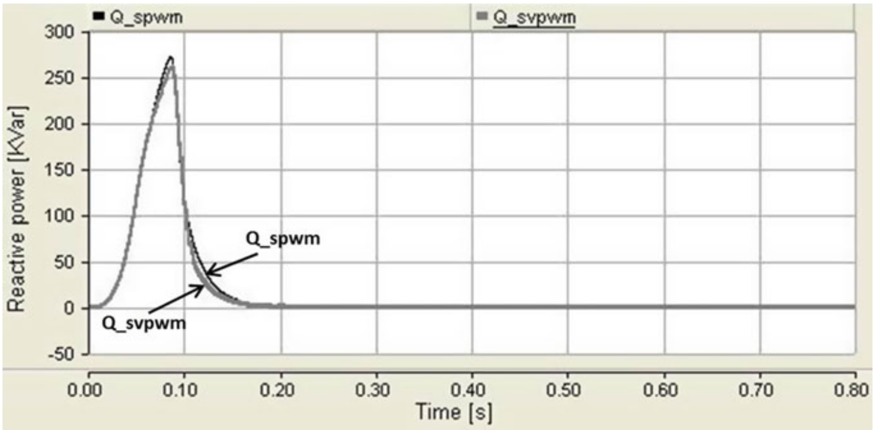


Fig. 5 Reactive power output in the source side converter

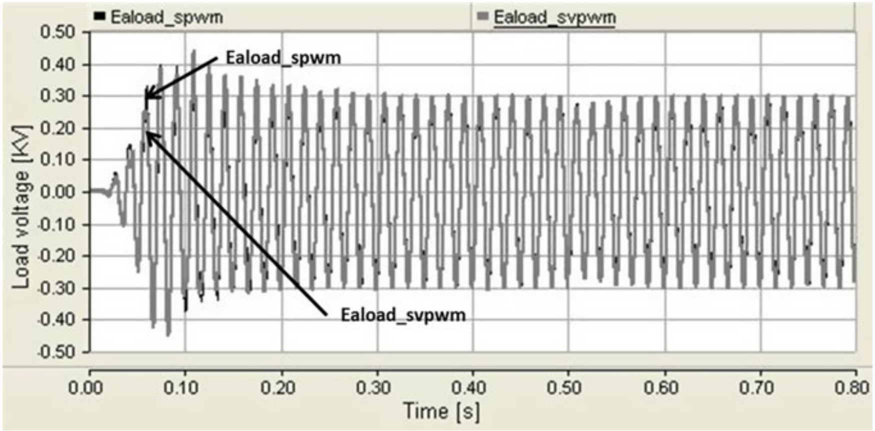


Fig. 6 Phase voltage output in load side converter

algorithms. The control strategy also allows regulating the power factor, which must remain closed to one, thus the reactive power is equal to zero. Figure 5 exhibits that the source side converter is operating at unity power factor, since the reactive power tracks the reference zero for both modulation strategies.

The control of the load side converter regulates the AC voltage and frequency. Figure 6 presents the voltage performance in phase A, for both algorithms. The modulation techniques show a fast evolution, allowing that the converters are operated properly and the interest variables reach the desired steady state value with fast transient response.

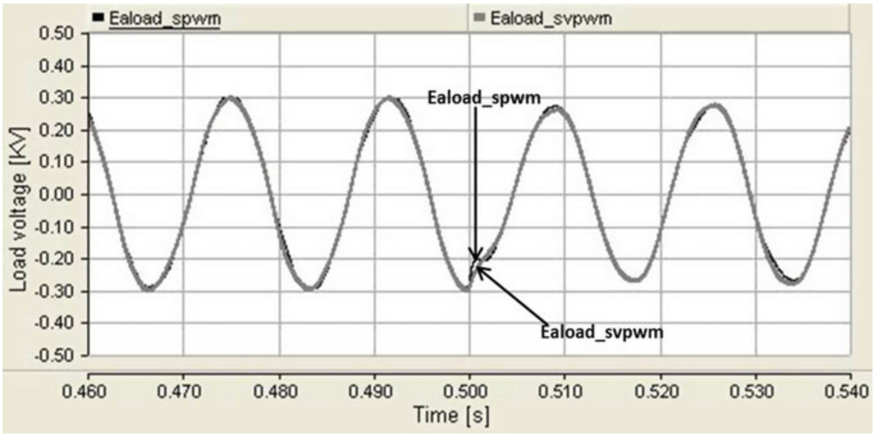


Fig. 7 Load voltage when the load is increased

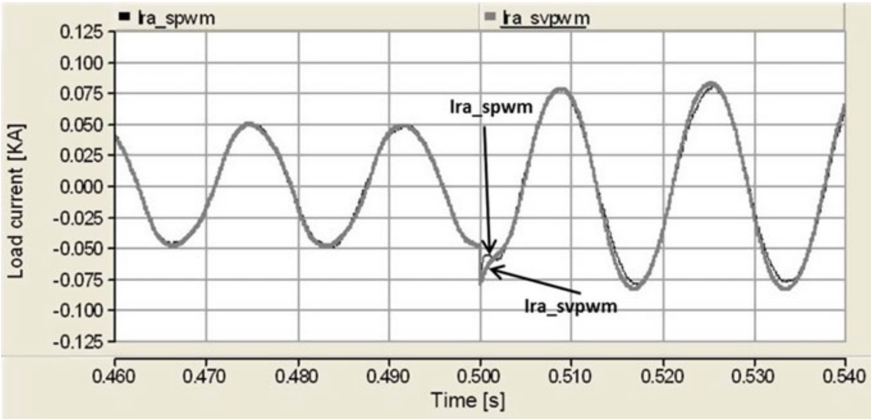


Fig. 8 Load current when the load is increased

4.2 Case 2

The second case illustrates the system's evolution when the load is increasing from 15 to 25 KW, at t = 0.5 s. One can see that the transient period is short and quickly response is exhibited before the steady state condition is achieved. The load voltage attains the reference value at 300 V, before and after the load changed, Fig. 7. By increasing the load value, the current takes a new value and it is established around 90A, Fig. 8. It is clear that the transient response is similar for both schemes and the settling time is same for two cases. The modulation

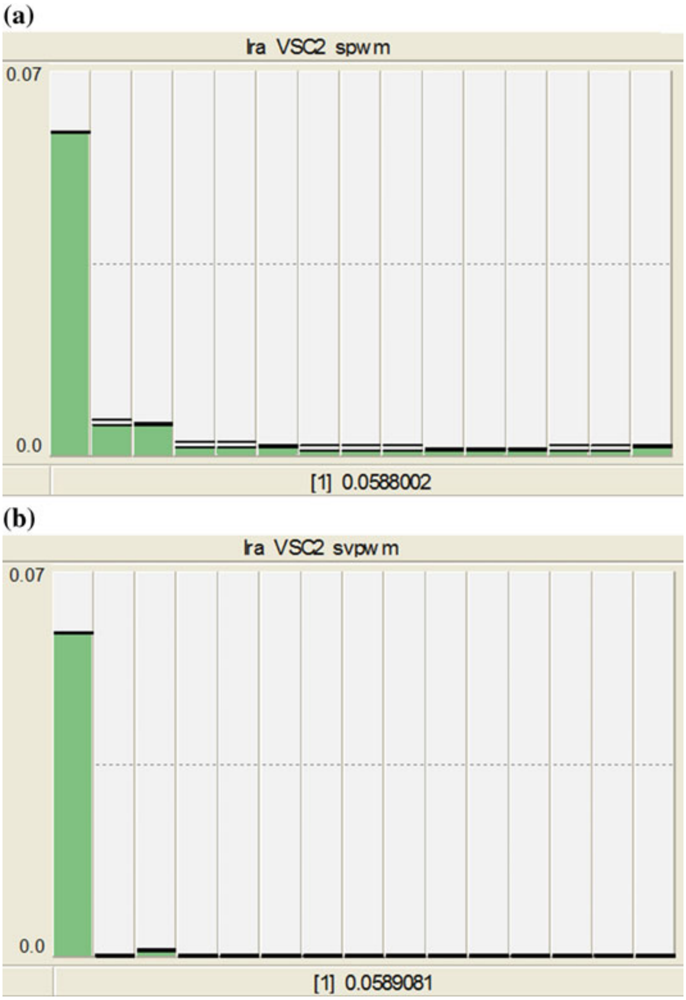


Fig. 9 FFT analysis of phase current A: **a** SPWM; **b** SVPWM

techniques display very well performance allowing that the converters operate properly and the interest variables reach the desired steady state value.

The waveform distortion according to FFT analysis in Fig. 9 is exhibited. Using the SVPWM scheme the harmonic current components of phase A is less that with the SPWM technique. In general, drive systems with low harmonic content are better than that with high content. The two cases validate the appropriate system evolution under perturbations. The simulations of back to back structure under different operating conditions demonstrate the effectiveness and robustness of PWM techniques. Transient and steady state response were analyzed.

5 Conclusion

This article developed the simulation and comparison of space vector pulse width modulation and sinusoidal pulse width modulation for back to back converters with a decoupling control strategy in PSCAD/EMTDC. Both steady state and transient system performance are presented. The simulation results show the control scheme effectiveness. The SPWM is simpler to implement than SVPWM in the software. Very similar results are presented with both control algorithms. However, the most obvious difference is displayed in the harmonic current content. The SVPWM has advantage of less harmonic content with this scheme has a value of 3.66 %, whereas 6.05 % with SPWM; it is useful principally to avoid malfunction of sensitive equipment by harmonic excess, as well as, problems of transformers and wiring overheating.

Acknowledgment This project has been partially funded by the CONACyT- Mexico grant CB-169062 and by the ECEST-SEP (Espacio Común de Educación Superior Tecnológica) Program under the mobility scheme for students, and was supported by PROMEP: Redes Temáticas de Colaboración under the project titled: Fuentes de Energías Alternas.

References

1. J. Segundo-Ramirez, A. Medina, Modeling of FACTS devices based on SPWM VSCs. IEEE Trans. Power Deliv. **24**(4), 1815–1823 (2009)
2. N.G. Hingorani, L. Gyugyi, *Understanding FACTS*. (IEEE Press Editorial Board, New York, 2000)
3. R. Majumder, A. Ghosh, G. Ledwich, F. Zare, Control of parallel converters for load sharing with seamless transfer between grid connected and islanded modes, in *IEEE Power and Energy Society General Meeting*, 2008
4. P. Roncero-Sánchez, E. Acha, J.E. Ortega-Calderon, V. Feliu, A. García-Cerrada, A versatile control scheme for a dynamic voltage restorer for power-quality improvement. IEEE Trans. Power Deliv. **24**(1), 277–284 (2009)
5. H.M. Nguyen, D.S. Naidu, Advanced control strategies for wind energy systems: an overview, in *IEEE PES Power Systems Conference and Exposition* (*PSCE*), 2011
6. A. Junyent-Ferré, O. Gomis-Bellmunt, A. Sumper, M. Sala, M. Mata, Modeling and control of the doubly fed induction generator wind turbine. Simul. Model. Pract. Theory **18**(9), 1365–1381 (2010)
7. G.O. Cimuca, C. Saudemont, B. Robyns, M.M. Radulescu, Control and performance evaluation of a flywheel energy-storage system associated to a variable-speed wind generator. IEEE Trans. Ind. Electron. **53**(4), 1074–1085 (2006)
8. S. Niu, K.T. Chau, J.Z. Jiang, C. Liu, Design and control of a new double-stator cup-rotor permanent-magnet machine for wind power generation. IEEE Trans. Magn. **43**(6), 2501–2503 (2007)
9. Y. Wang, X. Lie, Coordinated control of DFIG and FSIG-based wind farms under unbalanced grid conditions. IEEE Trans. Power Deliv. **25**(1), 367–377 (2010)
10. S. Mishra, Y. Mishra, F. Li, Z.Y. Dong, TS-fuzzy controlled DFIG based wind energy conversion systems, in *IEEE Power and Energy Society General Meeting*, pp. 1–7, 2009

11. F.M. Hughes, O. Anaya-Lara, N. Jenkins, G. Strbac, Control of DFIG-based wind generation for power network support. IEEE Trans. Power Syst. **20**(4), 1958–1966 (2005)
12. J. Alcalá, V. Cárdenas, E. Rosas, C. Nuñez, Control system design for bi-directional power transfer in single-phase back to back converter based on the linear operating region, in *Applied Power Electronics Conference and Exposition* (*APEC*), pp. 1651–1658, 2010
13. A.K. Gupta, A.M. Khambadkone, A space vector PWM scheme for multilevel inverters based on two-level space vector PWM. IEEE Trans. Ind. Electron. **53**(5), 1631–1639 (2006)
14. R. Pena, J.C Clare, G.M. Asher, Doubly fed induction generator using back-to-back PWM converters and its application to variable-speed wind-energy generation, IEE Proc. Electric Power Appl. **143**(3), 231–241 (1996)
15. J.N. Wani, A.W. Ng, Paths to sustainable energy, Intech, Chapter 14, 2010
16. B. Wu, Y. Lang, N. Zargari, S. Kouro, *Power Conversion and Control of Wind Energy systems, Chapter 2*. (IEEE Press, New York, 2011)
17. S.M. Muyeen, M.A. Mannan, M.H. Ali, Simulation technique and application of space-vector PWM method in PSCAD/EMTDC, in *International conference on information and communication technology, ICICT*, (2007)
18. A. Hernandez, R. Tapia, O. Aguilar, A. Garcia, Comparison of SVPWM and SPWM techniques for back to back converters in PSCAD, in *Lectures Notes in Engineering and Computer Science: Proceedings of the World Congress on Engineering and Computer Science 2013, WCECS 2013*, San Francisco, USA, pp. 236–240, 23–25 October 2013

Chapter 6
Robust Control Strategy for a WECS Based at PMSG

Omar Aguilar Mejía and Ruben Tapia Olvera

Abstract The wind energy boom in the world began in 1980's. This work presents the modeling and control of a Wind Energy Conversion Systems (WECS) based Permanent Magnet Synchronous Generator (PMSG). The WECS adopts a back-to-back converter system with Voltage Source Inverter (VSI). In the strategy, the generator-side converter is used to tracks the maximum power point and the grid-side converter is responsible for the control of power flow and control the dc-link voltage. The control scheme uses a B-spline artificial neural network for tuning controllers when the system is subjected to disturbances. The currents from VSI's are controlled in a synchronous orthogonal *dq* frame using an adaptive PI control. The B-spline neural network must be able to enhance the system performance and the online parameters updated can be possible. This work proposes the use of adaptive PI controllers to regulate the current and DC link voltage. The simulations results confirm that the proposed algorithm is remarkably faster and more efficient than the conventional PI. Comprehensive models of wind speed, wind turbine, PMSG and power electronic converters along with their control schemes are implemented in MATLAB/SIMULINK environment.

Keywords Current control · Neural network · Permanent magnet synchronous generator · Phase locked loop · Voltage source inverter · Wind power generation system

O. A. Mejía (✉) · R. T. Olvera (✉)
Engineering Department at Polytechnic, University of Tulancingo, Huapalcalco,
43629 Tulancingo, HGO, Mexico
e-mail: omar.aguilar@upt.edu.mx

R. T. Olvera
e-mail: ruben.tapia@upt.edu.mx

H. K. Kim et al. (eds.), *Transactions on Engineering Technologies*,
DOI: 10.1007/978-94-017-9115-1_6,

1 Introduction

The history of wind power goes back many centuries. Wind power is an unending energy source and clean. Compared with conventional energy, wind power does not requirement fuel and does not damage the environment. Wind energy technology has evolved rapidly over the last 30 years with increasing rotor diameters and the use of complex power electronics to allow operation at variable rotor speed [1].

A WECS is a physical system that has three primary components. The first one is rotor connected to blades. As wind goes through blades, it makes the rotor rotate and therefore creates mechanical power. The second one is a transmission (gear box) that transfers power from the rotor to generator. The last one is electric generator that converts mechanical power to electric power [2], as shown in Fig. 1.

The drivers behind these developments are mainly the ability to comply with Grid Code connection requirements and the reduction in mechanical loads achieved with variable-speed operation [1]. Different machine types have been used in WECS through the ages. These include the squirrel cage induction generator (SCIG), doubly fed induction generator (DFIG), and synchronous generator PMSG with power ratings from a few kilowatts to several megawatts. DFIG's are widely used as the generator in a variable speed wind turbine system. But, the DFIG needs a gearbox to match the turbine and rotor speed. The gearbox many times suffers from faults and requires regular maintenance, making the system unreliable [3].

The reliability of the variable speed wind turbine can be improved significantly using a direct drive-based PMSG. This generator has the following main characteristics: (a) full operating speed range; (b) brushless; (c) full scale power electronic converter; (d) complete control of active and reactive power exchanged with the grid [4].

Power electronics, being the technology of efficiently converting electric power, plays an important role in wind power systems. It is an essential element for integrating the variable speed wind power units to achieve high efficiency and high performance in power systems. In particular, VSI units are used to match the characteristics of wind turbines with the requirements of grid connections, including frequency, voltage, control of active and reactive power, harmonics, etc. [5].

Major techniques to regulate the VSI output current include either a variable switching frequency, such as the hysteresis control scheme, or fixed switching frequency schemes, such as the ramp comparison, stationary and synchronous frame proportional–integral (PI), optimal, nonlinear, predictive control, and soft computing, neural networks control and on the fusion or hybrid of hard and soft control techniques [6]. However, tuning alternatives are needed for electrical grid complexity. Some author's mixed neural networks and PID techniques to strengthen the linear controller. In [7, 8] a back propagation neural networks was used to adjust coefficients K_P, K_I, and K_D of PID controllers attaining power regulation of wind turbines. The similar PID tuning strategy using radial basis

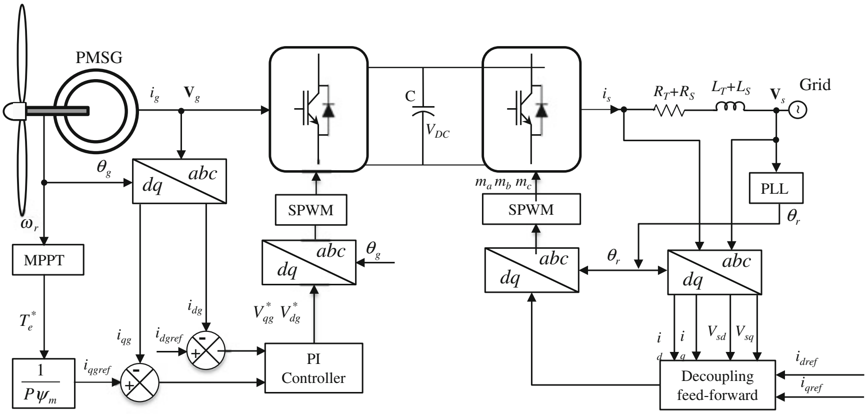


Fig. 1 Wind energy conversion system using PMSG

function (RBF) neural network for pitch angle control systems [9]. In these work we proposed a good adaptive tuning technique based on B-spline neural network (BSNN) [10].

Typically, the VSC synchronization is usually done by a phase-locked loop (PLL) system. Nevertheless, having a good synchronization permits a good monitoring of the grid voltage phase and amplitude, and enhancing the capability of injecting power into the grid. A good PLL can provide further advanced functionalities to the control system, as it is the case of the islanding detection mode for wind farms [11].

In this work a B-spline neural network (BSNN) is employed for two main tasks; one for PI simultaneous tuning, taking care of a key feature: the proposed controller must be able to enhance the system performance; the second the online parameters updated can be possible [10]. The strategy is proposed to update conventional PI parameters for currently operating in power converters that were tuned time ago.

2 Wind Energy Conversion System

A WECS is a structure that transforms the kinetic energy from the wind into electrical energy, the above system consists mainly of three parts: a wind turbine drive train, an electric generator, and two back-to-back VSCs, see Fig. 1. The wind turbine extracts the energy and makes that is then transferred to the generator. The generator transforms the mechanical energy into electrical energy. The output power of the electrical generator is supplied to the grid through a generator side converter and a grid side inverter [12].

2.1 Wind Turbine Model

A wind turbine is a power extracting mechanism. Wind turbine output power P_{wt} and wind turbine torque T_{wt} are given by the following equations [13]

$$P_{wt} = 0.5\rho\pi v^2 R^3 C_p(\lambda) \tag{1}$$

$$\Gamma_{wt} = \frac{0.5C_p(\lambda)\rho\pi v^2 R^3}{\lambda} \tag{2}$$

where ρ is the air density, R is the blade length, v is the wind speed and $\mathrm{C_p(\lambda)}$ is the turbine performance coefficient, $\lambda = \omega_h R/v$, ω_h is the angular rotor speed for the wind turbine. The performance coefficient C_p is a function of the tip-speed-ratio. Therefore, the $C_p(\lambda)$ performance curve gives information about the power. The torque coefficient is derived as

$$C_\Gamma(\lambda) = C_p(\lambda)/\lambda \tag{3}$$

The torque coefficient can be described by a polynomial function of the tip speed ratio as [13]

$$C_\Gamma(\lambda) = a_0\lambda^2 + a_1\lambda + a_2 \tag{4}$$

The output power characteristics of the wind turbine are depicted in Fig. 2. The wind turbine can produce maximum power when the turbine operates at maximum $C_p(\lambda_{opt})$. If the wind speed varies, the rotor speed should be adjusted to follow the change. The objective optimum power from a wind turbine can be written as

$$P_{wt\_opt} = k_{opt}\omega_{h\_opt}^3 \tag{5}$$

where

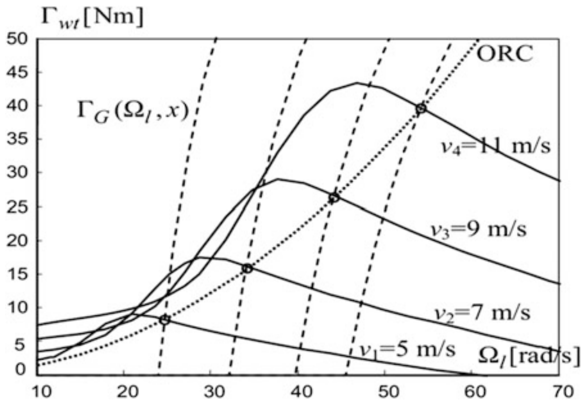
$$k_{opt} = 0.5\rho\pi v^2 R^5 \frac{C_p(\lambda_{opt})}{\lambda_{opt}^3} \tag{6}$$

Therefore, the objective optimum torque can be given by

$$\Gamma_{wt\_opt} = k_{opt}\omega_{h\_opt}^2 \tag{7}$$

The mechanical power generated by the turbine as a function of the rotor speed for different wind speed is shown in Fig. 2. The optimum power curve (P_{wt_opt}) shows how maximum energy can be captured from random wind. If the controller can properly follow the optimum curve, the wind turbine will produce maximum power at any speed within the allowable range [13, 14].

Fig. 2 Output power characteristics of wind turbine



2.2 Permanent Magnet Synchronous Generator Model

The PMSG is modeled under the following simplifying assumptions: (a) sinusoidal distribution of stator winding; (b) electric and magnetic symmetry; (c) negligible iron; (d) losses; (e) and unsaturated magnetic circuit. The voltage and electromagnetic torque equations of the PMSM in the *dq* reference frames are given by the following equations [4, 13, 14]:

$$L_d \frac{di_d}{dt} = -R_s i_d + L_q i_q \omega_s - v_d \tag{8}$$

$$L_q \frac{di_q}{dt} = -R_s i_q - (L_d i_d + \psi_m)\omega_s - v_q \tag{9}$$

where v_d and v_q are the *d-q* axis voltages, i_d and i_q are the *d-q* axis axis currents, R_s is the stator resistance, L_d and L_q are the *d-q* axis inductances, ω_s is the generator rotational speed, ψ_m is the permanent magnetic flux, and p is the number of pole pairs. Under the steady state condition, (8–9) reduces to

$$v_d = -R_s i_d + L_q i_q \omega_s \tag{10}$$

$$v_q = -R_s i_q - (L_d i_d + \psi_m)\omega_s \tag{11}$$

The electromagnetic torque is obtained as

$$T_e = \frac{3}{2} p \left[\psi_m i_q + \left(L_d - L_q\right) i_d i_q\right] \tag{12}$$

where p is the number of pole pairs. When the permanent magnets are mounted on the rotor surface, then $L_d = L_q$, therefore the electromagnetic torque is, $T_e = 1.5 p i_q \psi_m$. Using generator convention; the rotational speed of the generator and wind turbine driving torque as [13, 14]

$$J\frac{d\omega_h}{dt} = P(\Gamma_{wt} - T_e - B\omega_h) \tag{13}$$

where $\mathit{\Gamma}_{wt}$ is the turbine driving torque referred to the generator ($\mathit{\Gamma}_{wt} = P_{wt}/\omega_l$), B is the active damping coefficient representing turbine rotational losses and $\omega_h = i\omega_l$, where i is the ratio of a rigid drive train.

3 Back to Back Converter System

Figure 1 illustrates that back-to-back converter system can be considered as the composition of two VSC systems: (i) the right-hand side VSC system involves the real reactive power and DC-voltage controllers; (ii) and the left-hand side VSC system controlled the speed of the generator. The real-reactive power controller and the controlled DC-voltage power port are interfaced with grid and PMSG, respectively. The real-reactive-power controller and the controlled DC-voltage power port are connected in parallel from their DC-side terminals [10, 15].

The WECSs are requested to operate robustly in different grid locations and to keep ancillary services in order to behave as a conventional power plant. The control scheme for the WECS based PMSG is designed to satisfy the grid requirements.

3.1 VSC Dynamic Model

Figure 1 depicts the diagram of a three-phase three-wire VSC connected to the AC system represented by an equivalent Thevenin circuit via the inductance and resistance (L_T, R_T) of the coupling transformer. The converter DC terminal is connected to a shunt capacitance (C_{dc}) and resistance (R_{dc}), which represent losses switching losses. The three-phase AC side voltage balancing equations of the VSC are expressed as [16]:

$$\frac{d\mathbf{I}_{abc}}{dt} = \frac{1}{L_s + L_T}[-(\mathbf{R}_s + \mathbf{R}_T)\mathbf{I}_{abc} + \mathbf{V}_{Sabc} - \mathbf{V}_T] \tag{14}$$

where $\mathbf{I_{abc}} = [I_a\, I_b\, I_c]^T$ is the three-phase current vector, $\mathbf{V_{Sabc}} = [\mathrm{V}_{Sa}\, \mathrm{V}_{Sb}\, \mathrm{V}_{Sc}]^T$ is the three-phase AC source voltage vector, $\mathbf{V_T} = [\mathrm{V}_{Ta}\, \mathrm{V}_{Tb}\, \mathrm{V}_{Tc}]^T$ is the voltage source converter AC terminal three-phase voltage vector. The dc-side voltage dynamic expression is deduced based on power balance between the ac and dc-side as

$$V_{dc}(t)I_{dc}(t) = P(t) - P_L(t) \tag{15}$$

where P(t) is the instantaneous real power at point of common coupling voltage, and P_L(t) includes the total power loss, and $P_{dc} = V_{dc}I_{dc}$ is the transferred power from the dc side to the “converter” system. Loss components include: (i) capacitor dielectric; (ii) switching and on-state; and (iii) losses in the converter ac-side components as represented by R_{dc}. The DC current is:

$$C_{dc}\frac{dV_{dc}}{dt}=\left[-\frac{V_{dc}}{R_{dc}}-I_{dc}+\mathbf{I}_{\mathbf{abc}}^{\mathbf{T}}\frac{\mathbf{V}_T}{V_{dc}}\right] \tag{16}$$

Using a orthogonal transformation, the state equations for the back-to-back are [16]:

$$\frac{di_{d1}}{dt}=-a_1 i_{d1}+\omega_1 i_{q1}+b_1 V_{sd1}-b_1\mathrm{V}_{Td1} \tag{17}$$

$$\frac{di_{q1}}{dt}=-a_1 i_{q1}-\omega_1 i_{d1}+b_1 V_{sd1}-b_1\mathrm{V}_{Td1} \tag{18}$$

$$\frac{di_{d2}}{dt}=-a_2 i_{d2}+\omega_2 i_{q2}+b_2 V_{sd2}-b_2\mathrm{V}_{Td2} \tag{19}$$

$$\frac{di_{d2}}{dt}=-a_2 i_{d2}+\omega_2 i_{q2}+b_2 V_{sd2}-b_2\mathrm{V}_{Td2} \tag{20}$$

$$\frac{dV_{dc}}{dt}=\frac{1}{C_{dc}}\left[-\frac{V_{dc}}{R_{dc}}+\frac{1}{V_{dc}}\left(i_{d1}V_{Td1}+i_{q1}V_{Tq1}\right)+\frac{1}{V_{dc}}\left(i_{d2}V_{Td2}+i_{q2}V_{Tq2}\right)\right] \tag{21}$$

where $a = (R_s + R_T)/(Ls + L_T)$, $b = 1/(Ls + L_T)$, $\omega_1 = \omega_h$, ω_2 is grid frequency, $\mathrm{i}_{\mathrm{x}1}$ are currents of PMSG, $\mathrm{V}_{\mathrm{sx}2}$ are grid voltage, $\mathrm{V}_{\mathrm{Tx}2}$ are terminal voltage grid side converter.

3.2 Instantaneous Complex Power (d–q)

The α–β or Clarke transformation offers advantages in the dimension order reduction. However, for feedback control implementation is highly desirable that the signals to be constant. The inverse Park transform allows, from constant values generating signals in-phase or anti-phase, respect to the reference. That is especially useful for flexible AC transmission systems (FACTS) and power conditioners' control [17]. Instantaneous complex power are expressed by [14, 15, 17, 18]

$$S=\left(V_{Td}I_{Td}+V_{Tq}I_{Tq}\right)+j\left(V_{Tq}I_{Td}-V_{Td}I_{Tq}\right) \tag{22}$$

Equation (22) suggest that if $V_{Tq} = 0$, the active and reactive power components are proportional to i_d and i_q, respectively. This property is widely employed in the control of grid-connected three-phase VSC systems [14, 15, 17, 18]. The angle θ of the grid voltage is computed and provided by a phase-locked loop [17, 18].

4 Control WECS-Based on PMSG

The main advantage of the back-to-back converter is that it allows independently handle the active and reactive power flow between two AC systems with different characteristics (fundamental frequency, switching frequency, input voltage, etc.).

4.1 Generator Side Converter Control

The inner loops are constituted by two controllers, which regulate the *dq*-axis of the stator currents. The electromagnetic torque may be controlled directly by the *q*-axis current component i_q, therefore the speed can be controlled by changing *q* axis current, and d-axis current component i_d is set to zero to minimize the current and resistive losses for a given torque [3]. The outputs from the two current controllers are the *dq* axis stator voltage references, which are sent to the sinusoidal pulse-width modulation (SPWM) block as show in Fig. 3. The SPWM will generate the switching signals required by the IGBT elements of the converter.

4.2 Grid Side Converter Control

The control scheme of this strategy is exhibited in Fig. 4. The grid-side converter control aim is to supply a reliable electric power to the consumers, following a specific set of parameters such as voltage, frequency and harmonic levels. The control scheme contains two cascaded loops. The inner loops independently control the grid i_d and i_q currents, while the outer loops control the DC-link voltage, the reactive and active power. The feedback and feed-forward signals are first transformed to the *dq* frame and then processed by compensators to produce the control signals [18]. These control signals are transformed to the *abc* frame and sent to the grid side converter.

Under normal conditions the wind turbine's reactive power output is controlled under the range according to the grid codes. This work proposes the use of adaptive PI controllers to maintain the current and frequency AC load and DC link voltage constant. Reactive power flow should be maintained close to zero. Both parameters the proportional and integer gains are updated online to attain a proper performance under different operating conditions, without restructuring the control scheme.

5 Controller Tuning

The proposed can be achieved adding a B-spline neural network to update k_P and k_I gains in five PI controllers, Figs. 3, 4, where each PI transfer function is given by,

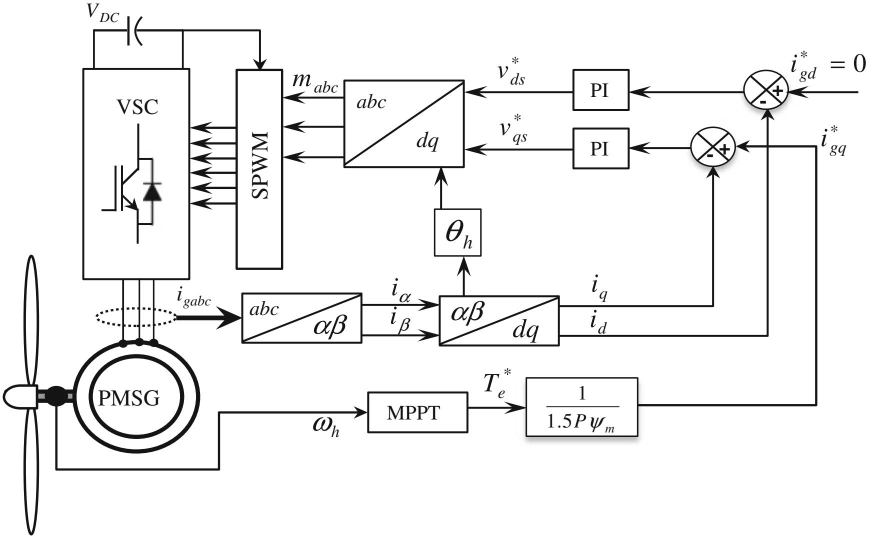


Fig. 3 Generator-side converter control system

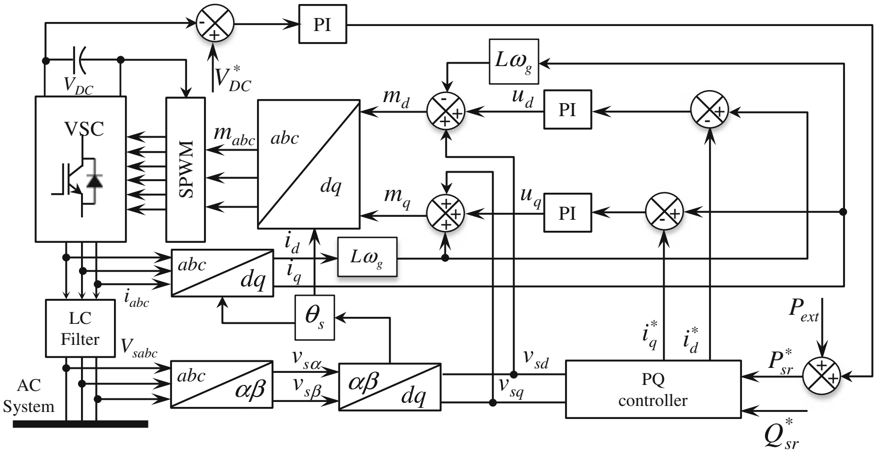


Fig. 4 Grid-side inverter control system

$$\frac{U(s)}{E(s)} = \frac{k_I + k_P}{s} \tag{23}$$

Thus, k_P and k_I are updated from a B-spline neural network at every sampled time.

The B-spline neural networks (BSNN's) are a particular case of neural networks that allow to control and model systems adaptively, with the option of carrying out such tasks on-line, and taking into account the power grid non-linearities. The BSNN's output can be described by [19],

$$y = \mathbf{a}^T \mathbf{w} \tag{24}$$

$$\mathbf{w} = \left[w_1\ w_2 \ldots w_\gamma\right]; \quad \mathbf{a} = \left[a_1\ a_2 \ldots a_\gamma\right] \tag{25}$$

where w_γ and a_γ are the γ-th weight and the γ-th BSNN basis function output, respectively; γ is the number of weighting factors.

In this paper it is proposed that k_P and k_I be adapted through one B-spline neural network, respectively, for each voltage source converter. The dynamic control parameters for back-to-back system can be described as follows:

$$k_P = NN_m(e_x, w_m) \tag{26}$$

$$k_I = NN_m(e_x, w_m) \tag{27}$$

where NN_m denotes the B-spline network which is used to calculate k_P and k_I; w_m is the corresponding weighting factor; $m = 1, 2, 3, 4, 5$ number of PI controllers. Figure 5 depicts a scheme of the proposed B-spline neural network.

The appropriate design requires the following a priori information: the bounded values of e_x, the size, shape, and overlap definition of the basis function. Such information allows to bound the BSNN input and to enhance the convergence and stability of the instantaneous adaptive rule [19]. Likewise, with this information the BSNN estimates the optimal weights' value. The neural network adaptive parameters, (26) and (27) are created by univariate basis functions of order 3, considering that e_x are bounded within $[-12, 12]$.

On-line learning of continuous functions, mostly via gradient-based methods on a differentiable error measure is one of the most powerful and commonly used approaches to train large layered networks in general [20], and for non-stationary tasks in particular.

In this application, the parameters' quick updating is looked for. While conventional adaptive techniques are suitable to represent objects with slowly changing parameters, they can hardly handle complex systems with multiple operating modes. The instantaneous training rules provide an alternative so that the weights are continually updated and reach the convergence to the optimal values. Also, conventional nets sometimes do not converge, or their training takes too much time [19, 20].

In this work, the neural network is trained on-line using the following error correction instantaneous learning rule [20],

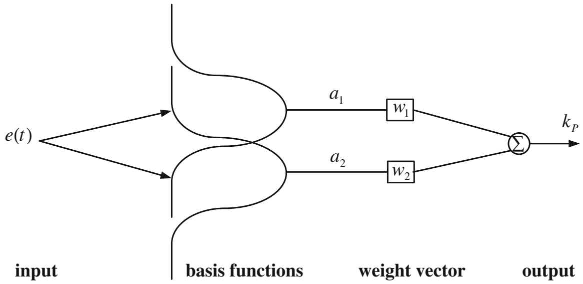


Fig. 5 Proposed BSNN for adapting k_p and k_i control parameters

$$w_i(t) = w_i(t-1) + \frac{\eta e_i(t)}{\|\mathbf{a}(t)\|_2^2} a_i(t) \tag{28}$$

where η is the learning rate and $e_i(t)$ is the instantaneous output error.

Respect to the learning rate, it takes as initial value one point within the interval [0, 2] due to stability purposes [20]. This value is adjusted by trial-and-error. If η is set close to zero, the training becomes slow. On the contrary, if this value is large, oscillations may occur. In this application, it settles down in 0.051 for k_P, and 0.0016 for k_I.

It is proposed that during the actualization procedure, a dead band is included to improve the learning rule convergence. The weighting factors are not updated if the error has a value below 0.1 %,

$$w_i(t) = \begin{cases} w_i(t-1) + \frac{\eta e_i(t)}{\|\mathbf{a}(t)\|_2^2} a_i(t), & \text{if } |e_i| > 0.0001 \\ w_i(t-1), & \text{otherwise} \end{cases} \tag{29}$$

This learning rule has been elected as an alternative to those that use, for instance, Newton's algorithms for updating the weights [21] that require Hessian and Jacobian matrix evaluation. Regarding the weights' updating, (29) should be applied for each input–output pair in each sample time; the updating occurs if the error is different from zero. That is the reason because it is said that the weights converge to optimal values [20]. Hence, the BSNN training process is carried out continuously on-line, while the weights' values are updated using only two feedback variables.

6 Test Results and Analysis

In order to demonstrate the feasibility of this proposition, a WECS-Based on PMSG is employed. Matlab-Simulink are used for simulation, the proposed tuning performance is exhibited. To analyze the results, simulations are developed under

Table 1 Parameters of PMSG wind power system

Blade length	$R = 90$ m
Rated line–line voltage	4000 V
Rated stator current	490 A
Rated power factor	0.7162
Rated mechanical torque	58.4585 kN m
Stator resistance	$R_s = 24.21$ mΩ
Number of poles	$p = 8$
Stator inductance	$L_d = L_q = 9.816$ mH
Magnet flux linkage	$\psi_m = 4.971$ Wb

different scenarios with PI controllers tuned by BSNN (dynamic parameters). Some operating conditions are taken into account. The models of the low-power (3-kW) rigid drive PMSG based WECS shown in Fig. 1 are included in the simulations.

Major system parameters are listed in Table 1 [13]. The power converter and the control algorithm are also implemented and included in the model. The sampling time used for the simulation is 20 μs. The wind speed profile is considered varying smoothly with step rate at different slopes, as see in Fig. 6. The system is subjected to the following sequence of events: until t = 0.033 s, $P_{ref} = 2200$ W, $Q_{ref} = 0$. At t = 0.033 s, P_{sref} is subjected to a step change from 2200 to 2700 W. At t = 0.066 s, P_{sref} is subjected to another step change from 2700 to 1800 W.

Figure 7 shows the simulation result of DC link voltage with the proposal and considering fixed parameters. Figure 8 exhibits the dynamic behavior of the reactive power at DC link bus, where the proposed scheme is worked. The transient response is diminished in terms of overshot magnitude without parameters update.

The adaptive neural network PI exhibits very well performance adapting itself to the new conditions. Figure 8 illustrates that P_s and $_{Qs}$ rapidly track P_{sref} and Q_{sref}, respectively.

Figure 9 shows the instantaneous load currents when the load changes. The load current is changing with the load variations as expected. From Fig. 9, it is seen that there is no significant rise in the current waveform during load transient. Figure 10 displays the proportional an integral gain's evolution for controller one. Quite similar results are exhibited for all adaptive parameters.

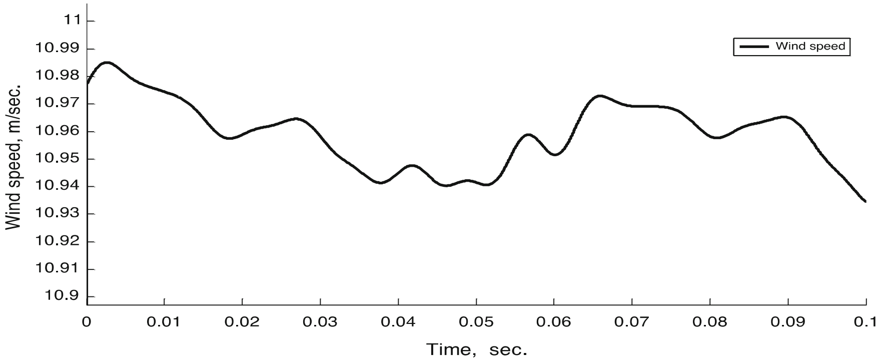


Fig. 6 Wind speed variation in m/s

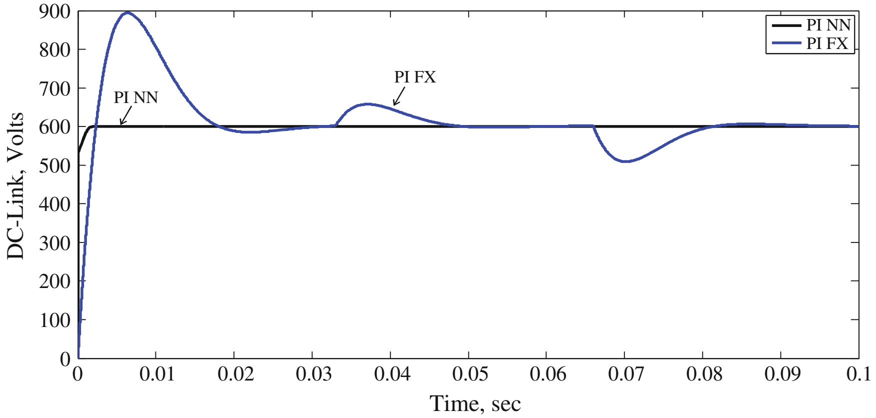


Fig. 7 DC link voltage performance

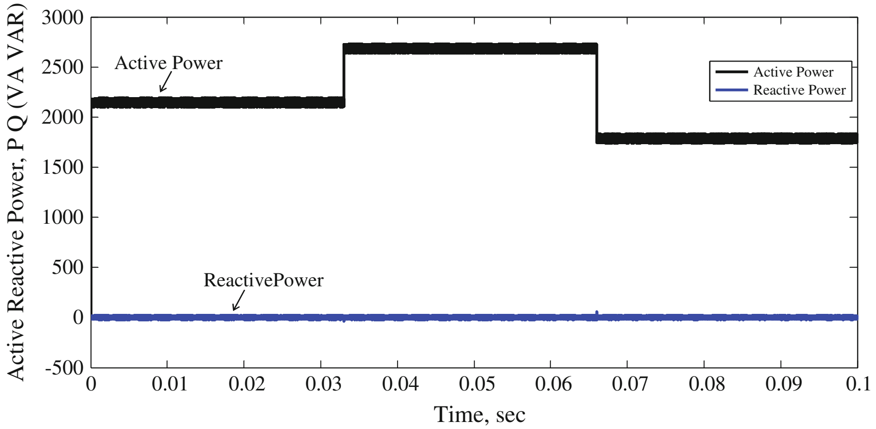


Fig. 8 Active and reactive power in WECS terminal

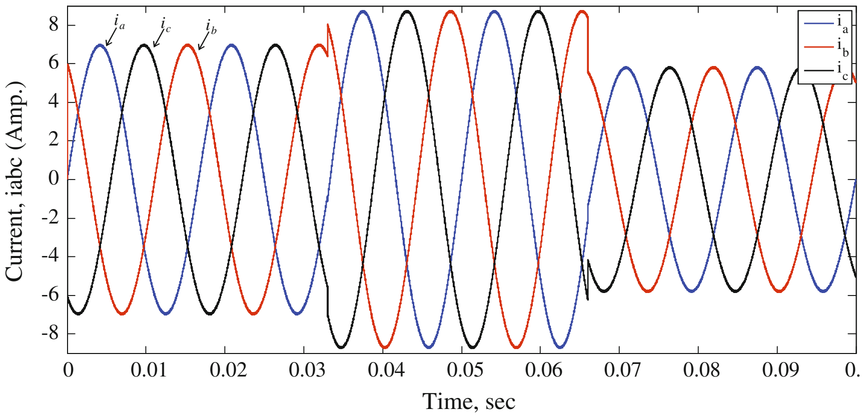


Fig. 9 Instantaneous output line current

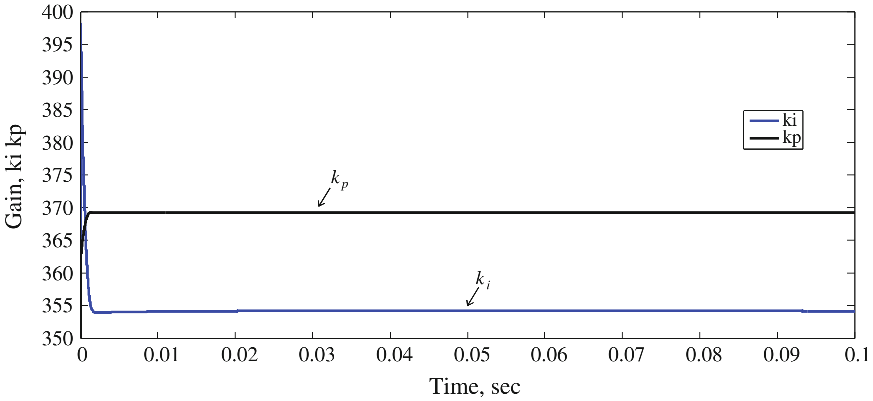


Fig. 10 Proposed BSNN for adapting k_I and k_P parameter, controller 3

7 Conclusion

The aim of the work is to show the performance of adaptive PI parameters as a mean to enhance VSC performance. In order to attain such purposes a BSNN control is proposed. With this neural adaptive scheme, the possibility to implement the on-line updating parameters is potential due to it has learning ability and adaptability.

The simulation results shown that the system has a stable operation at various load conditions. Unlike the conventional technique, the BSNN exhibits an adaptive behavior since the weights can be adapted on-line responding to inputs and error values as they arise.

References

1. O. Anaya-Lara, N. Jenkins, J. Ekanayake, P. Cartwright, M. Hughes, *Wind Energy Generation Modelling and Control* (Wiley, Chichester, 2009), pp. 110–112
2. T. Ackermann, *Wind Power in Power Systems* (Wiley, Chichester, 2005)
3. B. Wu, Y. Lang, S. Kouro, *Power Conversion and Control of Wind Energy Systems* (IEEE Press, New York, 2011)
4. Variable Boldea, *Speed Generators* (CRC Press, Boca Raton, 2006)
5. G. Abad, J. López, M.A. Rodríguez, L. Marroyo, G. Iwanski, *Doubly Fed Induction Machine* (Wiley, New Jersey, 2011), pp. 1–25
6. M. Hoa, D. Subbaram Naidu., Advanced Control Strategies for Wind Energy Systems: An Overview, in *International Conference on Power Systems, 2011*, Vol. 1, 2011, pp. 1–8
7. X. Yao, X. Su, L. Tian, Wind turbine control strategy at lower wind velocity based on neural network PID control, in *Intelligent Systems & Applications*, May 2009, pp. 1–5
8. Z. Xing, Q. Li, X. Su, H. Guo, Application of BP neural network for wind turbines. Intell. Comp. Technol. Autom. **1**, 42–44 (2009)
9. X. Yao, X. Su, L. Tian, Pitch angle control of variable pitch wind turbines based on NN PID, in *4th IEEE Conference on Industrial Electronics & Applications*, May 2009, pp. 3235–3239
10. O. Aguilar, M. Saucedo Jose, R. Tapia, "On-line control strategy for a WECS with permanent magnet synchronous generator", Lecture Notes in Engineering and Computer Science, in *Proceedings of The World Congress on Engineering and Computer Science 2013, WCECS 2013*, San Francisco, USA, 23–25 Oct. 2013, pp. 355–360 (2013)
11. A. Luna, J. Rocabert, G. Vazquez, P. Rodríguez, R. Teodorescu and F. Corcoles, Grid synchronization for advanced power processing and FACTS in wind power systems, in *IEEE Industrial Electronics Conference*, pp. 2915–2920
12. A. Valderrabano, J.M. Ramirez, Details on the implementation of a conventional StatCom's control, in *International Conference IEEE. Transmission and Distribution: Latin America, Bogota, Colombia*, 2008
13. I. Munteanu, A.I. Bratcu, N.A. Cutululis, E. Ceanga, Optimal Control of Wind Energy Systems. (London, Springer, 2008)
14. E. Haque, M. Negnevitsky, K.M. Muttaqi, A novel control strategy for a variable-speed wind turbine with a permanent-magnet synchronous generator. IEEE Trans. Ind. Appl. **46**(1), 331–339 (2010)
15. L. Shuhui, T.A. Haskew, R.P. Swatloski, W. Gathings, Optimal and direct-current vector control of direct-driven PMSG wind turbines. IEEE Trans. Power Electron. **27**(5), 2325–2337 (2012)
16. J. Arrillaga, Y. Liu, N. Watson, N. Murray, *Self-Commutating Converters for High Power Applications* (Wiley, Chichester, 2009)
17. J.C. Rosas, Simple Topologies for Power Conditioners and FACT's Controllers, Ph.D. dissertation, CINVESTAV, Gdl, 2009
18. A. Yazdani, R. Iravani, Voltage-Sourced Converters in Power Systems Modeling, Control, and Applications. (Wiley, New Jersey, 2010)
19. J. Brown, C. Harris, Neurofuzzy Adaptive Modeling and Control. (Prentice Hall International, London, 1994)
20. D. Saad, On-line learning in neural networks. (Cambridge University Press, Cambridge, 1998)
21. J.M. Ramirez, R.E. Correa-Gutierrez, N.J. Castrillon-Gutierrez, A study on multiband PSS coordination. Int. J. Emerg. Elect. Power Syst. **10**, 1–20 (2009)

Chapter 7
Optimum APC Placement and Sizing in Radial Distribution Systems Using Discrete Firefly Algorithm for Power Quality Improvement

Masoud Farhoodnea, Azah Mohamed and Hussain Shareef

Abstract This paper presents an improved solution to determine simultaneously the optimal location and size of active power conditioners (APCs) in distribution systems using the discrete firefly algorithm (DFA) for power quality enhancement. A multi-criterion objective function is defined to enhance voltage profile of the system, to minimize voltage total harmonic distortion and total investment cost. The performance analysis of the proposed DFA is performed in the Matlab software on the radial IEEE 34-bus test system to demonstrate its effectiveness. The DFA results are then compared with the standard firefly algorithm, standard particle swarm optimization (PSO), genetic algorithm, and discrete PSO. The simulation and comparison of results prove that the DFA can accurately determine the optimal location and size of the APCs in radial distribution systems.

Keywords Active power conditioner · Discrete optimization · Firefly algorithm · Optimal placement · Power quality · Voltage sag

1 Introduction

Due to the rapid development in the competitive deregulated electricity market, access to electrical energy is considered a basic right of every individual, which should be made available at all times under pre-determined baselines.

M. Farhoodnea (✉) · A. Mohamed · H. Shareef
Department of Electrical, Electronic and Systems Engineering, University Kebangsaan,
Bangi, Malaysia
e-mail: farhoodnea_masoud@yahoo.com

A. Mohamed
e-mail: azah@eng.ukm.my

H. Shareef
e-mail: shareef@eng.ukm.my

H. K. Kim et al. (eds.), *Transactions on Engineering Technologies*,
DOI: 10.1007/978-94-017-9115-1_7,

Consequently, any deviation from the baselines can be considered as a power quality disturbance and may cause failure of sensitive equipment at the utility and customer sides. Therefore, the electricity supply should be continuously analyzed, predicted and enhanced to ensure that the delivered power is within the pre-specified baseline. Among the power quality disturbances, voltage variation, voltage sag and harmonic distortion are the important power quality disturbances that can affect customers, cause interruption in the processing plants and cause financial losses. Hence, the best solution to mitigate power quality disturbances, especially voltage sag and harmonic distortion and to protect sensitive equipment is to install proper types of custom power devices (CPDs) such as the active power conditioner (APC). APC can be installed by an individual customer or a group of customers to mitigate deviations in power quality from the current baseline or to enhance power quality levels. The mitigation option, location, and sizing of the required APC should be determined based on economic feasibility, which is a major concern in the selection process and must be optimized.

Up to the present literature, several heuristic optimization techniques have been applied to solve the optimal placement and sizing problems of CPDs relative to different objectives and constraint functions. A genetic algorithm (GA)-based optimization technique was used to optimally place a dynamic voltage restorer and a thyristor voltage regulator to minimize the total power quality cost due to occurrence of voltage sag [1]. Another GA-based algorithm was applied to solve the optimal placement problem of several types of flexible alternating current transmission devices to minimize the overall financial losses and improve the overall network sag performance in transmission systems [2]. An improved GA method using the niching GA was then developed to explore a wider search space by maintaining the genes' diversity to decrease the probability of convergence in the local optima [3]. The gravitational search algorithm was also developed for solving the optimal placement problem of distribution static synchronous compensator (D-STATCOM) to improve reliability, power quality and sag performance of a power system [4]. In addition to the CPDs, other devices such as distributed generators and capacitor banks have also been considered to optimally improve the power quality of a system using particle swarm optimization (PSO) [5], GA and combined GA and PSO [6]. Considering the discrete nature of the placement and sizing problems, discrete optimization techniques such as the discrete non-linear programming [7] and discrete PSO (DPSO) [8] are also developed to mitigate harmonic distortion and improve power quality using optimal placement of APCs.

In this paper, a new heuristic optimization technique is proposed to determine the optimal size and location of APC using the discrete firefly algorithm (DFA) for the purpose of enhancing power quality. A multi-objective problem is formulated with respect to voltage harmonic distortion, voltage profile of a system and total investment cost which includes installation and incremental costs, where the voltage limits, APC capacity limits, and power flow limits are considered as constraints of the control variables. The performance of the proposed DFA is then evaluated on the radial IEEE 34-bus test system using Matlab programming. Considering the effectiveness of APC to compensate voltage sags caused by motor-starting effects or inrush currents, it is also considered in the simulation for voltage profile improvement. For comparison purposes, the DFA is compared with the results obtained by the standard firefly algorithm (SFA), standard PSO (SPSO), GA, and the DPSO.

2 Modeling of Active Power Conditioner

APC is a parallel multi-function compensating device, which, depending on the available controller design, is able to mitigate voltage sag and harmonic distortion, performs power factor correction, and improves the overall power quality [9]. The voltage-source converter is the main part of the APC, which converts the dc-link voltage into three-phase ac voltages with controllable amplitude, frequency and phase. Considering the steady-state APC losses such as transformer and inverter losses, an accurate load flow model of the APC should be obtained. Figure 1 shows the schematic diagram of an APC and its Thevenin equivalent circuit with respect to bus N. The injected current I_{APC} at bus N can be expressed as

$$I_{APC}^{h} = I_{L}^{h} - I_{S}^{h} = I_{L}^{h} - \frac{\left(V_{S}^{h} - V_{N}^{h}\right)}{Z_{S}^{h}} = \frac{\left(V_{APC}^{h} - V_{N}^{h}\right)}{Z_{APC}^{h}} \tag{1}$$

where, I_{APC} is the injected current by the APC with phase angle δ_{i}; V_{APC} is the Thevenin voltage with phase angle δ_{v}; I_S is the utility-side current with phase angle θ_{i}; V_S is the utility-side voltage with phase angle θ_{v}; I_L is the load-side current with phase angle λ_{i}; V_L is the load-side voltage with phase angle λ_{v}; Z_S, Z_{APC}, and h are the utility impedance, APC Thevenin impedance, and harmonic order, respectively.

Equation (1) shows that the injected APC current, I_{APC}^{h} can correct voltage sag, voltage variation and harmonic distortion at bus N by adjusting the voltage drop across the impedance, Z_{APC} in the fundamental and harmonic frequencies.

To present accurately the effects of APC in the bus voltage calculations in the fundamental and harmonic frequencies, we assume that the APC is added at a PQ bus with impedance Z_{APC}^{h} between bus N and virtual bus K in the M-bus system. Therefore, the new impedance matrix of the system should be modified based on Z_{APC}^{h} as,

$$Z_{bus-new}^{h} = \begin{bmatrix} & & & & Z_{1N}^{h} \\ & & & & Z_{2N}^{h} \\ & & Z_{old}^{h} & & \vdots \\ & & & & Z_{MN}^{h} \\ Z_{N1}^{h} & Z_{N2}^{h} & \cdots & Z_{NM}^{h} & Z_{NN}^{h} + Z_{APC}^{h} \end{bmatrix} \tag{2}$$

The new column accounts for the increase in all bus voltages because of impedance, Z_{APC}^{h}. As the virtual bus K is short-circuited to the reference node, the impedance can be eliminated using the Kron's reduction method in (2) as

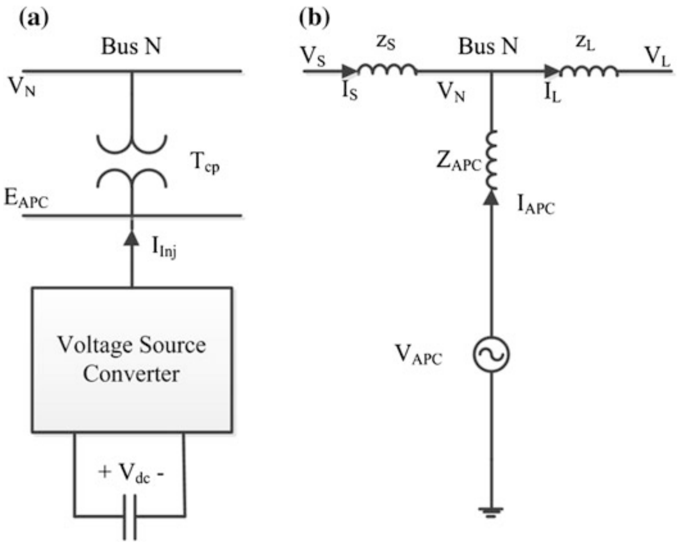


Fig. 1 **a** Schematic diagram of an APC. **b** Thevenin equivalent circuit

$$Z^h_{hi-new} = Z^h_{hi} - \frac{Z^h_{h(M+1)} Z^h_{(M+1)i}}{Z^h_{NN} + Z^h_{APC}} \tag{3}$$

Hence, the effects of APC on the system bus voltages in the fundamental and harmonic frequencies using the modified impedance matrix (2) and (3) can be calculated by the following equations using the backward/forward sweep method described in [9],

$$I^k_i = I^{rel}_i(V^k_i) + jI^{\mathrm{Img}}_i(V^k_i) = \left(\frac{P_i + jQ_i}{V^k_i}\right)^* \tag{4}$$

$$V^h = Z^h I^h \tag{5}$$

where V^k_i, I^k_i, I^{rel}_i, and I^{img}_i are the node voltage at the *k*th iteration, equivalent current injection at the *k*th iteration, and the real and imaginary parts of the equivalent current injection at the *k*th iteration, respectively. In addition, [V], [Z] and [I] are the bus voltage vector, system impedance matrix, and nodal injected current vector in the fundamental and harmonic frequencies.

It is noted that the values of *P* and *Q* in (4) are positive for conventional *PQ* buses and negative for the installed APC at bus *i*. Therefore, the bus voltage at bus *i* in the fundamental and harmonic frequencies, and consequently, the voltage total harmonic distortion (THD_V), can be changed by altering the rating of the installed APC during the optimization process.

3 Multi-objective Problem Formulation

The solution to the optimal APC placement problem in this paper aims to improve simultaneously the power quality of the system and to minimize the investment costs of the APCs. In this sense, the problem is essentially a multi-objective optimization problem where the objective function comprises of three sub-functions. There are also three constraints to the control variables as described in following subsections.

3.1 Objective Functions

3.1.1 Minimization of Average Voltage Deviation

The voltage improvement index of a power system is defined as the deviation of the voltage magnitudes of all the buses from unity. Thus, for a given system, the voltage improvement index for bus i is defined as

$$V_{dev-i} = \left(\frac{V_{i-ref} - V_i}{V_{i-ref}} \right)^2 \tag{6}$$

where $V_{i\text{-}ref}$ and V_i are the reference and actual voltages at bus i, respectively. Therefore, using the summation of normalized $V_{dev\text{-}i}$ for all buses, the average voltage deviation in the system in per unit (p.u.) can be expressed as

$$V_{dev-avr} = \frac{\sum_{i=1}^{M} V_{dev-i}^{norm}}{M} \tag{7}$$

where M is the total number of system buses. Equation (7) can be used to measure the deviation in the bus voltages from the reference voltage because of the unregulated voltage or the voltage sag that occurs because of motor starting in the system [10].

3.1.2 Minimization of Average $\mathrm{THD_V}$

To control the THD_V level of the whole system, the average of the normalized THD_V in the system buses is considered as

$$THD_{V-avr} = \frac{\sum_{i=1}^{M} THD_{V-i}^{norm}}{M} \tag{8}$$

where $THD_{V\text{-}i}^{norm}$ is the normalized THD_V in bus i.

3.1.3 Minimization of the Total Cost of APC

The total cost of an APC, which is composed of the installation and incremental costs, can be expressed in terms of the normalized total cost in a polynomial function as

$$C_{APC} = \frac{\sum_{i=1}^{k} (\alpha S_{APC-i}^{2} - \beta S_{APC-i} + C_{0-i})}{Cost_{\max}} \tag{9}$$

where, C_{APC}, C_0, and S_{APC} are the normalized total cost, fixed installation cost and operating range of the APC, respectively. In addition, α and β are fixed coefficients, which are assumed in this work as 0.0002466 and 0.2243, respectively.

3.2 Operational Constraints

3.2.1 Bus Voltage Limits

With respect to power quality and system stability considerations, each bus voltage V_i must be maintained around its nominal value $V_{i\text{-}nom}$ within a permissible voltage band, specified as [$V_{i\text{-}min}$, $V_{i\text{-}max}$], where $V_{i\text{-}min}$ and $V_{i\text{-}max}$ are the minimum and maximum permissible voltages at bus *i,* respectively. These limits can be expressed in terms of an inequality function as

$$V_{i-\min} \le V_i \le V_{i-\max} \tag{10}$$

3.2.2 APC Capacity Limits

Considering that the APC capacity is inherently limited by the energy resources at any given location, the capacity has to be constrained within a permissible band, specified as [$S_{APC\text{-}min}$, $S_{APC\text{-}max}$], where $S_{APC\text{-}min}$ and $S_{APC\text{-}max}$ are the minimum and maximum permissible values of each APC capacity, respectively. These limits can be expressed in terms of an inequality function as

$$S_{APC-\min} \le S_{APC} \le S_{APC-\max} \tag{11}$$

3.2.3 Power Flow Limits

The apparent power, S_l transmitted through branch *l* must not exceed its maximum thermal limit, $S_{l\text{-}max}$ in the steady state operation,

$$S_l \le S_{l-\max} \tag{12}$$

3.3 Overall Objective Function

The overall optimal APC placement problem can be configured as a constrained multi-objective optimization problem. Therefore, the weighted sum method is considered to combine the individual objective functions in terms of a single objective function. In addition, each constraint violation is incorporated in the overall objective function using the penalty function approach. Therefore, the final objective function to be minimized is expressed as

$$\begin{aligned} F &= f(Location, Size) \\ &= w_1 \frac{\sum_{i=1}^{M} V_{dev-i}^{norm}}{M} + w_2 \frac{\sum_{i=1}^{M} THD_{V-i}^{norm}}{M} + w_3 C_{APC} \\ &\quad + \lambda_V \sum_{i \in M} [\max(V_i - V_{i-\max}, 0) + \max(V_{i-\min} - V_i, 0)] \\ &\quad + \lambda_l \sum_{l \in L} \max(|S_l| - |S_{l-\max}|, 0) \\ &\quad + \lambda_{APC} \sum_{i \in P} [\max(S_{APC} - S_{APC-\max}, 0) + \max(S_{APC-\min} - S_{APC}, 0)] \end{aligned} \tag{13}$$

subject to

$$\sum_{i=1}^{3} w_i = 1 \quad and \quad 0 < w_i < 1 \tag{14}$$

where w_i and λ are the relative fixed weight factors assigned to the individual objectives and the penalty multipliers for violated constraints, which are large fixed scalar numbers. L, P, and M are the total line number, total APC number, and total bus number, respectively.

The weight factors are assigned to the individual objective functions based on their importance and may vary according to the desired preferences of the power system operators. In this paper, the proper weighting factors used are $w_1 = w_2 = 0.4$ and $w_3 = 0.2$, in which the first two objectives are assumed to be equally more important. In addition, because the scalarization of the objective functions is sufficient for Pareto optimality as long as $w_i > 0$, all the individual objective functions in (7) and (8) are scalarized by normalizing each component before summation to keep the average values between zero and 1.0. The same normalization process is applied on (9) by dividing the summation by the maximum cost.

4 Firefly Algorithm and Its Application

The background theories of the SFA and DFA are first described and then followed by the application of the DFA in solving the optimal location and sizing of APCs for power quality enhancement.

4.1 Standard Firefly Algorithm

Firefly algorithm is a novel nature-inspired metaheuristic algorithm that solves the continuous multi-objective optimization problems based on the social behavior of fireflies [10]. It is proven to be a very efficient technique to search for the Pareto optimal set with superior success rates and efficiency compared with the PSO and GA for both continuous and discrete problems. In SFA, two important issues arise, namely, the variation in light intensity I and the formulation of the attractiveness β. In the simplest form and considering a fixed light absorption coefficient γ, light intensity I, which varies with distance r, can be expressed as

$$I(r) = I_0 \exp(-\gamma r^2) \tag{15}$$

where I_0 is the light intensity at $r = 0$.

Considering the firefly's attractiveness as proportional to the light intensity seen by adjacent fireflies, the attractiveness β can be expressed as

$$\beta(r) = \beta_0 \exp(-\gamma r^2) \tag{16}$$

where β_0 is the attractiveness at $r = 0$.

The distance between any two fireflies i and j at x_i and x_j, respectively, can be calculated using the Euclidean distance as

$$r_{ij} = \left\| x_i - x_j \right\| = \sqrt{\sum_{d \in D} \left(x_{i,d} - x_{j,d} \right)^2} \tag{17}$$

where $x_{i,d}$ is the dth component of the spatial coordinate x_i of the ith firefly and D is the dimension of the problem. Therefore, the movement of firefly i to another more attractive (brighter) firefly j can be expressed as

$$x_i^{k+1} = x_i^k + \beta_0 e^{-\gamma r_{ij}^{k2}} \left(x_j^k - x_i^k \right) + \alpha \xi_i \tag{18}$$

where α is the randomization parameter and ξ_i is a vector of random numbers with Gaussian or uniform distributions.

4.2 Discrete Firefly Algorithm

In practice, the Pareto optimal solution using the SFA can be achieved by changing the positions of the fireflies to a more attractive position and decreasing the distance between them to zero in further iterations. Therefore, the convergence speed and performance of SFA can be enhanced using the logistic sigmoid function to constrain the position of the fireflies to the interval [0, 1]. By changing the positions of the fireflies to a more attractive position and decreasing the distance, the probability, S which is given in (19) decreases. When the distance of the fireflies are very far at a specific position, the probability of moving x_i^k in (20) to a new location x_i^{k+1} is very high, whereas by decreasing the distance in further iterations, the probability of moving x_i^k to a new location decreases.

$$S(r_{ij}^k) = \frac{1}{1 + e^{-r_{ij}^k}} \tag{19}$$

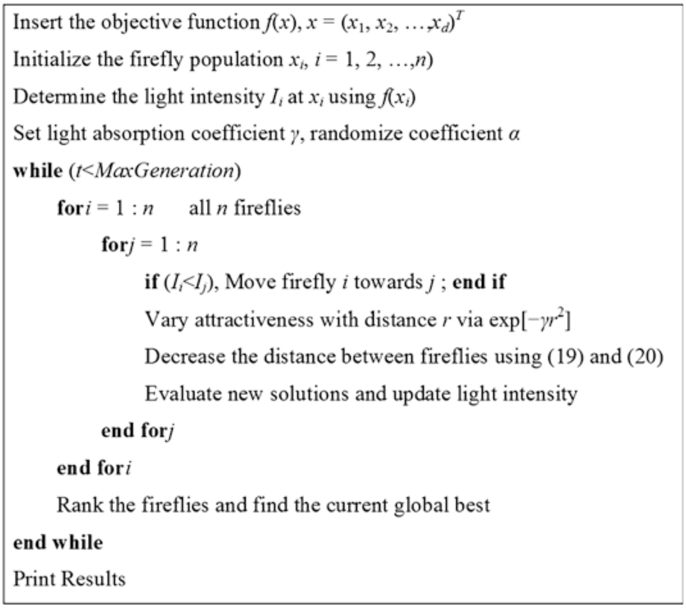
$$x_i^{k+1} = \begin{cases} x_i^k + \beta_0 e^{-\gamma r_{ij}^{k2}} \left(x_j^k - x_i^k \right) + \alpha \xi_i & \text{if} \quad r \quad \text{and} \quad < S\left(r_{ij}^k \right) \\ x_i^k & \text{else} \end{cases} \tag{20}$$

where $S(r_{ij}^k)$ is the probability of distance r_{ij}^k to be one, k is the iteration number, and *rand* is a random number in the interval [0,1].

The implementation procedure of DFA is described as shown in Fig. 2.

4.3 Application of DFA in Solving the Optimal Location and Sizing of APC

To solve the optimal location and size of the APC problem in radial distribution systems, DFA is applied to minimize the objective function (13). Initially, the number of APCs and the system specifications, including the bus and line data, should be considered as inputs of the DFA. The number of APCs can be chosen as any integer number between 1.0 and the maximum number of system buses. By increasing the number of APCs, the assigned power to each APC decreases; however, the total APC cost may increase because of the fixed installation cost, C_0 compared with a fewer number of APCs. The variables for optimization are the location of the APCs and the real and imaginary APC powers at the fundamental and harmonic frequencies. It is noted that all buses of the system are considered as a possible candidate bus for APC location, but only one APC can be installed at any of these buses. After initializing the locations and sizes of the APCs in terms of the firefly populations, as shown in Fig. 2, the bus voltages in the fundamental and harmonic frequencies in (4) and (5) are obtained using the backward/forward sweep method. The voltage variations and THD_V of each bus can be calculated



```
Insert the objective function $f(x)$, $x = (x_1, x_2, \ldots, x_d)^T$
Initialize the firefly population $x_i$, $i = 1, 2, \ldots, n$)
Determine the light intensity $I_i$ at $x_i$ using $f(x_i)$
Set light absorption coefficient $\gamma$, randomize coefficient $\alpha$
while ($t<MaxGeneration$)
    for $i = 1 : n$      all $n$ fireflies
        for $j = 1 : n$
            if ($I_i<I_j$), Move firefly $i$ towards $j$ ; end if
            Vary attractiveness with distance $r$ via $\exp[-\gamma r^2]$
            Decrease the distance between fireflies using (19) and (20)
            Evaluate new solutions and update light intensity
        end for $j$
    end for $i$
    Rank the fireflies and find the current global best
end while
Print Results
```

Fig. 2 Implementation procedure of DFA

using the computed bus voltages to calculate the objective function (13). Hence, using the obtained result from (13), the fireflies can be ranked to determine the current global solution, and we proceed to the DFA for the next iteration. The convergence criteria are set to $t = MaxGeneration$ or when the current global solution does not change for a specific number of iterations. If convergence is not achieved, the algorithm continues with the next generation. Figure 3 shows the schematic diagram of the procedures used in solving the optimal APC placement and sizing problem using the DFA.

5 Simulation and Results

To verify the effectiveness and applicability of the proposed DFA on radial distribution systems, the modified IEEE 34-bus test system are used, as shown in Fig. 4. The system is balanced and composed of several linear and non-linear loads with a total power of 0.925 MVA. In addition to the non-linear loads, which distort the voltage and current waveforms of the system, a heavy induction motor is installed at bus 25, which creates voltage variation and voltage sag problems in the systems.

To solve the optimal APC placement and sizing problem and improve the general power quality of the system, three APCs with power rating limits of [0 1.4]

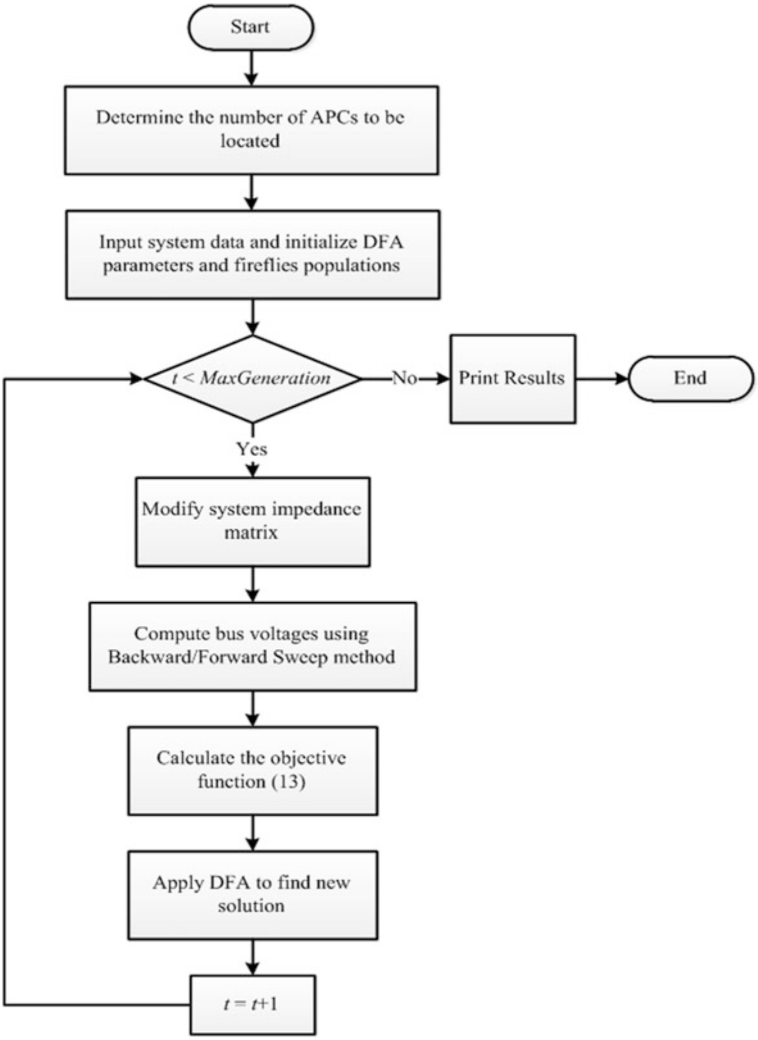

Start
Determine the number of APCs to be located
Input system data and initialize DFA parameters and fireflies populations
t < MaxGeneration
No
Print Results
End
Yes
Modify system impedance matrix
Compute bus voltages using Backward/Forward Sweep method
Calculate the objective function (13)
Apply DFA to find new solution
t = t+1

Fig. 3 DFA implementation to solve the APC placement and sizing problem

p.u. with base power of 100 kVA are considered for placement in the system, and minimum and maximum voltage limits are considered as 0.95 and 1.05 p.u, respectively. The objectives are to mitigate the harmonic distortion and to improve the voltage profile of the system using the proposed DFA method. The results are compared with the obtained results using SFA, SPSO, GA, and the recently proposed DPSO [8]. Table 1 shows the optimization results for the 34-bus test system, using the different optimization methods.

From the results shown in Table 1, the DFA achieves the best performance in determining the optimal location and size of the APCs and minimizing *F*. In

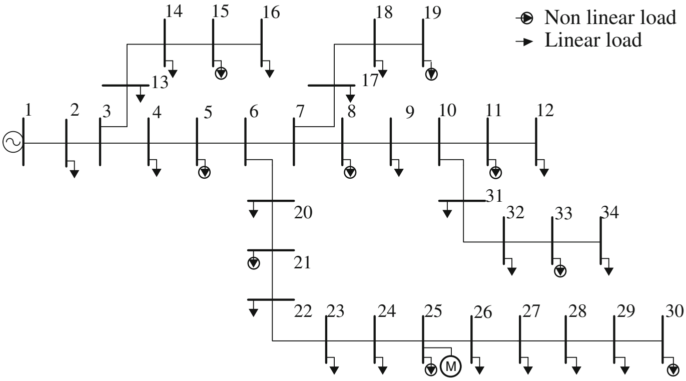


Fig. 4 Single line diagram of the IEEE 34-bus test system

Table 1 Optimization results of the 34-bus test system

Solver	Location (Bus)			Rating (p.u.)			Objective function, *F*	APC total cost (US $)
GA	27	33	24	0.209	0.293	0.416	0.2344	1,887,000
SPSO	18	15	30	0.272	0.115	0.556	0.2719	1,998,000
DPSO	26	15	10	0.491	0.565	0.217	0.2325	1,365,000
SFA	30	33	29	0.400	0.276	0.362	0.2518	1,932,000
DFA	8	30	15	0.206	0.597	0.399	0.2285	1,114,000

Table 2 SD and mean values at different initial values

Optimization method	SD (%)	Mean	Fmax	Fmin
GA	5.29	0.2323	0.2344	0.2314
SPSO	6.12	0.2713	0.2719	0.2666
DPSO	4.58	0.2319	0.2325	0.2301
SFA	4.97	0.2509	0.2518	0.2486
DFA	4.32	0.2281	0.2285	0.2269

addition, the obtained results of DFA and DPSO are very close, whereas those of the other methods returned a larger *F* value because of the trapping in the local minima. Furthermore, the obtained total APC cost using DFA is significantly smaller than those of the others, which presents a lighter installation and operational burden on utilities for further system development.

To investigate the sensitivity of the proposed method to the randomness of the initial values, the standard deviation (SD) and the mean value are calculated for 35 run times of the algorithm with the optimization method parameters being kept constant. Table 2 shows the obtained SD and the mean value and the comparison with the SFA, SPSO, GA, and DPSO methods. The comparison of the reported SD and mean values in Table 2 shows that the proposed DFA-based method has the

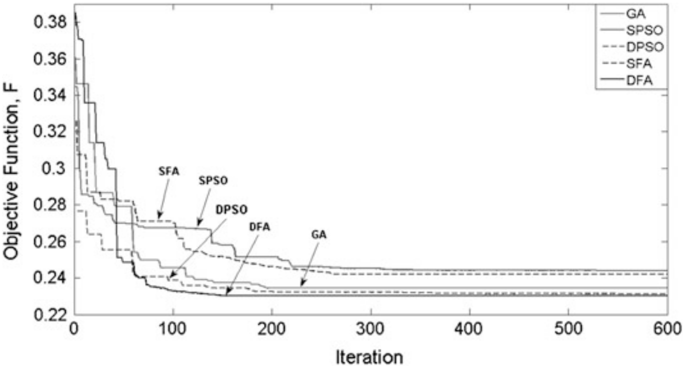


Fig. 5 Convergence characteristics of the optimization techniques

Table 3 System performance before and after the APC installation in the 34-bus test system

Bus no.	Bus voltage (p.u.)		THD_V (%)		Voltage deviation (%)	
	Before	After	Before	After	Before	After
1	1.000	1.000	0.000	0.000	0.000	0.000
2	0.986	1.002	1.984	0.450	1.444	0.240
9	0.912	1.001	10.056	0.492	8.815	0.112
12	0.910	0.999	10.163	0.488	9.002	0.059
19	0.915	1.000	9.811	0.506	8.549	0.007
20	0.906	1.001	11.811	0.447	9.354	0.053
24	0.847	1.000	20.520	0.712	15.309	0.038
25	0.829	1.002	23.127	0.869	17.146	0.158
26	0.827	1.011	23.169	0.860	17.318	1.100
30	0.825	1.041	23.221	0.836	17.527	4.088
31	0.910	1.000	10.225	0.484	8.976	0.035
32	0.910	0.999	10.317	0.480	9.012	0.067
33	0.910	0.999	10.384	0.476	9.030	0.084
34	0.910	0.999	10.385	0.476	9.036	0.089

smallest SD and mean value than the other optimization methods, which proves the higher accuracy and robustness of DFA in solving the optimal placement and sizing problem of the APCs for radial distribution systems.

The convergence characteristics of the DFA is shown and compared with those of the other methods in Fig. 5. As shown in the figure, the DFA converges to the final solution in fewer iteration steps compared with the other optimization methods.

The voltage profile and THD_V level in the 34-bus test systems before and after the optimal APCs placement for the selected buses are measured and shown in Table 3. Clearly, after the optimal APC placement, the voltage profile of the system significantly improves, even when a voltage sag with a depth of approximately 10 % occurs. In addition, the voltage harmonic distortion of all the system buses is also mitigated to meet the IEEE Std 519 requirements.

6 Conclusion

This paper presented an improved method to determine the optimal location and size of APCs in distribution systems. The method was based on DFA to solve the problem using a multi-objective function, defined to enhance the voltage profile of the system and to minimize the THD_V and the total investment cost. The performance of the system was analyzed using the Matlab software on the radial IEEE 34-bus test system. The results were compared with the SFA, SPSO, GA, and DPSO algorithms to verify the superiority of the proposed DFA over the other methods. The simulation and comparison results proved that the proposed DFA is able to find the most effective location and optimal size of the APCs in radial distribution systems.

References

1. C.S. Chang, Y. Zhemin, Distributed mitigation of voltage sag by optimal placement of series compensation devices based on stochastic assessment. IEEE Trans. Power Syst. **19**(2), 788–795 (2004)
2. Z. Yan, J.V. Milanovic, Voltage sag cost reduction with optimally placed FACTS devices, in *Proceedings of the 9th International Conference on Electrical Power Quality and Utilisation*, 2007
3. Y. Zhang, J.V. Milanovic, Global voltage sag mitigation with FACTS-based devices. IEEE Trans. Power Deliv. **25**(4), 2842–2850 (2010)
4. N. Salman, A. Mohamed, H. Shareef, Reliability improvement in distribution systems by optimal placement of DSTATCOM using binary gravitational search algorithm. Przegl Ąd Elektrotech. **88**(2), 295–299 (2012)
5. O. Amanifar, M.E.H. Golshan, Mitigation of voltage sag by optimal distributed generation placement and sizing in distribution systems with economic consideration using particle swarm optimization, in *Proceedings of the International Power System Conference*, Tehran, 2011
6. M. Moradi, M. Abedini, A combination of genetic algorithm and particle swarm optimization for optimal DG location and sizing in distribution systems. Int. J. Electr. Power Energy Syst. **34**(1), 66–74 (2011)
7. W. Chang, W. Grady, Minimizing harmonic voltage distortion with multiple current-constrained active power line conditioners. IEEE Trans. Power Deliv. **12**(2), 837–843 (1997)
8. I. Ziari, A. Jalilian, Optimal placement and sizing of multiple APLCs using a modified discrete PSO. Int. J. Electr. Power Energy Syst. **43**(1), 630–639 (2012)
9. T. Jen-Hao, C. Chuo-Yean, Backward/forward sweep-based harmonic analysis method for distribution systems. IEEE Trans. Power Deliv. **22**(3), 1665–1672 (2007)
10. M. Farhoodnea, A. Mohamed, H. Shareef, Optimum active power conditioner placement for power quality enhancement using discrete firefly algorithm, in *Proceedings of the World Congress on Engineering and Computer Science (WCECS 2013)*. Lecture Notes in Engineering and Computer Science, San Francisco, pp. 247–251, 23–25 Oct 2013

Chapter 8
Evaluation of Water Consumption and Neuro-Fuzzy Model of the Detergent Leavings Kinetics' Removal in a Clean in Place System

Rodrigo Sislian, Valdir M. Júnior, Leo Kunigk, Sérgio R. Augusto, Ricardo A. Malagoni, Ubirajara C. Filho and Rubens Gedraite

Abstract This work has focused on describing the water consumption and the Neuro-Fuzzy model of the detergent leavings kinetics' removal of a CIP System, based on the pH measured. The plant dynamics has been identified for different operational conditions. A flowrate value of 10.5 L.min^{-1} has been proved to be effective in order to provide the minimum required rinse water volume to execute the stage of the CIP system, which means that it is possible to optimize the process reducing energy, water and steam consumption as well as the time of unused machinery bringing productivity gains. The obtained models, allowed the prediction of the system dynamics behavior. The results were validated when compared with the experimental data. Three triangular membership functions for the input data modeled accordingly the pH dynamics with an error of 0.011 when comparing the validation data and the obtained model.

R. Sislian (✉)
Avenida Santa Inês, 881, apto. 102-B, São Paulo 02415-001, Brazil
e-mail: rodrigo.sislian@gmail.com; rodrigo@ifsp.edu.br

V. M. Júnior · L. Kunigk · S. R. Augusto
Praça Mauá 01, São Caetano do Sul 09580-900, Brazil
e-mail: vmelerojr@gmail.com

L. Kunigk
e-mail: kunigk@maua.br

S. R. Augusto
e-mail: sergio@pratica.com.br

R. A. Malagoni · U. C. Filho · R. Gedraite
Avenida João Naves de Ávila 2121—Bloco 1K, Uberlândia 38408-100, Brazil
e-mail: malagoni@feq.ufu.br

U. C. Filho
e-mail: ucfilho@gmail.com

R. Gedraite
e-mail: rgedraite@gmail.com

H. K. Kim et al. (eds.), *Transactions on Engineering Technologies*,
DOI: 10.1007/978-94-017-9115-1_8,

Keywords CIP system · Kinetics' removal · Modeling · Neuro-fuzzy · Optimization · Productivity gains

1 Introduction

The contact of food with poorly sanitized surfaces may increase the incidence of microorganisms affecting its quality. The presence of leavings also cause operational issues on the equipment, since it causes yields decreasing in thermal exchange performance and increases the system pressure drop. These factors justifies the value of a correct cleaning procedures of the inputs used in food processing. Usually companies perform the cleaning procedures poorly because it requires breaks in production.

The Clean in Place—CIP process is the most common cleaning procedure used in industry to ensure that the pipes/equipment are free of organic and inorganic contaminants. Thus, its study and optimization is fundamental, through the establishment of the residual kinetic removal in each step of the process [1].

Issues related with the production machinery, including the heat exchangers, are common in dairy industry. Nowadays, the process equipment involved with the milk and other dairy industry products manipulation must be completely stopped for a long time during the cleaning procedures. Such cleaning procedure involves the cleaning of the equipment with caustic detergent promoting the reduction of the residue (fat, lactose, proteins and minerals). After this step, the equipment must be rinsed to remove the sanitizing compound.

The aim of this work is the dynamic behavior identification of the residual concentration of sanitizing agent in the effluent rinse water of the equipment using Neuro-Fuzzy systems; this identification will be performed in function of the operating flow [2]. Models using the ANFIS (Adaptative Network-based Fuzzy Inference System) system were obtained which allows predicting the system's dynamic behavior [3].

2 Theoretical Framework

Cleaning the equipment properly may bring economical benefits for industry. According with [4], costs are related with production breaks for maintenance and equipment repairs, production losses, the excessive water usage and energy losses.

The knowledge and consequent optimization of a cleaning process allows its cost reduction [5]. The understandings on how residues are generated, as well as the temporal behavior of its removal, are necessary to optimize a cleaning procedure.

The residues kinetic removal knowledge may contribute to optimize the heat exchangers cleaning operations resulting, most of the times, in operational costs reduction of those equipments on the order of at least 50 % [6]. The kinetic removal establishment is obtained by the mathematical model of the system, and one of the techniques is the Neuro-Fuzzy modeling.

An adaptative neural network is characterized by a network formed by nodes and connections, where each node consists of a process unit associated with a function, and the connections represents inputs and outputs of each node. Each connection of the network indicates a relationship between the connected nodes. The group of nodes may be divided in two subgroups [7].

(1) Adaptative nodes, where the outputs depends not only on the inputs, but also on modifiable parameters within the model.
(2) And the other case, nodes, which function depends only on the inputs, called non-adaptative.

One of the approaches that use hybrid methods is the Adaptative Network-based Fuzzy Inference System (ANFIS) proposed by [8]. The ANFIS model works equivalently to the fuzzy inference systems, and its adaptative capabilities makes it applicable into a wide range of areas of study, for instance, in data classification and, in the case of this study, feature extraction from response curves. One of the properties of the ANFIS model is that the group of parameters may be decomposed to be used in a more efficient hybrid learning rule than the traditional techniques found in publication [9].

The ANFIS model is available in MATLAB® tool which supports only orders zero or one Takagi-Sugeno systems. The tool allows multiple input variables, but is limited to one output variable; in other words, for MISO (Multiple Input Single Output) studies [9].

3 Materials and Methods

3.1 Experimental System

The executed experiments to obtain the data used for the mathematical modeling of the residue kinetic removal of the heat exchanger used in this study were proceeded using the system showed in the Fig. 1 [10].

The system works as described in the sequence (referring to the Fig. 1). The product flows through the heat exchanger piped, by a positive displacement pump (1), making it possible for the process fluid to pass into the heat exchanger tubes (4 tubes pass). The centrifugal pump (2) is responsible for the heating agent movement through the heat exchanger shell. The heating agent temperature control is proceeded by the control valve (3), which is responsible to adjust the quantity of steam.

A simplified general diagram of the studied system is presented in Fig. 2, where:

(1) TT1 represents the temperature measurement of the process fluid on the heat exchanger output.
(2) TT2 the temperature of the heating agent output.

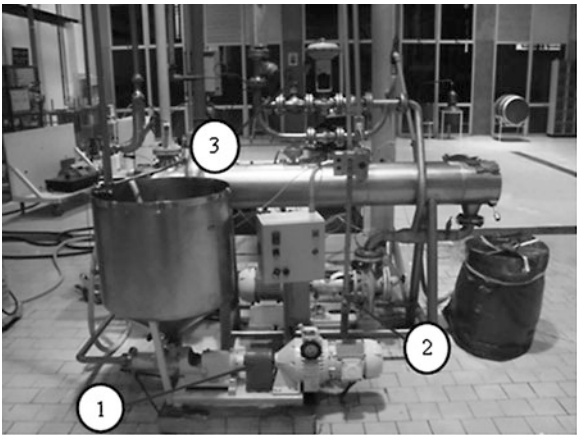


Fig. 1 Shell and tube heat exchanger for the studied system

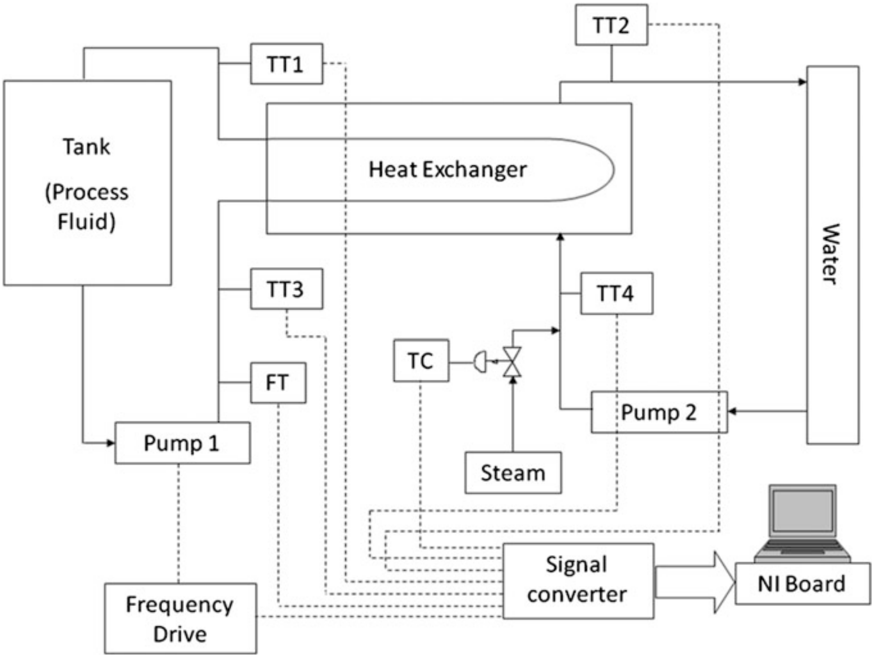


Fig. 2 Instrumentation flowchart for the studied system

(3) TT3 the temperature of the process fluid on the heat exchanger input.
(4) TT4 the temperature of the heating agent input.
(5) FT the process fluid flow on the heat exchanger input.
(6) TC the heating agent temperature control.

The electronic system used to collect data is basically composed of a microcomputer, a National Instruments® data acquisition board (model: NI PCI-6259) and the Labview® tool used to monitor, acquire data and control the process in real time.

Also, it has been developed an application using Labview® dedicated to collect data acquired in the proceeded experiments, from which the model of the studied system has been developed.

3.2 Sediment Build Up and Clean Up Stage with Detergent

Initially, milk at 85 °C and 9.0 L.min^{-1} was circulated during 90 min inside the tubes of the tube bundle of the heat exchanger to promote the desired deposits formation, for subsequent chemical cleaning stage. After the end of the heat exchanger incrustation stage, the milk was drained. Afterwards, such milk protein deposits was kept to rest for nearly 1 h considering the consolidation of the deposit process on the inner walls of the heat exchanger tubes.

Following the previous stage, the cleaning stage with detergent solution (sodium hydroxide, 0.5 % w/w) on the inner tubes of the heat exchanger has been started. Firstly, the solution was put in circulation in the inner tubes of the heat exchanger, with temperature and flow controlled respectively at 50 °C and 9.0 L.min^{-1} for 1 h, returning the output product to the tank of hot detergent. The monitored process variables were the values of pH, temperature, and flow of the solution and heating agent versus time.

3.3 Rising Process

Once the cleaning stage with detergent solution was finished, the rinsing process starts with potable water to remove residual detergent attached to the inner tube walls, used in the previous stage. The equipment is placed in steady state operation, feeding the inlet nozzle of the exchanger with water, pre-heated at nearly 50 °C, keeping the temperature controlled at 50 °C and flow at 9.0 L.min^{-1}, and directing the output of the process fluid to a small intermediate tank, used for the measurement of pH without turbulence.

The experimental procedure mentioned in the previous paragraphs was repeated for temperature of 50 °C and different values of flow of rinse water (4.0, 6.0, 7.5, 9.0, 10.5, 12, 14 and 16 L.min^{-1}). The flow was disturbed considering step changes imposed on the set–point of the flow controller, keeping the temperature controller operating in automatic mode.

3.4 Behavior of pH in Function of the Flow Fluctuation and Water Consumption Analysis

Based on the procedure described on the item 3.3, flow disturbances has been applied during the rinsing process, keeping the process fluid temperature constant and equals to 50 °C. In this test, as already mentioned, steps from 0 to 4.0 L.min^{-1}, 6.0, 7.5, 9.0, 10.5, 12, 14 and 16 L.min^{-1} has been applied.

Analyzing the obtained response curves, it has been concluded that the cleaning time decreases while the flow increases, but not linearly.

For the water consumption analysis, it has been used the developed empirical models in [10] and those were compared the experimental rinse time. A comparison between experimental data and the model results for rinse time and rinse water volume are presented in Figs. 3 and 4, respectively.

A flowrate value of 10.5 L.min^{-1} ($\text{Re} = 2071$) has been proved to be effective in order to provide the minimum required rinse water volume to execute the rinse stage of the CIP system and flow rates higher than 12 L.min^{-1} ($\text{Re} = 2367$) has been proved to not contribute in a significant way to improve the removal process [10].

3.5 Modeling Using the Adaptative Network-Based Fuzzy Inference System (ANFIS)

From the obtained values during the tests, the modeling of the pH behavior relative to the sodium hydroxide solution concentration present on the rinse water in the heat exchanger output over time has been performed with the Simulink/MATLAB® tool, using the ANFIS inference system. For that, it was used the "Fuzzy Logic" toolbox of the MATLAB® (performed using the Sugeno inference system).

For the neural network training, it has been used the pH responses for steps from 0 to 4.0 L.min^{-1}, 6.0, 9.0, 10.5, 12, 14 and 16 L.min^{-1}.

As described in [7] the data for the neural network training were sequentially inserted. It has been used three triangular membership functions for the fuzzy inference system generation: the current flow, F[k], and the pH delayed one sample, pH[k−1], as the inputs and the current pH, pH[k], as the output, with a selected tolerance for the error of 0.01 and 10 epochs of iteration. The "Grid Partition" algorithm has been used for the membership functions generated. For the neural network training, a 0.0099 average error has been achieved in 4 epochs (generated by MATLAB®).

The membership function obtained for the flow (F[k]) and the pH[k−1] inputs are shown in the Figs. 5 and 6, respectively.

The response surface obtained is presented in the Fig. 7.

It has been obtained nine rules (listed below). The linguistic terms used in the membership functions for the pH[k−1] input and the pH[k] output were Neutral,

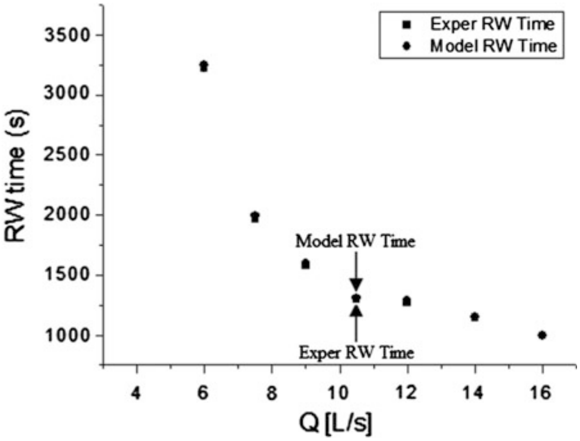


Fig. 3 Rinse time behavior for different flow rates at constant temperature of 50 °C

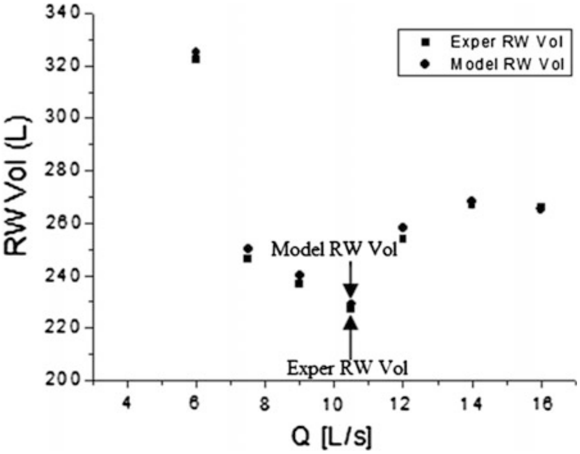


Fig. 4 Water volume behavior for different flow rates at constant temperature of 50 °C

Neutral-Alkaline and Alkaline; and the linguist terms for the flow (F[k]) input were Low, Average and High.

The obtained rules are:

(1) If (pH[k−1] is Neutral) and (Flow is Low) then (pH is Neutral)
(2) If (pH[k−1] is Neutral) and (Flow is Average) then (pH is Neutral-Alkaline)
(3) If (pH[k−1] is Neutral) and (Flow is High) then (pH is Alkaline)
(4) If (pH[k−1] is Neutral-Alkaline) and (Flow is Low) then (pH is Neutral)
(5) If (pH[k−1] is Neutral-Alkaline) and (Flow is Average) then (pH is Neutral-Alkaline)
(6) If (pH[k−1] is Neutral-Alkaline) and (Flow is High) then (pH is Alkaline)
(7) If (pH[k−1] is Alkaline) and (Flow is Low) then (pH is Neutral)
(8) If (pH[k−1] is Alkaline) and (Flow is Average) then (pH is Neutral-Alkaline)
(9) If (pH[k−1] is Alkaline) and (Flow is High) then (pH is Alkaline)

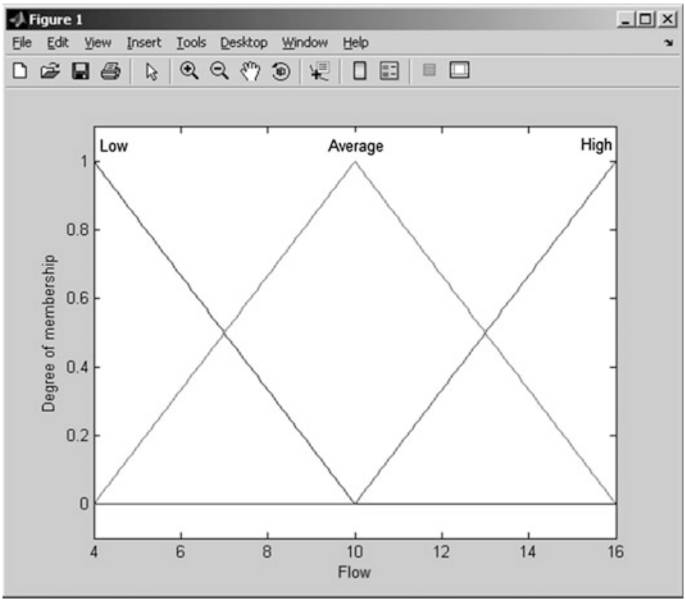


Fig. 5 Flow (F[k]) input for the neural network training

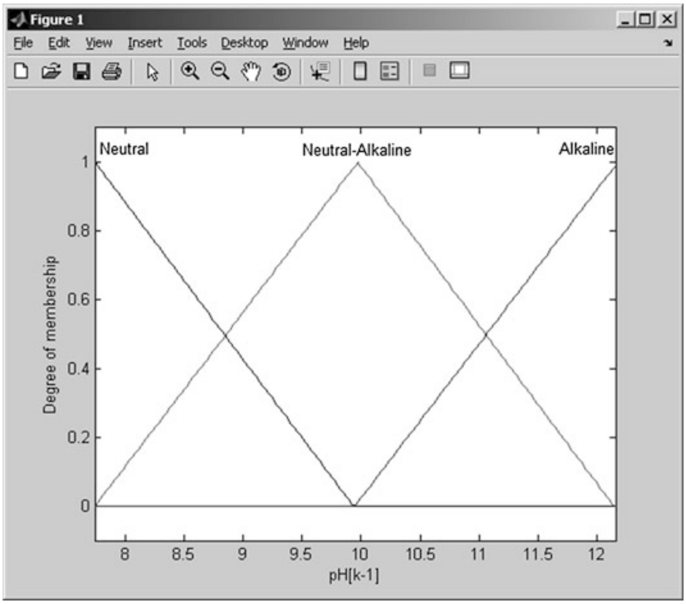


Fig. 6 pH delayed 1 sample (pH[k−1]) input for the neural network training

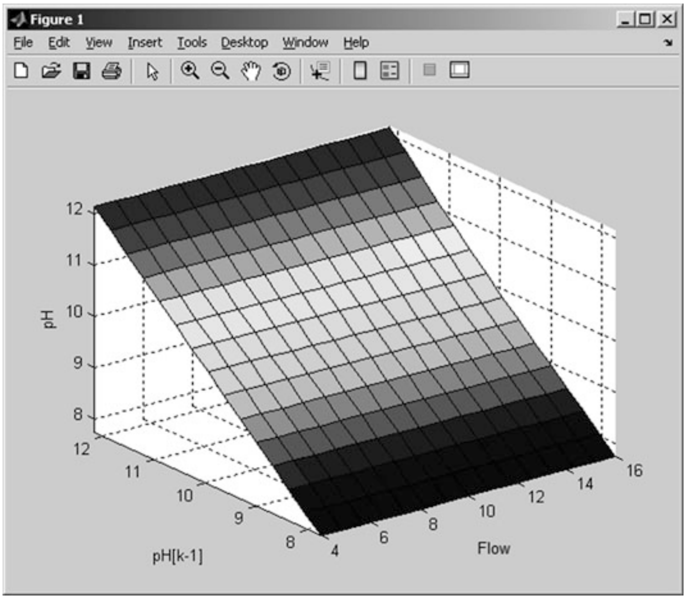


Fig. 7 Surface response of the model with 2 inputs and 1 output

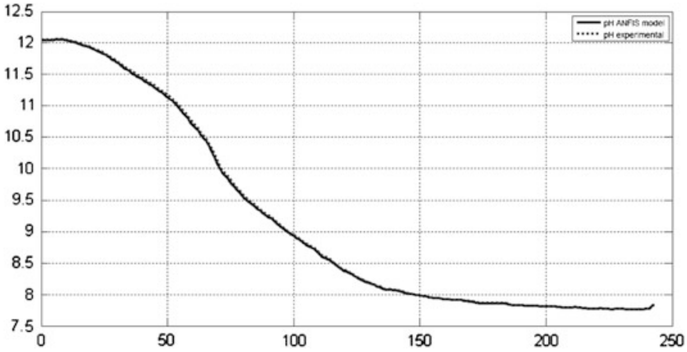


Fig. 8 Model versus validation data of the pH response for a step from 0 to 7.5 L.min^{-1} in the flowrate

For the model validation it has been used the pH response for a step from 0 L.min^{-1} to 7.5 L.min^{-1}; the average error between the model and the validation data was 0.011. The model (Fig. 8) represents adequately the system response with a small error.

4 Conclusion and Future Work

The implementation and instrumentation of a typical CIP station, with an appropriate monitoring, control and data acquisition of the process variables has been successfully done, as well as the CIP process dynamics identification (in the rising process). This identification has been done using the ANFIS inference system available in the "Fuzzy Toolbox" on MATLAB®/Simulink tool.

The advantage observed in the ANFIS system modeling is possibility on using a tool to easily and quickly generate a model independently of the production, operational condition and equipment used.

The empirical models, used for the water consumption analysis, showed that the flowrate value more suitable for operation considering the minimum required rinse water volume to execute the rinse stage of the CIP station used is nearly 10.5 $L.min^{-1}$.

The development of a virtual sensor based on the obtained models for the process variable considered in this paper, reveals an attractive industrial application perspective. An optimized control strategy based on a virtual sensor, in which a target objective function is related to the cost minimization, represents a potential application for process control industries.

Acknowledgment The authors acknowledge the financial support provided by Fapemig (Fundação de Amparo a Pesquisa do Estado de Minas Gerais) research fund and by Mauá Institute of Technology.

References

1. M.R. Bird, M. Barlett, CIP optimization for the food industry: Relationships between detergent concentration, temperature and cleaning time. Inst. Chem. Eng. **73c**, 63–70 (1995)
2. L. Gormezano, Desenvolvimento e implementação de sistema para avaliar a cinética de remoção de resíduos presentes nos tubos de trocador de calor feixe tubular. M.S. Dissertation, Department of Chemical Engineering, CEUN-IMT, 2007
3. R. Sislian et. al., Neuro-Fuzzy Model of the Detergent Leavings Kinetics' Removal in a Clean in Place System, in *Lecture Notes in Engineering and Computer Science: Proceedings of the World Congress on Engineering and Computer Science*, Vol II, WCECS 2013, pp. 617–620, San Francisco, 23–25 Oct 2013
4. S. Jun, V.M. Puri, A 2D dynamic model for fouling performance of plate heat exchangers. J. Food Eng. **75**, 364–374 (2006)
5. T.J.M. Jeurnink, D.W. Brinkman, The cleaning of heat exchangers and evaporators after processing milk or whey. Intern. Dairy J. **4**, 347–368 (1994)
6. P. Jong, Impact and control of fouling in milk processing. Trends Food Sci. Tech. **8**, 401–405 (1997)
7. R. Sislian, Estudo de Sistema de Limpeza CIP Usando Identificação de Sistemas, M.S. Dissertation, Department of Chemical Engineering, Unicamp, Mestrado 2012
8. J.S. Jang, ANFIS: Adaptive-Network-based Fuzzy Inference Systems. IEEE Trans. Syst. Man Cybern. **23**(3), 665–685 (1993)
9. L.C. Benini, M.M. Junior, Modelagem Neuro-Fuzzy com apoio do Matlab. Support Material, Presidente Prudente 2008
10. V. Melero Jr., Instrumentação e Identificação de um Processo de Sanitização Cinética CIP, M.S. Dissertation, Department of Chemical Engineering, CEUN-IMT 2011

Chapter 9
In Situ Photocatalytic Reduction for Contaminated Soil with Hexavalent Chromium by Titanium Dioxide

A. López-Vásquez, N. Ramirez and R. López-Vásquez

Abstract The photocatalytic reduction of hexavalent chromium was conducted in laboratory environment in order to evaluate the alternative use of this technology for in situ decontamination. The soil samples used had initial concentrations of Cr(VI) of 651.0, 308.0 and 112.0 mg kg^{-1}, with loads of catalyst TiO_2 which were exposed to UV irradiation through black light. Different loads of catalyst of 0.1–2 % (w/w) were tested in soil contaminated by hexavalent chromium with a concentration of 651 mg kg^{-1} for a period of 88 h of exposure. In addition, we examined the effect of alkalinity $Ca(OH)_2$ (10 % w/w). The rise in the pH due to $Ca(OH)_2$ addition shows no measurable effect on the chromium reduction. The photocatalytic remediation using TiO_2 combined with UV light showed their effectiveness in the reduction of Cr(VI) at 2.0, 4.0 and 6.0 cm of depth of contaminated soil, moreover also showed mobility of the contaminant towards to surface.

Keywords Chromium (VI) · Flood-affected soil · In situ treatment · Photocatalytic reduction · Remediation soil · Titanium dioxide

A. López-Vásquez (✉) · N. Ramirez
Environmental Engineering Department, Universidad Libre, Cra. 70 #53-40, Of. A 208, Bogota, Colombia
e-mail: andresf.lopez@unilibrebog.edu.co; andres.lopez.vasquez@correounivalle.edu.co

A. López-Vásquez
Chemical Engineering School, Universidad del Valle, Cl. 13 #100-00, Ed. 336, Cali, Colombia

R. López-Vásquez
Faculty of Engineering, Engineering Department, Universidad de Caldas, Cl. 65 #26-10, Manizales, Colombia
e-mail: rafael.lopez@ucaldas.edu.co

H. K. Kim et al. (eds.), *Transactions on Engineering Technologies*,
DOI: 10.1007/978-94-017-9115-1_9,

1 Introduction

Over time, variations in global temperature have caused climatic changes. One of these changes is represented in rainfall levels. For example, in Colombia, heavy rains in Colombia are typically a phenomenon due to "El Niño" or "La Niña" period when the sea surface temperatures increase in the Pacific Ocean near South America.

When the rains levels increase dramatically, can promote alterations in river basin leading to flooding in the lower grounds causing environmental disasters. The major changes that can occur in flooding soils are changes in the processes of erosion, nutrient leaching and densification or compaction. Such alterations generate negative impacts on productivity temporarily or permanently. In flooding areas, such phenomena can be noticed as runoff from higher areas, temporary ponding and flooding of depressed areas. This situation causes soil erosion on slopes and sedimentation in low relief areas generating an unlikely remediation. Similarly, in soils flooded for long periods of time, a reduction in air due to the lack of oxygen causes affectations in biological activity. The loss of fertility is affected by salinization levels that can restrict plant growth modifying soil properties. Each soil according to their characteristics reacts differently when a large amount of contaminants in the water, as in the case of floods deposited on it. Prolonged contact causes this resource alterations are reversible in some cases but not in others. The main negative effects that leave contaminants in the soil are altered in their resilience, the biogeochemical cycle and alteration in activities that develop microorganisms for plants. This affects the quantity and kind, thereby increasing and decreasing crop degradation due to possible changes in the composition of the products generating health risk consumer.

One of these disasters occurred in the Sabana de Bogota–Colombia, where the river exceeded the levels flood affecting great land extensions principally productive sectors as flower crops, livestock and grass. These phenomena reduced the biological activity due to the decreased oxygen available to plants and microorganisms [1]. Bogota is home to a large concentration of dense urban areas, mixed intensive and extensive agricultural areas, recreational areas, and many industrial sites [2]. Due to human activities over river basin, pollution levels are increased. The main pollution activities are related with waste disposal and are discharged to environment, especially in the aquatic ecosystem [3]. A typical example of human intervention can be seen in the upper part of the Bogota river basin (Villapinzon–Colombia), where is develops the bigger leather tannery industry of country based on wet-blue process. This condition require great amounts of water to develop its industrial activities and therefore produces important levels of waste affecting the natural sources, principally the river because wastewater are disposal in the Bogota river. Figure 1 shown the influence river area affected by flooding.

It's well known that this kind of process produce an industry wastes which poses serious environmental impact on water, terrestrial and atmospheric systems due its high oxygen demand, discolouration and toxic chemical constituents [5].

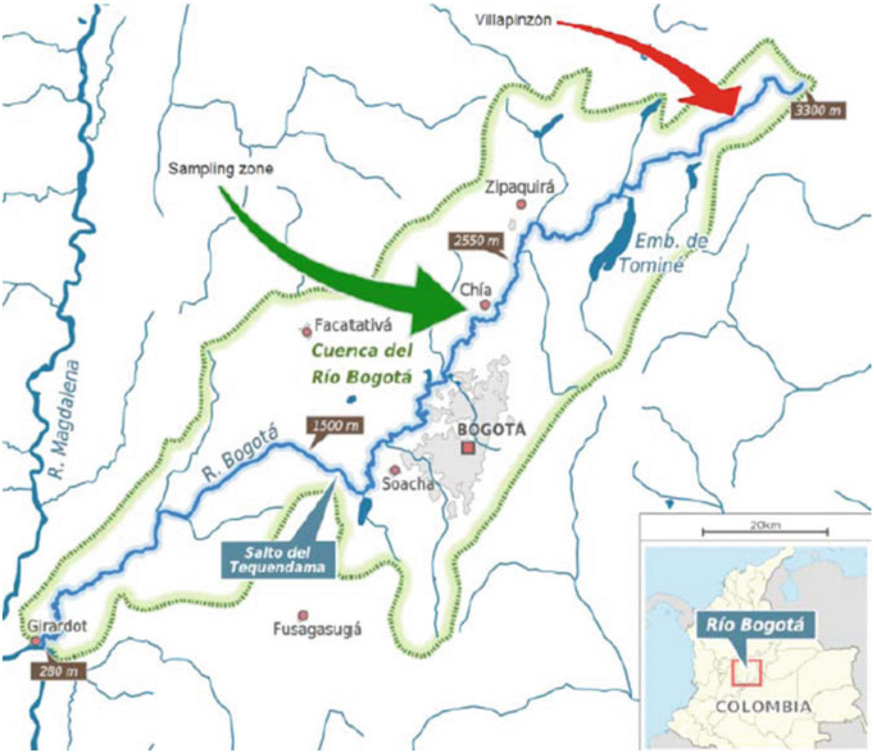


Fig. 1 Influence Zone River. Affected areas by flooding of contaminated Bogota river (sampling zone). Disposal of waste from tannery industry (Villapinzon). Adapted from [4]

Their wastes contain a mixture of both organic and inorganic pollutants such as chlorinated phenols and chromium. Other pollutants of concern within the tanning industry include azodyes, cadmium compounds, cobalt, copper, antimony, barium, lead, selenium, mercury, zinc, arsenic, polychlorinated biphyenls (PCB), nickel, formaldehyde resins and pesticides residues [6]. So, a lot of industrial dumping is discharged without any type of treatment into the river Bogota increasing the pollution levels due principally to organic pollutants and chromium (permissible limit 0.01 mg l^{-1}) [7].

1.1 Soil Contaminant (Chromium)

Chromium is the most representative metallic specie on the river which causes one of pollution problems when the river floods lands nearby to river basin. When this occurs, the chromium suffers a phase change, depositing into the ground and altering the metal natural concentrations. This generates raise in permissible levels according to environmental regulations [8]. It has been founded that soils contaminated with this kind of metals prevents the normal development of plants and crops causing decreasing in productivity levels. Cr(VI) has particular interest to scientist due to its toxicity and mobility therefore, is a great challenge its removal from wastewater. It is a strong oxidizing agent which is carcinogenic, mutagenic and diffuses easily through soil and aquatic environments. Cr(VI) does not form insoluble compounds in aqueous solutions therefore its separation is not feasible. Some oxyanions are very mobile and toxic in the environment, Cr(III) cations are not. Like many metal cations, Cr(III) forms insoluble precipitates. Thus, reducing Cr(VI) to Cr(III) simplifies its removal from effluent and also reduces its toxicity and mobility [9]. The chromium toxicity depends of its oxidation states. Hexavalent ions are highly mobile, soluble and bioavailable however trivalent form exhibits other properties. For example, Cr(III) exists in precipitates or is strongly adsorbed by inorganic and organic colloids in soils, and is one of the micronutrient elements in humans. Thus, many methods have been explored to utilize reductants or cause the transformation of Cr(VI) to Cr(III) in Cr(VI) contaminated soils [10]. Because it is only weakly sorbed onto inorganic surfaces, Cr(VI) is mobile in nature.

In soil, Cr(III) is adsorbed to components, which prevents chromium leaching into groundwater or its uptake by plants and in soils and basically is present as insoluble $Cr(OH)_3$ [11]. Depending of pH value, chromium is present in soil as Cr complexes. Under acidic conditions can be identified species such as $Cr(H_2O)_6^{+3}$, $CrOH^{+2}$ and $HCrO_4$. These forms are easily adsorbed by macromolecular clay compounds. This condition can be due to increased negative charge of soil or by protonation of successive ligating groups in clay. Under neutral or alkaline conditions in soils, Cr(VI) exists in soluble forms as chromates. Under oxidation and reduction, chromium converts to Cr(III) to Cr(VI) and viceversa. This depends on, O_2 concentration, presence of reducers and mediators (ligands or catalyst) [12].

Because tannery wastewater contains a complexity of pollutants including chromium and chlorinated phenols as indicated earlier, it is vital to dissect the toxic nature of such wastewater both to understand its environmental impacts and identify potential remediation strategies [13].

1.2 Treatment for Contaminated Soil by Photocatalysis

Due to the great number of sources and environmentally hazardous chemical composition of these contaminants, implementation of innovative treatment processes for contaminated soil remediation is a matter of pressing concern. Numerous potential technologies exist for soil remediation and degradation processes. These include biological, physical and chemical treatments.

Advanced oxidation processes are promising techniques for degrading an extensive variety of hazardous compounds in remediation of soil at waste disposal and spill sites. Unlike non-destructive traditional methods, like volatilization or adsorption onto a solid phase, this process has the advantage of destroying the organic compounds, in–situ, by redox reactions on the catalyst surface. Photocatalytic processes can be applied both in situ (to soil in place) and ex situ (after soil excavation). Matching the remedial oxidant and technology of delivery to the contaminant of concern and site conditions is an extremely important step in the successful remediation of contaminated soil.

The use of advanced oxidation process in the treatment of contaminated soil has been demonstrated recently [14]. Specifically, the use of ZnO and CdS in the photocatalytic reduction of Cr(VI) to Cr(III) has also been reported. Cr(VI) has a toxicity one hundred times higher than that of Cr(III). The photoreduction of Cr(VI) to Cr(III) can be achieved via a photocatalytic process with a simplified mechanism represented by Eqs. 1–5 [15]:

$$ZnO + hv \rightarrow ZnO(h^+ + e^-) \tag{1}$$

$$Cr_2O_7^{-2} + 14H^+ + 6e^- \rightarrow 2Cr^{+3} + 7H_2O \tag{2}$$

$$2H_2O + 4h^+ \rightarrow O_2^{-2} + 4H^+ \tag{3}$$

$$H_2O + h^+ \rightarrow OH \cdot + H^+ \tag{4}$$

$$Cr(VI) + H_2O_2 + H^+ \rightarrow Cr(III) + H_2O + O_2 \tag{5}$$

UV light illumination on ZnO produces hole–electron pairs (Eq. 1) at the surface of the photocatalyst. After the hole–electron pairs being separated, the electrons can reduce Cr(VI) to Cr(III) (Eq. 2), and the holes may lead to generation of O_2 in the absence of any organics (Eq. 3). In the photocatalytic reaction process, reduction reagents are important to eliminate the oxidative pollutants such as Cr(VI).

Another option is the use of titanium dioxide (TiO_2) semiconductor. This has the ability to reflect radiation visible and absorbs the electromagnetic radiation that is close to the UV region [16]. TiO_2 is the most used photocatalytic semiconductor in environmental applications. For its application, parameters like the pH (pzc $\sim$ 7.0), which must be above or below this value to the catalyst was negative or positively charged, catalysts loading and contaminant concentration must be considered to specific treatment. The TiO_2 reduction has been reported to be effective for the removal of Cr(VI). The mechanism is shown in Fig. 2.

In this way, have been tested that the addition of small amounts of TiO_2 enhanced the photodegradation of p,p-DDT on soil surfaces significantly [17]. The photocatalytic treatment using TiO_2 combined with solar light was very efficient in the destruction of pesticide Diuron in the top 4 cm of contaminated soils. The organic contaminants were destroyed in a relatively short time when the contaminated soils containing atrazine, 2-chlorophenol, 2,7-dichlorodibenzodioxin were mixed with TiO_2 and exposed to simulated solar radiation [18].

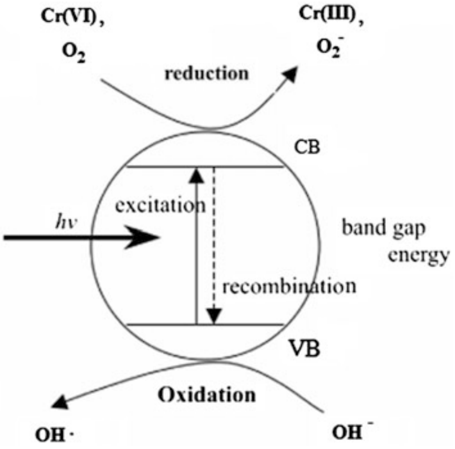


Fig. 2 Photocatalytic mechanism over TiO_2 surface. *CB* Conduction band and *VB* Valence band

In this study, a series of batch experiments at lab scale were carried out to investigate the effect of soil types on Cr(VI) reduction with or without irradiation. The effects of chromium concentration, catalyst loading, pH and soil depth were considered based on previous methodologies.

2 Methods and Materials

2.1 Materials

Titanium dioxide (Degussa, AEROXIDE P25) was used as received. Calcium hydroxide $Ca(OH)_2$, diphenilcarbazide, hydrogen peroxide and nitric, sulphuric and phosphoric acid were reagent grade.

2.2 Types of Soil

Four types of soils from Chia (Cundinamarca)–Colombia were selected for this study. Table 1 shown the sampling sites georeferencing.

The soils (0–40 cm) were sampled, dried in air, disaggregated by hand to reduce clumping, and sieved. The fraction less than 100 mesh size, which was mixed and split several times to obtain representative homogenized sample, was used in the experiments.

Table 1 Sampling sites coordinates of soil in Chia (Cundinamarca)—COLOMBIA

Type of soil	Sample	Coordinates
Affected by flooding	Suba–Cota	N 4°47′57″ W 74°5′44″
	Hato Grande	N 4°55′18″ W 74°0′21″
	La Chavela	N 4°51′14″ W 74°1′58″
Non-affected by flooding (N.A)	Chia–Cajica	N 4°54′53″ W 74°0′58″

N.A Non–affected

2.3 Irradiation Experiments to Contaminated Soil

The experiments were carried out to lab scale conditions, where 10 grams of soil located in Petri dishes and adjusted to experimental conditions. The plates were exposed to UV artificial light during 88 h, so all graphs showing the irradiation time refers to the amount of exposure to black light in sequential time.

2.3.1 Effect of the Contaminant Concentration

To Cr(VI) determination, 0.5 grams of soil were taken after 16, 24, 40, 48, 64, 72 and 88 h of artificial irradiation and analyzed by digestion for Cr(VI). The catalyst loading was made by adding the studied amount of TiO_2 in the soil, followed by manual shaking for 30 min in a glass vessel.

2.3.2 Effect of the Photocatalyst Concentration

To study the optimum concentration on the reduction of Cr(VI), three plates were prepared with contaminated soil samples. In each plate 0.0, 0.1 and 2.0 % TiO_2 (w/w) were incorporated into the soil and submitted to photocatalytic treatment. The influence of metal concentration was also evaluated by means of samples Suba–Cota, Hato Grande and La Chavela, followed by the addition of 1.0 % (w/w) catalyst.

2.3.3 Effect of Soil Alkalinity

The influence of soil alkalinity by $Ca(OH)_2$ in photocatalytic degradation was also assessed for Cr(VI) reduction. The addition of $Ca(OH)_2$ is performed to adjust the pH to alkaline conditions, was made by weighing of the amount of reagent and mixing to the soil.

2.3.4 Effect of Soil Depth

Considering that the artificial radiation penetrates only a few centimeters of the soil surface, we assume that the photocatalytic treatment is effective only in the top of soil. For these reason one experiment was carried out, to evaluate the reduction rate and mobility of Cr(VI) as a function of soil depth. Two plastic cylinder (3.5 cm in diameter and 10 cm high) containing soil contaminated with 1.0 % TiO_2 (w/w) and humidity (30 % w/w) were irradiated with artificial UV light on a 24 and 48 h period. Each cylinder was cutted at intervals of 2.0 cm down from top until cylinder bottom (three slices in total). Each slice was analyzed for Cr(VI).

2.4 Hexavalent Chromium Analysis

The metal ion present on soil surfaces were extracted by 0.5 M HNO_3, hydrogen peroxide and distilled water. The suspension was mixed in a shaker bath at 60 ± 2 °C for 30 min. and was filtered. Cr(VI) concentration was measured using the 1,5 diphenylcarbazide colorimetric method, using phosphoric acid buffer to control pH for the color development. The absorbance at 540 nm was measured in a 1.0 cm cell on a UV–Pharo 3000 spectrophotometer (Merck). A 781 Methrom pH/Ion, after two point calibration, was used to determine pH values [19].

3 Results and Discussion

The selected properties in Suba–Cota, Hato Grande, La Chavela and Chia–Cajica (N.A.) soils by flooding are listed in Table 2.

The biological characterization was a microorganism's count for fungi and bacteria totals. These are shown in Fig. 3.

3.1 Effect of the Hexavalent Chromium Concentration

Figure 4 shows the results obtained for the reduction when the concentration of Cr(VI) in the soil varied among 651.0, 308.0 and 112.0 mg kg^{-1} using 1.0 % (w/w) of TiO_2 and exposed to artificial UV radiation. Reduction follows a pseudo–first order kinetics with a half–life about 24 h to all Cr(VI) concentration levels. In the absence of catalyst, there is photolysis effect after 10.0 h of treatment until 88 h of irradiation being more noticeable at higher loads of pollutant. Cr(VI) reduction in the absence of the catalyst may be due to both photolysis and the

Table 2 Chemical and physical characteristics of soils

Item	Suba–Cota	Hato Grande	La Chavela	Chia–Cajica
True density, g/ml	1.73	2.23	1.68	2.49
Apparent density, g/cm^3	1.25	1.52	1.57	1.3
Humidity, %	50.01	39.95	18.3	9.87
Color	10 YR 4/3	7.5 YR 4/2	7.5 YR 3/2	7.5 YR 5/2
Textural kind	Clay	Clay loam	Silt clay	Silt loam
Liquid limit, %	86.9	88.3	62.8	54.9
Plastic limit, %	78.3	81.5	53.5	40.3
Plasticity index, %	9.97	11.8	9.1	13.1
pH	7.0	4.38	7.73	6.2
Electric conductivity, µs cm^{-1}	1.14	1.58	1.43	1.39
Organic carbon, %	9.93	7.82	8.01	4.2
Organic matter, %	16.55	13.18	14.25	7.31
Cationic exc. capacity, Cmol+ kg^{-1}	31.2	34.8	27.3	21.7
N (Org. Mat.), %	0.85	0.66	0.68	0.32
P, mg kg^{-1}	70.04	82.3	71.3	44.3
Zn, mg kg^{-1}	973	896	516	223
K, meq 100 g^{-1}	0.24	0.21	0.17	0.25
Mg, meq 100 g^{-1}	0.05	0−06	0.08	0.08
Ca, meq 100 g^{-1}	1.6	0.96	1.9	1.24
Na, meq 100 g^{-1}	0.9	0.94	0.85	0.76
Fe, mg kg^{-1}	488	269	345	230
Mn, mg kg^{-1}	63	260	36	66
Cu, mg kg^{-1}	345	62	32	12
Cr(VI), mg kg^{-1}	651	308	112	N.D
Total Cr, mg kg^{-1}	883	525	251	N.D

N.D Non detectable

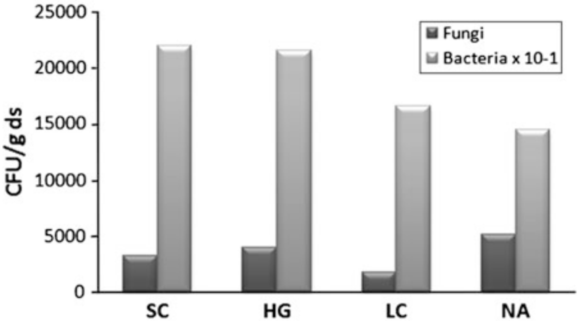


Fig. 3 Fungi and bacteria totals for flooding soils

presence of iron species which promote the ion conversion. The results shows that for these levels of pollution, soil loaded with 1.0 % of catalyst is enough to reach an efficiency about 50.0 % for Cr(VI) reduction.

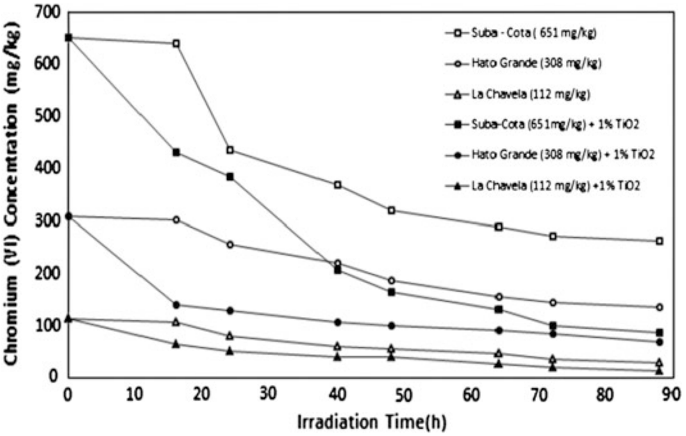


Fig. 4 Influence of Cr(VI) concentration on the photocatalytic reduction in soil after flooding in presence and absence of TiO_2 to treated soils

3.2 Effect of Load Catalyst in Cr(Vi) Reduction in Soil

The studies were development with Suba–Cota soil (651.0 mg kg^{-1}). Changes in the concentration of TiO_2 from 0.1 to 2.0 % (w/w) have not significant effect in the chromium reduction after 24 h, as shows Fig. 5.

The results show that with catalyst loading of 2.0 % promotes maximum rates reduction of Cr(VI), but not a result showing a significant difference. It is clear that before 24 h there is a significant degradation in both soils about 70.0 % less contamination.

Again, was presented the reduction of Cr(VI) in absence of photocatalyst. According to Fig. 5, catalyst load was selected as 1.0 % for the others experiments.

3.3 Effect of Alkalinity with and Without the Presence of Load Catalyst

Figure 6 shows that alkaline conditions by adding $Ca(OH)_2$ to the soil does not significant effect in Cr(VI) photoreduction when compared to natural soil.

This results shows that, the limitant mechanism for hexavalent chromium reduction using heterogeneous photocatalysis is more related to suitable conditions of mobility the metal ion to the catalyst surface, or increasing the transport of the pollutant. It's remarkable that the experiments in soil where carried out at lab scale using in situ decontamination.

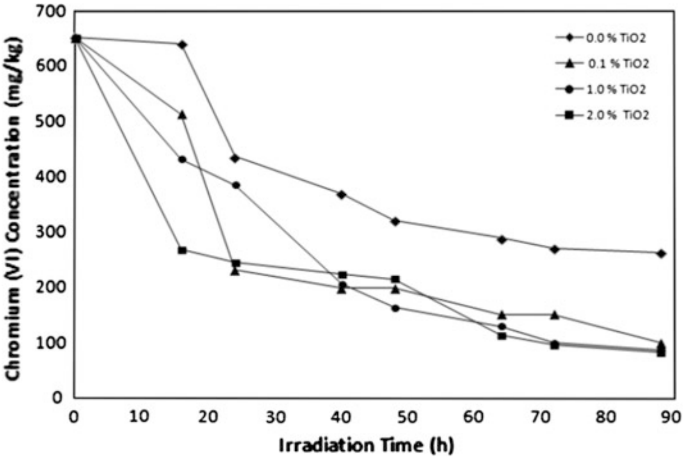


Fig. 5 Catalyst loading effect on the reduction of Cr(VI) in Suba–Cota soil after flooding (651.0 mg kg^{-1})

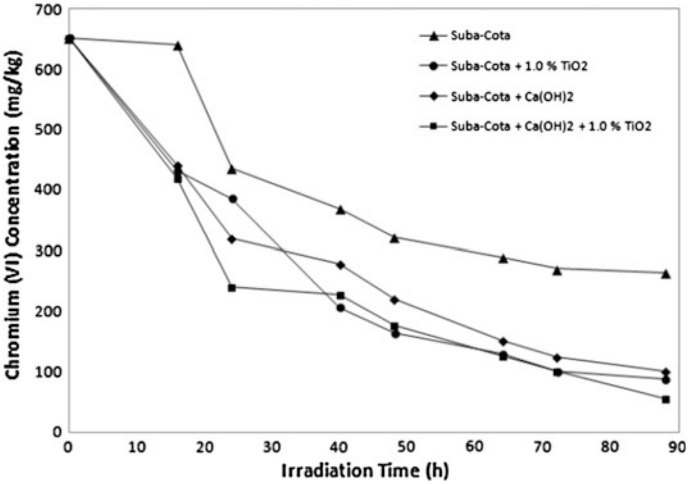


Fig. 6 Catalyst loading effect on the reduction of Cr(VI) in Suba–Cota soil after flooding (651.0 mg kg^{-1})

3.4 Depth Soil Effect

One of photocatalytic process limitants is the optical thickness. In the soil case, the process has a limited application, since it acts only in the uppermost lawyer. As shows in Fig. 7, in the profile of Cr(VI) reduction in soil contaminated (Suba–Cota) illuminated by UV radiation after 48 h of illumination, the concentration of the Cr(VI) in the top 2.0 cm of soil was about 200 mg kg^{-1} (60.0 % reduction), with a decrease through soil depth.

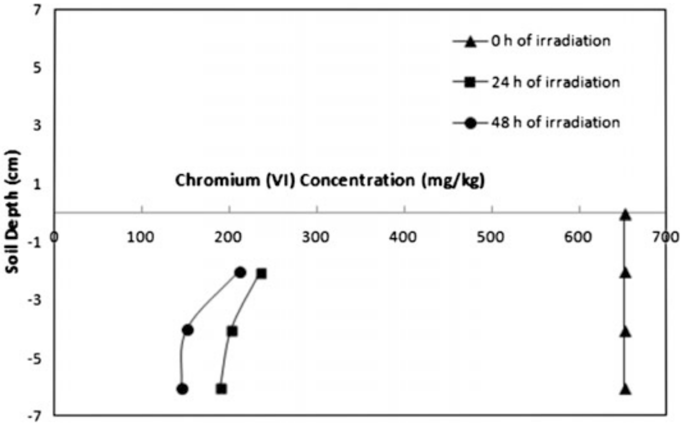


Fig. 7 Photocatalytic reduction of Cr(VI) in soil after flooding at different depths

At 4.0 cm depth, chromium reduction reached values around 80.0 % of natural concentration in contaminated soil.

4 Conclusions

The aim of this work was to assess the feasibility of using TiO_2 for in–situ soil remediation. To evaluate the influence of Cr(VI) concentration on the photocatalytic reduction in soil after flooding in presence and absence of TiO_2, different soils were used (Suba–Cota, Hato Grande and La Chavela). However to evaluate the effect of load catalyst, pH and depth soil in Cr(VI) reduction, only Suba–Cota zone soil sample was tested.

The major conclusions of this study are as follows: The results shows at these levels of Cr(VI), loading the contaminated soil with 1.0 % of catalyst was enough to reach efficiencies about 50.0 % for the reduction. In–situ process was easy and the Cr(VI) half life don't exceed 20 h under best conditions. The results show that with catalyst loading of 2.0 % promotes maximum rates reduction of Cr(VI), but not a result showing a significant difference. It is clear that before 24 h there is a significant degradation in both soils about 70.0 % less contamination.

Alkaline conditions by adding $Ca(OH)_2$ to the soil does not significant effect in Cr(VI) photoreduction when compared to natural soil. On the other hand, after 48 h of illumination the concentration of the Cr(VI) in the top 2.0 cm of soil was about 200 mg kg^{-1} (or around 60.0 % reduction), with a sharp increase in the reduction process in the soil column. At 4.0 cm depth, chromium reduction reached values around 80.0 % of natural concentration in flooding soil. At greater depths there is a remarkable transport effect, which can refer to hexavalent chromium mobility towards the surface.

References

1. A. López–Vásquez, L.N. Ramírez Q, E. Benavides–Contreras, R. López–Vásquez, In–situ photocatalytic reduction of hexavalent chromium in contaminated soil, in *Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering and Computer Science*, WCECS 2013, pp. 632–635, San Francisco, 23–25 Oct 2013
2. World Bank (2012) Integrated Urban Water Management Case Study: Bogota, Blue Water Green Cities, The World Bank, http://water.worldbank.org/laciuwm
3. C.A. Pereira, J. Tavares, G.G. Cavalcante, F.F. Vieira, Aplicação de radiação UV artificial e solar no tratamento fotocatalítico de efluentes de curtumbre. Quim. Nova **30**(5), 1082–1087 (2007)
4. Descontaminación del río Bogota, http://riitobogota.blogspot.com
5. Z. Song, C.J. Williams, R.J. Edyvean, Sedimentation of tannery wastewater. Water Res. **34**, 2171–2176 (2000)
6. M. Mwinyihija, in Ecotoxicological Diagnosis in the Tanning Industry,ed. by G.F. Marx. Main Pollutants and Environmental Impacts of the Tanning Industry, (Springer, Heidelberg, 2010), pp. 17–35
7. Secretaría Distrital de Ambiente Bogota D.C, Colombia, Resolución No 3956 de 2009, Junio 19 de 2009
8. Secretaria de Medio Ambiente y Recursos Naturales, Mexico. NORMA Oficial Mexicana NOM-147-SEMARNAT/SSA1-2004. March 2007
9. C.E. Barrera-Díaz, V. Lugo-Lugo, B. Bilyeu, A review of chemical, electrochemical and biological methods for aqueous Cr(VI) reduction. J. Hazard. Mater. **223–224**, 1–12 (2012)
10. X. Tian, X. Gao, F. Yang, Y. Lan, J.D. Mao, L. Zhou, Catalytic role of soils in the transformation of Cr(VI) to Cr(III) in the presence of organic acids containing α-OH groups. Geoderma **159**(3–4), 270–275 (2010)
11. J. Kotas, Z. Stasicka, Chromium occurrence in the environment and methods of its speciation. Environ. Pollu. **107**(3), 263–283 (2000)
12. R.J. Bartlett, J.M. Kimble, Behaviour of chromium in soils: I trivalent forms. J. Environ. Qual. **8**(1), 31–35 (1976)
13. O. Tunay, D. Orhon, I. Kabdasli, Pretreatment requirements for leather tanning industry wastewaters. Water Sci Tech **29**(9), 121–128 (1994)
14. M.M. Higarashi, W.F. Jardim, Remediation of pesticide contaminated soil using TiO_2 mediated by solar light. Catal. Today **76**(2–4), 201–207 (2002)
15. A. Assadi, M.H. Dehghani, N. Rastkari, S. Nasseri, A.H. Mahvi, Photocatalytic reduction of hexavalent chromium in aqueous solutions with zinc oxide nanoparticles and hydrogen peroxide. Environ. Prot. Eng. **38**(4), 5–16 (2012)
16. A.F. Lopez-Vasquez, J.A. Colina-Marquez, F. Machuca-Martinez, Multivariable analysis of 2,4-D herbicide photocatalytic degradation. Dyna **78**(168), 119–125 (2011)
17. X. Quan, X. Zhao, S. Chen, H.M. Zhao, J.W. Chen, Y.Z. Zhao, Enhancement of p, p-DDT photodegradation on soil surfaces using TiO2 induced by UV-light. Chemosphere **60**(2), 266–273 (2005)
18. D. Dong, P. Li, X. Li, Q. Zhao, Y. Zhang, C. Jia, P. Li, Investigation on the photocatalytic degradation of pyrene on soil surfaces using nanometer anatase TiO_2 under UV irradiation. J. Hazar. Mater. **174**(1–3), 859–863 (2010)
19. US Environmental Protection Agency (EPA), Methods for Chemical Analysis of Water and Wastes. Method 7196 A, http://www.epa.gov/osw/hazard/testmethods/sw846/pdfs/7196a.pdf

Chapter 10
An Improved Cross-Correlation Velocity Measurement Method Based on Fraction Delay Estimation

Xinggan Zhang, Yechao Bai and Xiaoli Chen

Abstract For target detection, identification and imaging, velocity estimation of high-speed targets is very important. Since the target echoes may not locate at the same range bin, and the Doppler velocity is seriously ambiguous, it is difficult to apply the traditional Doppler velocity method into estimation of high speed targets. Currently a wideband cross-correlation method is widely used to estimate the velocity. However, the measurement accuracy is limited by the sampling interval. An improved cross-correlation velocity measurement method based on the fractional delay estimation is proposed in this paper, and corresponding simulations are carried out. The results show that the estimation accuracy of the proposed method is within 1 m/s in the situation of wideband. The proposed method breaks through the limitation of integer sampling interval, as well as results in higher estimation accuracy.

Keywords Cross-correlation · Fractional delay estimation · High-speed targets · Target detection · Velocity measurement · Wideband

1 Introduction

The detection and parameter estimation of high-speed targets has become a research hotspot in radar signal processing [1–3], especially with the rapid development of ballistic missile technology and astrospace detection technology.

X. Zhang · Y. Bai (✉) · X. Chen
School of Electronic Science and Engineering, Nanjing University, Nanjing 210023,
People's Republic of China
e-mail: ychbai@nju.edu.cn

X. Zhang
e-mail: zhxg@nju.edu.cn

X. Chen
e-mail: njukelly@gmail.com

H. K. Kim et al. (eds.), *Transactions on Engineering Technologies*,
DOI: 10.1007/978-94-017-9115-1_10,

The velocity estimation of high-speed targets is becoming more and more important for target detection, identification and imaging [4–6].

As is well-known, the velocity of moving targets can be estimated by the phase and envelope of echoes [7–9]. Traditionally, the Doppler frequency measurement method is frequently used to estimate the velocity [9–11], which applies Fourier transform to the same range bin in different repetition pulse echoes to estimate the Doppler frequency. The estimation range and accuracy of the Doppler frequency are mainly determined by pulse repetition frequency and the coherent processing time. As the echo coherence decrease quickly with the changing range, high-speed targets cannot be chronically accumulated. Wideband radars are usually used for high range resolution radar [12]. However, this is hard to apply for high speed targets, since the target position may migrate through resolution cell at adjacent pulses.

Some other estimation algorithms such as the Hilbert transform correlation [13] and an estimator based on a particle filter [14] also have been proposed. A method based on straight line fitting to the phase of the cross-spectrum has been presented in [15, 16], and a generalized phase spectrum method which is an improvement of the algorithm presented in [16] has been introduced in [17], where the amplitude square weight is employed to improve the straight line fitting. Consequently, the source signal waveform must be known at first. In addition, the reason of weight choosing is not described in detail in the paper and the phase unwrapping is needed in the estimation.

In this paper, the velocity measurement method with cross-correlation is described firstly. Then considering the limitation of estimation accuracy due to the sampling interval, an improved method based on fractional delay estimation is proposed. After that, corresponding simulations are also well carried out and analyzed in this paper. Finally, the conclusion is given out.

2 The Velocity Measurement Method with Cross-Correlation

The principle of the cross-correlation method is to fully utilize the change rate of relative distance between the target and the radar to measure the velocity. The change of target characteristics in any two adjacent echoes is not significant, which results in strong correlation between signals in any two adjacent frames. The delay between the signals in any two adjacent echoes can be estimated by the maximum value position of the cross-correlation result of the two echoes. Thereby, the moving distance of the target in the interval of the two echoes can be calculated. And then the instantaneous radial velocity of the target with the pulse repetition period can be obtained. Multiple scattering points may be contained in one target, when wideband signal is employed. Scattering points in different directions have amplitude fluctuation, which is also disturbed by the clutter and the noise. Correlation utilizes the information from every scattering point comprehensively. In any two adjacent periods, the correlation coefficient of the target echoes is much larger than that of the noise and clutter. So this method has strong anti-interference ability.

The wideband radar has high range resolution, which leads to more accurate cross-correlation peak position. Therefore, the velocity estimation method with cross-correlation is particularly suitable for wideband radar.

Assuming the adjacent echoes after dechirp are $s_1(n)$ and $s_2(n)$, $n = 0, 1, 2, \ldots, N - 1$ the cross-correlation can be achieved in the frequency domain as follows:

$$\begin{aligned} R_{12}(k)_N &= \sum_{n=0}^{N-1} s_1^*(n)s_2(n+k) \\ &= IFFT[(FFT(s_1))^* \cdot FFT(s_2)] \end{aligned} \tag{1}$$

where the superscript * denotes the complex conjugate.

If the peak position of cross-correlation is k_0, the relationship between pulse repetition period and range difference of adjacent pluses is

$$\frac{2vT_r}{c} = \frac{k_0}{f_s} \tag{2}$$

where T_r denotes the pulse repetition period, and f_s denotes the sampling frequency.

The velocity can be obtained as

$$v = \frac{k_0 c}{2f_s T_r} \tag{3}$$

According to the former analyses, the error of the estimated velocity with the cross-correlation method is mainly affected by sampling frequency, pulse repetition period. In the case of high SNR, the highest estimation accuracy of cross-correlation peak position is half of the range bin. The maximum error of estimated velocity is

$$\Delta v = \frac{1}{2} \cdot \frac{1}{2} cT_s/(2T_r) = \frac{c}{8f_s \cdot T_r} \tag{4}$$

For example, when the bandwidth $B = 1$ GHz, the sampling frequency $f_s = 2$ GHz, and the pulse repetition frequency is 300 Hz, then the maximum velocity error is 5.625 m/s.

The sampling interval affects the accuracy of the peak position estimation of cross-correlation, which limits the velocity estimation accuracy. In this case, an improved cross-correlation velocity measurement method that based on the fractional delay estimation is proposed to break through the sampling interval limitation.

3 Improved Velocity Measurement Method Based on Fraction Delay Estimation

In general, to reduce the measuring error caused by range quantization, interpolation which realizes fractional delay estimation is usually applied. However, this method suffers heavy computational loads, and is also unpractical for engineering realization.

To be more practical, we have developed a fractional delay estimation method in [18]. The key idea is that the fractional part of the time delay is calculated based on weighted straight line fitting to the cross-spectrum phase. The fractional delay estimation method is utilized to improve the estimation accuracy of targets velocity.

Suppose the two adjacent echo signals are respectively as follows

$$\begin{aligned} r_1(i) &= s(i) + n_1(i) \\ r_2(i) &= s(i-D) + n_2(i), i = 0, 1, \ldots N-1 \end{aligned} \tag{5}$$

The parameter D represents the fractional delay to be estimated between the two echoes. $s(i)$ denotes the discrete sample of target signal. $n_1(i)$ and $n_2(i)$ are zero-mean stationary Gaussian processes with variances of σ^2, being uncorrelated with each other and with $s(i)$. N is the number of sampling. And without loss of generality, let N be even.

Let $R_1(k)$ and $R_2(k)$ be the discrete Fourier transform of $r_1(t)$ and $r_2(t)$ respectively, and their cross-spectrum can be expressed as

$$\begin{aligned} R_1(k)R_2^*(k) &= |S(k)|^2 \exp\left(j\frac{2\pi}{N}kD\right) + S(k)N_2^*(k) + S^*(k)N_1(k)\exp\left(j\frac{2\pi}{N}kD\right) + N_1(k)N_2^*(k) \\ &= |S(k)|^2 \exp\left(j\frac{2\pi}{N}kD\right) + W(k), \\ k &= -N/2+1, -N/2+2, \ldots N/2 \end{aligned} \tag{6}$$

where the superscript * denotes complex conjugate, $S(k)$, $N_1(k)$, $N_2(k)$ are the Fourier transforms of $s(i)$, $n_1(i)$, and $n_2(i)$ respectively, and the noise component is

$$W(k) = S(k)N_2^*(k) + S^*(k)N_1(k)\exp\left(j\frac{2\pi}{L}kD\right) + N_1(k)N_2^*(k) \tag{7}$$

From (6) we can see that the time delay appears in the cross-spectrum as a phase function

$$\varphi(k) = \frac{2\pi}{N}kD \tag{8}$$

The phase is a linear function of the time delay *D*. If there is noise, the cross-spectrum phase fluctuates along the line. Then the time delay can be estimated by fitting a straight line to the cross-spectrum phase. Considering the symmetric cross-spectrum of real functions, the positive frequency part is taken into account only. According to the least squares approach, *D* is estimated as

$$\hat{D} = \frac{N}{2\pi} \frac{\sum_{k=1}^{N/2-1} \Phi(k)k}{\sum_{k=1}^{N/2-1} k^2} \tag{9}$$

where Φ(*k*) is the phase of cross-spectrum.

Each phase contributes the same in (9). Nevertheless, the SNR of cross-spectrum is different at different frequencies, thus it is more reasonable to give different weights when fitting.

The cross-spectrum is normalized firstly to find the weights for straight line fitting as

$$\frac{R_1(k)R_2^*(k)}{|S(k)|^2} = \exp\left(j\frac{2\pi}{N}kD\right) + \frac{W(k)}{|S(k)|^2}, \quad k = 1, 2, \ldots \frac{N}{2} - 1 \tag{10}$$

$N_1(k)$ and $N_2(k)$, $k = 1, 2, \ldots, N/2 - 1$ follow independent and identical zero-mean complex Gaussian distribution with variance $N\sigma^2$. With high SNR, the last item of *W*(*k*) can be omitted. Then *W*(*k*) becomes zero-mean complex Gaussian distribution with variance $2\ N|S(k)|^2\sigma^2$.

Now, we build a model from (10) to find the weights. Since exp(j2π*kD*/*N*) has been used to fit a straight line and the task is to find the weights now, exp(j2π*kD*/*N*) can be replace with a constant *A* which we want to estimate by the best linear unbiased estimator (BLUE). The model is shown as

$$x(k) = A + W_n(k), k = 1, 2, \ldots N/2 - 1 \tag{11}$$

where *x*(*k*) and $W_n(k)$ denote the item at the left side of equal in (10) and the last item of (10) respectively. $W_n(k)$ follows zero-mean complex Gaussian distribution with variance $\sigma_w(k)^2 = 2\ N\sigma^2/|\mathrm{S(k)}|^2$. The best linear unbiased estimator (BLUE) for *A* is

$$\hat{A} = \frac{\sum_{k=1}^{N/2-1} \frac{x(k)}{\sigma_w(k)^2}}{\sum_{k=1}^{N/2-1} \frac{1}{\sigma_w(k)^2}} \tag{12}$$

The weight of the point with the smallest variance is the largest. If the variances for $x(k)$ at each point are the same, it is clear that the estimator of A is the averaging of $x(k)$, and the weights for $x(k)$ are all $1/(N/2-1)$. Therefore, the weights caused by different variances are

$$p(k)=\frac{\frac{1}{\sigma_w(k)^2}(\frac{N}{2}-1)}{\sum\limits_{k=1}^{N/2-1}\frac{1}{\sigma_w(k)^2}}=\frac{|S(k)|^2}{\sum\limits_{k=1}^{N/2-1}|S(k)|^2}(\frac{N}{2}-1),\\ k=1,2,\ldots\frac{N}{2}-1 \tag{13}$$

These weights act to pre-whiten $x(k)$ before averaging.

The estimated time delay D is all contained in A in the model, and A, which equals $\exp(j2\pi kD/N)$, is exactly used to fit the straight line. So the weights shown in (13) can be used in the straight line fitting to the cross-spectrum phase. The weighted least squares approach is used to estimate D. The least squares error is

$$\hat{D}=\frac{N}{2\pi}\frac{\sum\limits_{k=1}^{N/2-1}p(k)\Phi(k)k}{\sum\limits_{k=1}^{N/2-1}p(k)k^2}=\frac{N}{2\pi}\frac{\sum\limits_{k=1}^{N/2-1}\Phi(k)|S(k)|^2k}{\sum\limits_{k=1}^{N/2-1}|S(k)|^2k^2} \tag{14}$$

From (8), it is clear that the phase will wrap for $k=1,2,\ldots,N$ if D is larger than 1. Even the positive frequency part is employed only, D cannot be larger than 2 to avoid wrapping. The integer part of time delay can be estimated by locating the maximum position of cross-correlation first. Consequently, the fractional part can be estimated by weighted straight line fitting without wrapping. Also, it is hard to obtain $|S(k)|^2$ directly, so the absolute value of the cross-spectrum is used instead. Referring to (6), it is seen that this similarity is advisable when SNR is not very low.

4 Simulation Results and Analyses

To compare the performance of the proposed algorithm to that of other existing algorithms, a series of Matlab simulations have been conducted as follows.

Simulation parameters are set as follows: the center frequency of radar signal $f_0=35$ GHz, the pulse width $T=10\mu s$, the pulse repetition frequency $PRF=300$ Hz. The bandwidth B varies between 20 MHz-1 GHz, the sampling frequency $f_s=2B$, and the $SNR=0$ dB. Besides, the target distance $r_0=100$ km, the velocity $v=14$ km/s.

The traditional cross-correlation method and the improved method are simulated and compared. The results are showed in detail in Figs. 1 and 2 as follows.

Fig. 1 The estimated velocity of traditional cross-correlation method and the improved cross-correlation method

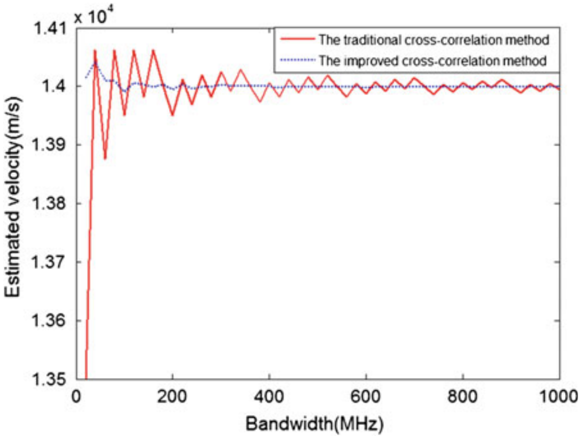
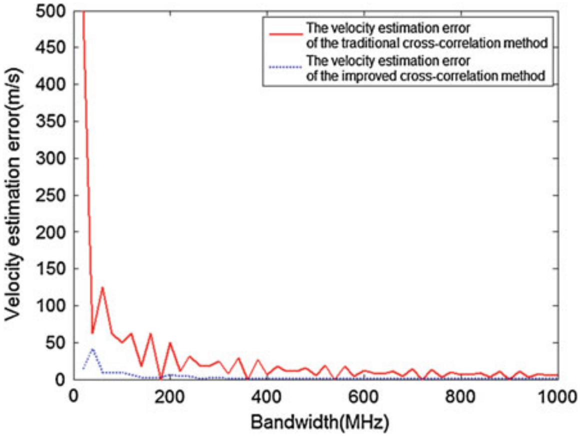


Fig. 2 The velocity estimation error of traditional cross-correlation method and the improved cross-correlation method

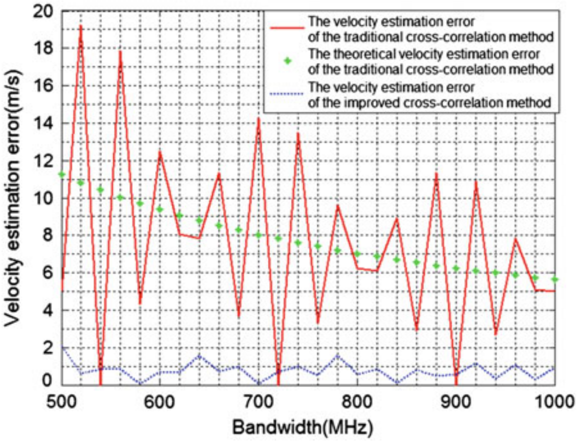


Figures 1 and 2 show that the velocity estimation method with cross-correlation is more suitable for wideband signals. The estimating accuracy is much higher in the condition of wideband radar. Comparing the two methods, the proposed method can improve the estimation accuracy of cross-correlation peak position by breaking through the limitation of the integer range bin. When the bandwidth varies from 500 MHz to 1 GHz, the details about velocity measurement error of the two methods are shown in Fig. 3.

According to the above theoretical analyses, the mean cross-correlation velocity estimation error is

$$\int_{-1/2f_s}^{1/2f_s} \left|\frac{c\triangle}{2T_r}\right| f_s d\varDelta = \frac{c}{8f_s \cdot T_r} \tag{15}$$

Fig. 3 The velocity estimation error of the two methods when bandwidth varies from 500 MHz to 1 GHz



As can be seen in Fig. 3, the trend of the velocity estimation error is consistent with theoretic analyses. However, the velocity estimation error of the improved method is smaller by breaking through the limitation of integer range bin. The velocity estimation error is within 1 m/s in the condition of wideband.

The utilized fractional delay estimation method approaches the Cramer-Rao lower bound for wideband signals [18]. The performance of velocity estimation is improved by applying the fractional delay estimation method.

5 Conclusion and Future Work

In this paper, an improved method, which is based on the fractional delay estimation with weighted linear fitting of the phase of cross power spectrum, is proposed. It can improve the estimation accuracy of the velocity measurement method with cross-correlation by breaking through the limitation of integer range bin. Meanwhile, corresponding simulations and analyses of both the traditional and proposed methods are accomplished in this paper. The results show that, in the situation of wideband, the velocity estimation accuracy of the proposed method is within 1 m/s, and the velocity estimation accuracy can be well improved with wider bandwidth.

Acknowledgments This work was supported by Industry-academic Joint Technological Innovations Fund Project of Jiangsu Province (BY2012187).

References

1. X. Zhang, Y. Bai, X. Chen, An improved cross-correlation method based on fractional delay estimation for velocity measurement of high speed targets, in *Lecture Notes in Engineering and Computer Science*: *Processing of The World Congress on Engineering and Computer Science 2013, WCECS 2013, San Francisco, USA,* pp. 651–654, 23–25 Oct 2013

2. C.S. Pang, T. Ran, A High Speed Target Detection Approach Based on STFrFT, in *IEEE International Conference on Instrumentation, Measurement, Computer, Communication and Control*, pp. 744–747 (2011)
3. S.L. Wen, Q. Yuan, Velocity measurement technique for high-speed targets based on digital fine spectral line tracking. IEEE J. Syst. Eng. Electron. **17**(1), 6–12 (2006)
4. R. Axelsson, J. Sune, Estimation of target position and velocity using data from multiple radar stations, in *IEEE International Conference on Geoscience and Remote Sensing Symposium*, vol. 7, pp. 4140–4143 (2003)
5. J.J. Chen, J. Chen, S.L. Wang, Detection of ultra-high speed moving target based on matched fourier transform, in *IEEE International Conference on Radar*, pp. 1–4 (2006)
6. T. Thayaparan, L. Stankovic, C. Wernik, M. Dakovic, Real-time motion compensation, image formation and image enhancement of moving targets in ISAR and SAR using S-method-based approach. IET Signal Proc. **2**(3), 247–264 (2008)
7. Y. Liu, H. Meng, G. Li, X. Wang, Range-velocity estimation of multiple targets in randomised stepped-frequency radar. Electron. Lett. **44**(17), 1032–1034 (2008)
8. A. Budillon, V. Pascazio, G. Schirinzi, Estimation of radial velocity of moving targets by along-track interferometric SAR Systems. IEEE Geosci. Remote Sens. Lett. **5**(3), 349–353 (2008)
9. A. Ludloff, M. Minker, Reliability of velocity measurement by MTD radar. IEEE Trans. Aerosp. Electron. Syst. **AES-21**(4), 522–528 (1985)
10. L.C. Perkins, H.B. Smith, D.H. Mooney. The development of airborne pulse Doppler radar. IEEE Trans. Aerosp. Electron. Syst. **AES-20**(3), 292–303 (1984)
11. W.D. Hu, H.J. Sun, X. Lv, S.Y. Li et al. Stepped frequency millimeter-wave signal ISAR processing, in *Global Symposium on Millimeter Wave*, pp. 63–65 (2008)
12. X.Z. Wei, R. Zhang, B. Deng, A recognition algorithm for high voltage transmission lines at horizontal polarization millimeter-wave radar, in *ICIA International Conference on Information and Automation*, pp. 1172–1175 (2008)
13. A. Grennberb, M. Sandell, Estimation of subsample time delay differences in narrowband ultrasonic echoes using the hilbert transform correlation. IEEE Trans. Ultrason. Ferroelectr. Freq. Control **41**(5), 588–595 (1994)
14. E.A. Lehmann, Particle filtering approach to adaptive time-delay estimation, in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, Toulouse, France, pp. 1129–1132 (2006)
15. V. Mosorov, Phase spectrum method for time delay estimation using twin-plane electrical capacitance tomography. Electron. Lett. **42**(11), 630–632 (2006)
16. A. Piersol, Time delay estimation using phase data. IEEE Trans. Acoust. Speech Signal Process. **29**(3), 471–477 (1981)
17. Z. Zhao, Z.Q. Hou, The generalized phase spectrum method for time delay estimation. ACTA ACUSTICA **10**(4), 201–215 (1985) (in Chinese)
18. Y. Bai, X. Zhang, X. Qiu, Subsample time delay estimation based on weighted straight line fitting to cross-spectrum phases. Chin. J. Electron. **19**(4), 553–556 (2010)

Chapter 11
A New Common Subexpression Elimination Algorithm for Constant Matrix Multiplications Over Binary Field

Ning Wu, Xiaoqiang Zhang, Yunfei Ye and Lidong Lan

Abstract In this work, a new multi-term common subexpression elimination (CSE) algorithm is proposed. The new algorithm aims to reduce area-delay-production (ADP) in VLSI designs of constant matrix multiplication (CMM) over binary field. For promoting delays optimization, a gate-level delay computing method is used to compute the delays based on the transformed constant matrices. The new algorithm also takes a greedy algorithm to search the minimal ADP result. The worst case computational complexities of the delay computing method and the new CSE algorithm are analyzed, respectively. Experimental results have shown that the new CSE algorithm has more efficient in ADP reduction in VLSI designs of binary CMM.

Keywords Area-delay-production · Binary field · Common subexpression elimination · Constant matrix multiplication · Critical path delay · Delay computing method · Greedy algorithm · Multiple constants multiplication

N. Wu (✉) · X. Zhang · Y. Ye · L. Lan
College of Electrical and Information Engineering, Nanjing University
of Aeronautics and Astronautics, Nanjing 210016, China
e-mail: wunee@nuaa.edu.cn

X. Zhang
e-mail: zxq198111@qq.com

Y. Ye
e-mail: yunfye@nuaa.edu.cn

L. Lan
e-mail: lanlidongdt@126.com

H. K. Kim et al. (eds.), *Transactions on Engineering Technologies*,
DOI: 10.1007/978-94-017-9115-1_11,

1 Introduction

Constant matrix multiplications (CMM) over binary field are widely used in cryptography and coding theory. In high level synthesis of VLSI design, the optimization of binary CMM can lead to significant improvements in various design parameters like area or power consumption. Common subexpression elimination (CSE) is an optimization procedure that is often used in solving these problems. The idea of CSE is to identify patterns (common subexpressions) that are present in expressions more than once and replace them with a single variable. With this, each of these patterns needs only to be computed once, thus reducing the size of the hardware required in VLSI implementation. However, how to select a pattern to eliminate for achieving optimal results is an NP-complete problem [1]. Over the years many CSE algorithms have been introduced for the multiplierless implementation of binary CMM.

Both area and throughput are significant for VLSI design. However, most of previous CSE algorithms only focused on the area reducing without providing complete control on the delays, which ultimately determines throughput. The straightforward realization of the binary CMM has the shortest critical path delay (CPD) when it is constructed by using complete binary-trees of adders [2]. In this case, the gate-level delay is easy to be determined by

$$d_i = \lceil \log_2 k_i \rceil \tag{1}$$

where $\lceil x \rceil$ represents the smallest integer larger than or equal to x, and k_i is the number of input variables in the expression of the output variable y_i. The CPD of the binary CMM circuit is the maximum of $\{d_i\}$.

Most of CSE algorithms try to reduce the area complexity by transforming the constant matrices. Thus, the first challenge for the CPD control is how to compute the delay according to the transformed constant matrices. We have proposed a delay computing method based on a delay matrix in [3]. In this paper, the worst case computational complexity of the delay computing method is analyzed. Based on the delay computing method, a new CSE is proposed. The new algorithm also employs a greedy algorithm to search the best set of patterns with the minimal area-delay-product (ADP), so as to optimize both the area and the delays in binary CMM blocks.

2 The Basic Principles of CSE Algorithm

Binary CMM can be expressed as a linear transform $\boldsymbol{Y} = \boldsymbol{MX}$, where $\boldsymbol{Y}$ and $\boldsymbol{X}$ are n- and m-dimensional binary column vectors, respectively, and $\boldsymbol{M}$ is an $n \times m$ binary constant matrix [1]. Clearly, such a transform requires only additions. Because additions over binary field are referring to XOR operation, and hence binary CMM can be implemented by a combinational logic with XOR gates only. The linear transform $\boldsymbol{Y} = \boldsymbol{MX}$ can also be expressed in bit-level expressions. A CSE algorithm can be exploited to extract the common factors in all the bit-level expressions, in order to reduce the area cost of combinational logic implementation. In general, any CSE algorithm involves the following general steps [3]:

1. Identify patterns present in the expressions.
2. Select a pattern for elimination.
3. Remove all the occurrences of the selected pattern.
4. The eliminated pattern is computed only once.
5. Repeat Step 1–4 until none recurrent pattern is present.

CSE algorithm can be divided into two categories according to the number of term in pattern [4]: *multi-term-pattern CSE algorithm* and *two-term-pattern CSE algorithm*. Since the *two-term* pattern can be regard as a special case of *multi-term* pattern, we take *multi-term* pattern into account in this paper. In [5], a *multi-term-pattern* CSE algorithm was proposed to select the *most-term* pattern with highest occurring frequency to eliminate at each iteration. While this algorithm was not explicitly proposed for computations over the binary field. Zeng et al. [6] extended it by using a greedy algorithm to check all candidate patterns, and used it to reduce the area in VLSI designs of the binary CMM in AES S-box implementation.

Consider the computation as

$$\begin{bmatrix} y_0 \\ y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} x_0 \\ x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} x_0 + x_1 + x_2 + x_3 + x_4 \\ x_0 + x_1 + x_3 + x_4 \\ x_0 + x_1 \\ x_0 + x_1 + x_3 \\ x_2 \end{bmatrix} \tag{2}$$

The computation in Eq. (2) can be optimized by the CSE algorithm in [6], and the optimization process is

$$\begin{aligned}
&\begin{bmatrix} \boxed{1} & \boxed{1} & 1 & \boxed{1} & \boxed{1} \\ \boxed{1} & \boxed{1} & 0 & \boxed{1} & \boxed{1} \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{bmatrix}
\xrightarrow{z_0 = x_0 + x_1 + x_3 + x_4}
\begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ \boxed{1} & \boxed{1} & 0 & \boxed{1} & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ \boxed{1} & \boxed{1} & 0 & \boxed{1} & 1 & 0 \end{bmatrix} \\
&\xrightarrow{z_1 = x_0 + x_1 + x_3}
\begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ \boxed{1} & \boxed{1} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ \boxed{1} & \boxed{1} & 0 & 1 & 0 & 0 & 0 \end{bmatrix} \\
&\xrightarrow{z_2 = x_0 + x_1}
\begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}
\end{aligned} \tag{3}$$

The CSE algorithm eliminates the *most-term* pattern with highest occurring frequency at each iteration. It took three iterations in this example. Patterns $x_0 + x_1 + x_3 + x_4$, $x_0 + x_1 + x_3$ and $x_0 + x_1$ (denoted by squares in Eq. 3) are selected to eliminate at these three iterations, respectively, and they are replaced by new variables z_0, z_1 and z_2, respectively. The eliminated patterns are appended to the bottom as new expressions to be further optimized.

A straightforward realization of the linear transformation requires 10 XOR gates. However, there are only 4 XOR gates required after optimization by the CSE algorithm. The reduction of XOR gates can be up to 60 %.

3 The Proposed Delay Computing Method

We proposed a gate-level delay computing method in [3] to compute the delays for the transformed matrices. For clarity, the details of the method are rewrite as follows.

The proposed delay computing method consists of three steps

1. Transform the primitive constant matrix to initial delay matrix.
2. Update the delay matrix according to each transformation of the constant matrix.
3. Determine the CPD based on the delay matrix.

3.1 Initial Delay Matrix Establishment

In a linear transform $\boldsymbol{Y} = \boldsymbol{MX}$, a row of binary constant matrix $\boldsymbol{M}$ corresponds to an output variable and a column corresponds to an input variable. Assuming the delay for input x_i is 0, and we use -1 to represent the invalid delay. Then the initial delay matrix $\boldsymbol{M_d}$ can be obtained from constant matrix $\boldsymbol{M}$ by

$$\mathbf{M_d} = \mathbf{M} - \mathbf{M_1} \tag{4}$$

where $\boldsymbol{M_1}$ is a matrix that has the same dimension as $\boldsymbol{M}$ and all of its elements are 1. For example, the initial delay matrix of the computation in Eq. (2) can be computed as

$$\begin{aligned} M_d &= \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{bmatrix} - \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix} \\ &= \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & -1 & 0 \\ 0 & 0 & -1 & -1 & -1 \\ -1 & -1 & 0 & -1 & -1 \end{bmatrix}. \end{aligned} \tag{5}$$

3.2 Update Process of Delay Matrix

The delay matrix will be updated accordingly as the constant matrix $\boldsymbol{M}$ transformed by the CSE algorithm. The delay matrix update process consists of the following two steps at each iteration:

1. Compute the delay of the new variable.
2. Update the delay matrix according to the transformed constant matrix.

3.2.1 Delay Computation

Suppose that a pattern $x_i + x_j + \cdots + x_k$ is replaced by a new variable z_{new} and appended to the bottom of constant matrix as an additional row at an iteration of CSE algorithm. Then the delay of z_{new} is can be obtained by using **Algorithm 1**.
ALGORITHM 1: COMPUTE DELAY

1. Input a row of delay d_{xi} d_{xj} … d_{xk}
2. Sort the row in an increasing order.
3. Select the smallest two valid delays d_s and d_{s+1}.
4. Update d_s and d_{s+1} by

$$\begin{cases} d_{s+1} = \max(d_s,\ d_{s+1}) + 1 \\ d_s = -1 \end{cases} \tag{6}$$

5. Repeat steps 2-4 until there is only one valid delay in the row.
6. Output the valid delay as d_{znew}.

Where d_{xi}, d_{xj}… d_{xk} and d_{znew} are the delays of x_i, x_j… x_k and z_{new}, respectively.

3.2.2 Delay Matrix Updating

The updated process of d_{znew} in the delay matrix is described in **Algorithm 2**.
ALGORITHM 2: UPDATE DELAY

1. Append new row with d_{xi}, d_{xj}… d_{xk} according to the transformation of constant matrix $\boldsymbol{M}$.
2. Compute the delay d_{znew} by **Algorithm 1** with the appended row.
3. Append new column with d_{znew} according to the transformation of constant matrix $\boldsymbol{M}$.
4. If the new variable z_{new} is an input of previous appended variable z_p, then re-update the delay d_{zp} by taking Step 2 and Step 3, then go to Step 5. If not, finish updating.
5. Take variable z_p as z_{new} to repeat Step 4.

3.2.3 Examples

We take the linear transform in Eq. (2) as a example to illustrate the update process of delay matrix. The updated process of constant matrix $\boldsymbol{M}$ of the linear transform is shown in Eq. (3) and the initial delay matrix can be got from Eq. (5).

We update the delay matrix in Eq. (5) by using **Algorithm 2**, and the corresponding updated process is expressed as

$$\begin{aligned}
&\begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & -1 & 0 \\ 0 & 0 & -1 & -1 & -1 \\ -1 & -1 & 0 & -1 & -1 \end{bmatrix}
\xrightarrow{\text{update d}_{z0}}
\begin{bmatrix} -1 & -1 & 0 & -1 & -1 & 2 \\ -1 & -1 & -1 & -1 & -1 & 2 \\ 0 & 0 & -1 & -1 & -1 & -1 \\ 0 & 0 & -1 & 0 & -1 & -1 \\ -1 & -1 & 0 & -1 & -1 & -1 \\ 0 & 0 & -1 & 0 & 0 & -1 \end{bmatrix} \\
&\xrightarrow{\substack{\text{update d}_{z1} \\ \text{re}-\text{update d}_{z0}}}
\begin{bmatrix} -1 & -1 & 0 & -1 & -1 & 3 & -1 \\ -1 & -1 & -1 & -1 & -1 & 3 & -1 \\ 0 & 0 & -1 & -1 & -1 & -1 & -1 \\ -1 & -1 & -1 & -1 & -1 & -1 & 2 \\ -1 & -1 & 0 & -1 & -1 & -1 & -1 \\ -1 & -1 & -1 & -1 & 0 & -1 & 2 \\ 0 & 0 & -1 & 0 & -1 & -1 & -1 \end{bmatrix} \\
&\xrightarrow{\substack{\text{update d}_{z2} \\ \text{re}-\text{update d}_{z1} \\ \text{re}-\text{update d}_{z0}}}
\begin{bmatrix} -1 & -1 & 0 & -1 & -1 & 3 & -1 & -1 \\ -1 & -1 & -1 & -1 & -1 & 3 & -1 & -1 \\ -1 & -1 & -1 & -1 & -1 & -1 & -1 & 1 \\ -1 & -1 & -1 & -1 & -1 & -1 & 2 & -1 \\ -1 & -1 & 0 & -1 & -1 & -1 & -1 & -1 \\ -1 & -1 & -1 & -1 & 0 & -1 & 2 & -1 \\ -1 & -1 & -1 & 0 & -1 & -1 & -1 & 1 \\ 0 & 0 & -1 & -1 & -1 & -1 & -1 & -1 \end{bmatrix}
\end{aligned} \tag{7}$$

At the first iteration, the pattern $x_0 + x_1 + x_3 + x_4$ is replaced by a new variable z_0 in the constant matrix $\boldsymbol{M}$. Then the delays of x_0, x_1, x_3, and x_4 are appended to the bottom of delay matrix as a new row. By using **Algorithm 1**, we can get that the delay of z_0 is 2 adder delays. According to the transform of constant matrix $\boldsymbol{M}$ in Eq. (3), the delay of z_0 is appended to the right delay matrix as a new column.

At the second iteration, the pattern $x_0 + x_1 + x_3$ is replaced by z_1. The updated process of delay matrix is similar to the one at first iteration, except re-updating d_{z0}. Because variable z_1 is an input of z_0 at this iteration, the delay d_{z0} should be re-updated. From Eq. (3), we can get that the appended variable z_0 has been changed to $z_0 = x_4 + z_1$ at the second iteration. Then d_{z0} depends on d_{x4} and d_{z1} at the second iteration. By using **Algorithm 1**, we can get the new d_{z0} is 3 adder delays.

At the last iteration, the updated process of delay matrix is similar to the one at second iteration.

3.3 Determine the CPD

When the optimization process of the CSE algorithm is finished, the delay of each output variable can be determined by using **Algorithm 1**. Consider the linear transform in Eq. (2), the delays of output variables are

$$\begin{bmatrix} -1 & -1 & 0 & -1 & -1 & 3 & -1 & -1 \\ -1 & -1 & -1 & -1 & -1 & 3 & -1 & -1 \\ -1 & -1 & -1 & -1 & -1 & -1 & -1 & 1 \\ -1 & -1 & -1 & -1 & -1 & -1 & 2 & -1 \\ -1 & -1 & 0 & -1 & -1 & -1 & -1 & -1 \end{bmatrix} \xrightarrow{Algorithm\ 1} \begin{bmatrix} -1 & -1 & -1 & -1 & -1 & -1 & -1 & 4 \\ -1 & -1 & -1 & -1 & -1 & -1 & -1 & 3 \\ -1 & -1 & -1 & -1 & -1 & -1 & -1 & 1 \\ -1 & -1 & -1 & -1 & -1 & -1 & -1 & 2 \\ -1 & -1 & -1 & -1 & -1 & -1 & -1 & 0 \end{bmatrix} \tag{8}$$

From Eq. (8), we can get that the CPD of the optimized implementation structure for Eq. (3) is 4 adder delays and the delay of y_0 is the maximum delay. By using Eq. (1), we can obtain that the CPD of the straightforward realization of binary CMM in Eq. (2) is equal to 2 adder delays. Instead of using Eq. (1), the CPD of the straightforward realization can also be determined by using **Algorithm 1** based on the initial delay matrix $\boldsymbol{M}_d$ in Eq. (5).

3.4 Computational Complexities Analysis

The input vectors for the proposed delay computing method are the constant matrix $\boldsymbol{M}$ and its transformed forms. Hence the computational complexities of proposed delay computing method depend on the dimensions of $\boldsymbol{M}$ only. Suppose the constant matrix $\boldsymbol{M}$ is an $n \times m$ binary matrix, then the worst case computational complexities of proposed delay computing method are analyzed in the following.

1. The establishment of initial delay matrix in Step 1 of the method requires nm comparisons, then the computational complexity is $O(nm)$.
2. In Step 2 of the method, **Algorithm 2** is used to update the delay matrix, so the computational complexity for step 2 of the method is equal to the computational complexity of **Algorithm 2**. **Algorithm 1** is used to compute the delay in Step 2 of **Algorithm 2**. In the following, we firstly analyze the computational complexity of **Algorithm 1**.

   (1) The input vectors for **Algorithm 1** are the rows of transformed delay matrices. At the s-th iteration, the dimension of input rows is $s + m - 1$,

The computational complexity of sorting operations for a row is at most $O((s + m - 1)^2)$ in Step 2. In Step 3, it requires at most $s + m - 1$ comparisons to find the smallest valid delay. The computational for delay computing in Step 4 is negligible. There are at most m valid delays in a row, and then the complexity for **Algorithm 1** is

$$O\Big(m\Big((s+m-1)^2+(s+m-1)\Big)\Big) \approx O(ms^2 + 2m^2 s + m^3) \tag{9}$$

(2) Next, the computational complexity of **Algorithm 2** is analyzed as follow. The input vectors for **Algorithm 2** are the transformed constant matrices and the transformed delay matrices. At the s-th iteration, the dimensions of both the transformed constant matrices and the transformed delay matrices are $(s + m - 1) \times (s + m - 1)$. Since there are at most n rows that include the selected patterns and the selected pattern has at most m terms, the computational complexity for eliminating and appending the delays in the Step 1 is at most $O(nm)$. The computation of delay in Step 2 relies on **Algorithm 1**, so the computational complexity for Step 2 is $O(ms^2 + 2m^2 s + m^3)$. The computational complexity for appending new column in Step 3 is $O(n + s)$. There are at most $s - 1$ delays of previous appended variables may be re-updated in Step 4 and Step 5. Then the computational complexity for **Algorithm 2**, for Step 2 of the method also, is at most

$$\begin{aligned} &O(nm + s((ms^2 + 2m^2 s + m^3) + (n + s))) \\ &\approx O(ms^3 + 2m^2 s^2 + m^3 s + ns + mn) \end{aligned} \tag{10}$$

3. In Step 3 of the method, **Algorithm 1** is used to compute the final delays of each output variables with the first n rows of transformed delay matrix. Suppose there are at most S patterns identified by the CSE algorithm, so the dimension of input rows is $S + m - 1$. Therefore the computational complexity is

$$O(n(mS^2 + 2m^2 S + m^3)) = O(nmS^2 + 2m^2 nS + m^3 n) \tag{11}$$

- In summary, the computational complexity for the proposed method is

$$\begin{aligned} &O(mn) + \sum_{s=1}^{S} O\big(ms^3 + 2m^2 s^2 + m^3 s + ns + mn\big) + O\big(nmS^2 + 2m^2 nS + m^3 n\big) \\ &\approx O\left(\frac{mS^4}{4} + \frac{2m^2 S^3}{3} + \frac{m^3 S^2}{2} + nmS^2 + 2m^2 nS + m^3 n\right) \end{aligned} \tag{12}$$

The number of identified patterns S is related to the dimensions of the constant matrix. For an $n \times m$ matrix, there are at most $nm/2$ patterns identified by a CSE algorithm [1]. Then computational complexity of proposed delay computing method is

$$O\left(\frac{mS^4}{4} + \frac{2m^2S^3}{3} + \frac{m^3S^2}{2} + nmS^2 + 2m^2nS + m^3n\right)_{S=\frac{nm}{2}} \approx O\left(\frac{n^4m^5}{64}\right) \quad (13)$$

4 The New CSE Algorithm

Based on the proposed gate-level delay computing method, a new CSE algorithm is proposed in this section to optimize both the area and the delay in implementation of binary CMM. In the new CSE algorithm, the *most-term* patterns with the highest frequency of occurrence are selected to eliminate, and a greedy algorithm to check all possible patterns. Obviously, the proposed algorithm is similar to the one in [6]. The difference is that the greedy algorithm is used to find out the minimal ADP, while in [6], the greedy algorithm is used to find out minimal area. The details of proposed CSE algorithm are summarized in **Algorithm 3**.
ALGORITHM 3: NEW CSE ALGORITHM

1. Establish the initial delay matrix based on the primitive constant matrix.
2. Compute the occurrence frequency of N-term patterns based on the constant matrix.
3. Establish a list for the N-term patterns with the highest frequency.
4. Select a pattern from the highest frequency list to eliminate.
5. Update the constant matrix and the delay matrix.
6. Repeat Steps 2–5 for the next iteration until no recurring N-term patterns exist.
7. Let $N = N - 1$, repeat Step 2–6 until $N < 2$.
8. Compute the ADP value based on the transformed constant matrix and the transformed delay matrix.
9. Select a different set of patterns from the highest frequency list to eliminate by repeating Steps 2–8, until all sets of patterns have been checked.
10. Find out the minimal ADP value.

Where the initial value of N is the number of input variables.

4.1 Worst Case Computation Complexities Analysis

The computational complexities of **Algorithm 3** only depend on the dimensions of $\boldsymbol{M}$. Suppose the dimensions of constant matrix $\boldsymbol{M}$ are $n \times m$, then the worst case computational complexities of **Algorithm 3** are analyzed in the following.

1. As mention above, the computational complexity for Step 1 in the algorithm is $O(nm)$.
2. At s iteration, the dimensions of both the transformed constant matrices and transformed delay matrices are $(s + n - 1) \times (s + m - 1)$. Since there are at most C_m^N patterns that need to compute the occurrence frequency in Step 2, then the computational complexity for Steps 2 is $O((s + n - 1) \times (m!/(N! \times (m - N)!)))$.
3. The computational complexity for Step 3 is $O(m!/(N! \times (m - N)!))$.
4. Because there are no complicate operations in Step 4, so the computational complexity for it can be neglected.
5. In Step 5, the computational complexity for update of constant matrix and delay matrix are $O(s + n - 1)$ and $O(ms^3 + 2m^2s^2 + m^3s + ns + mn)$, respectively.

- Steps 2–5 are used to eliminate a pattern at an iteration. Then total worst case computational complexity for a pattern elimination in Step 2–5 is

$$\begin{aligned} &O\left((s+n-1)\frac{m!}{(m-N)!N!}\right) + O\left(\frac{m!}{(m-N)!N!}\right) + O(s+n-1) \\ &\qquad + O\left(ms^3 + 2m^2s^2 + m^3s + ns + mn\right) \\ &\qquad \approx O\left((s+n-1)\frac{m!}{(m-N)!N!} + \left(ms^3 + 2m^2s^2 + m^3s + ns + mn\right)\right) \end{aligned} \tag{14}$$

6. According to [1], there are at most $nm/2$ identified patterns for an $n \times m$ constant matrix $\boldsymbol{M}$. But for a *N-term* pattern, there are at most $(m!/(N! \times (m - N)!))$ possibilities, hence the identified *N-term* patterns are at most

$$s_{most}(N) = \min\left(\frac{m!}{(m-N)!N!},\ nm/2\right) \tag{15}$$

Then the total computational complexity for N-term patterns elimination in Step 6 is

$$\begin{aligned} &O\left(\sum_{s=s_{most}(N+1)+1}^{s_{most}(N+1)+s_{most}(N)} \left((s+n-1)\frac{m!}{(m-N)!N!} + \left(ms^3 + 2m^2s^2 + m^3s + ns + mn\right)\right)\right) \\ &< O\left(\sum_{s=1}^{nm/2} \left((s+n-1)\frac{m!}{(m-N)!N!} + \left(ms^3 + 2m^2s^2 + m^3s + ns + mn\right)\right)\right) \\ &\approx O\left(\left(\frac{m!}{(m-N)!N!}\right)\left(\frac{n^2m^2}{8}\right) + \frac{n^4m^5}{64}\right) \end{aligned} \tag{16}$$

where $s_{most}(N + 1) = 0$ when $N = m$.

7. Then the computational complexity for Step 7 is

$$\begin{aligned}&\sum_{N=m}^{2}\left(O\left(\left(\frac{m!}{(m-N)!N!}\right)\left(\frac{n^2m^2}{8}\right)+\frac{n^4m^5}{64}\right)\right)\\&\approx O\left(\left(\frac{n^2m^2}{8}\right)\sum_{N=m}^{2}\left(\frac{m!}{(m-N)!N!}\right)+\frac{n^4m^6}{64}\right)\end{aligned} \tag{17}$$

8. The computation of ADP in Step 8 is composed of two parts: the computation of area and the computation of CPD. The computational complexities for the two parts are at most $O(n^2m^2/4)$ and $O(n^3m^3/4)$, respectively. Then the computational complexity for Step 8 is at most $O(n^3m^3/4)$.

- **Algorithm 3** completes once cycle by taking Steps 2–8, then the total computational complexity for Steps 2–8 is

$$\begin{aligned}&O\left(\left(\frac{n^2m^2}{8}\right)\sum_{N=m}^{2}\left(\frac{m!}{(m-N)!N!}\right)+\frac{n^4m^6}{64}\right)+O\left(\frac{n^3m^3}{4}\right)\\&\approx O\left(\left(\frac{n^2m^2}{8}\right)\sum_{N=m}^{2}\left(\frac{m!}{(m-N)!N!}\right)+\frac{n^4m^6}{64}\right)\end{aligned} \tag{18}$$

9. Because there are at most $(nm/2)$ patterns identified in a cycle of algorithm, then there are at most $(nm/2)!$ cycles in the algorithm. Therefore, the worst case computational complexity for Step 9 is

$$O\left(\frac{nm}{2}!\times\left(\frac{n^4m^6}{64}+\left(\frac{n^2m^2}{8}\right)\sum_{N=m}^{2}\left(\frac{m!}{(m-N)!N!}\right)\right)\right) \tag{19}$$

10. The worst case computational complexity for Step 10 is $O((nm/2)!)$.

Table 1 Comparison of area reduction and ADP reduction by the CSE algorithms

	Area	Reduction (%)	ADP	Reduction (%)
Straight forward implementation	23.4167	–	69.4583	–
[5]	13.8333	40.9255	54.2500	21.8956
[6]	13.5000	42.3488	52.3750	24.5950
[3]	13.9583	40.3917	50.7083	26.9946
Proposed	13.5417	42.1708	48.4583	30.2340

- As analysis above, the worst case computational complexity of **Algorithm 3** is

$$\begin{aligned} & O\left(\frac{nm}{2}! \times \left(\frac{n^4m^6}{64} + \left(\frac{n^2m^2}{8}\right)\sum_{N=m}^{2}\left(\frac{m!}{(m-N)!N!}\right)\right)\right) + O\left(\frac{nm}{2}!\right) \\ & \approx O\left(\frac{nm}{2}! \times \left(\frac{n^4m^6}{64} + \left(\frac{n^2m^2}{8}\right)\sum_{N=m}^{2}\left(\frac{m!}{(m-N)!N!}\right)\right)\right) \end{aligned} \tag{20}$$

4.2 Performance Test and Comparison

In this test, a number of binary CMMs in Advanced Encryption Standard (AES) S-box implementation are used to evaluate the efficiency of the proposed CSE algorithm. The S-box implementation with composite field arithmetic can be described as

$$\mathbf{Y} = \mathbf{M}(\boldsymbol{\beta}^{-1}(\beta\mathbf{X})^{-1}) + \mathbf{C} \tag{21}$$

where $\boldsymbol{\beta}$ is an 8 × 8 binary matrix of isomorphism function and $\boldsymbol{\beta}^{-1}$ is an 8 × 8 binary matrix of inverse isomorphism function. The inverse isomorphism matrix $\boldsymbol{\beta}^{-1}$ and affine transformation matrix $\boldsymbol{M}$ are both linear transformations, therefore they are merged into a new matrix $\boldsymbol{L} = \boldsymbol{M}\boldsymbol{\beta}^{-1}$ to reduce the gate count in many cases.

We take 8 different isomorphism matrices ($\boldsymbol{\beta}_0 \sim \boldsymbol{\beta}_7$) that are generate by the **Algorithm 1** in [7] in the case of $\{\boldsymbol{\phi} = (11)_2, \lambda = (1100)_2\}$, and the corresponding inverse isomorphism matrices ($\boldsymbol{\beta}_0^{-1} \sim \boldsymbol{\beta}_7^{-1}$), and the corresponding merged matrices $\boldsymbol{L}_0 \sim \boldsymbol{L}_7$ to test the performance of **Algorithm 3**. The average of area reduction and ADP reduction achieved by **Algorithm 3** for the 24 test matrices are listed and compared with other algorithms in Table 1. The ADP reduction of proposed algorithm can be up to 30.2340 % compared with the straightforward implementation. And it is improved by 8.3384 % compared to [5], and by 5.6390 % compared to [6], and by 3.2394 % compare to [3]. The CSE algorithms proposed in [5] and [6] only take the area into account. Compared with other algorithms, the algorithm in [6] achieves more efficient on area reduction, by taking a greedy algorithm to search the minimal area. The CSE algorithm in [3] is used to select the combination of matrices with minimal ADP, but it does not optimize the delays in a single matrix.

5 Conclusions

The aim of this paper is to propose a CSE algorithm with efficient ADP reduction for VLSI designs of binary CMM. For promoting delays optimization, a gate-level computing method is used in this algorithm. The method computes the delays based on the transformed constant matrices. The method can be used not only for a *multi-term-pattern* CSE algorithm but also for a *two-term-pattern* CSE algorithm. Note that the *two-term-pattern* CSE algorithm can be regard as a special case of the *multi-term-pattern* CSE algorithm. The delay computing method may be improvement by reducing the computational complexities according the special properties of *two-term-pattern* CSE algorithm.

The new CSE algorithm employs a greedy algorithm to find out the minimal ADP results. But it increases the computational complexities to do so. The proposed algorithm may take days for 12 × 12 matrices and higher dimensional matrices. Another more efficient way in ADP control is to reduce area under delay constrains, which is one of further works for us.

Acknowledgements This work was supported by the National Natural Science Foundation of China under Grant 61376025, the Industry-academic Joint Technological Innovations Fund Project of Jiangsu under Grant BY2013003-11.

References

1. N. Chen, Z.Y. Yan, Cyclotomic FFTs with reduced additive complexities based on a novel common subexpression elimination algorithm. IEEE Trans. Signal Process. **57**(3), 1010–1020 (2009)
2. A. Chandrakasan, M. Potkonjak, R. Mehra, J. Rabaey, R.W. Brodersen, Optimizing power using transformations. IEEE Trans. Comput.-Aided Design Integr. Circuits Syst. **14**(1), 12–31 (1995)
3. N. Wu, X.Q. Zhang, Y.F. Ye, L.D. Lan, in *Proceedings of The World Congress on Engineering and Computer Science 2013, WCECS 2013*. Improving common subexpression elimination algorithm with a new gate-level delay computing method. Lecture Notes in Engineering and Computer Science, 23–25 October. (San Francisco, USA, 2013), pp. 677–682
4. O. Gustafsson, M. Olofsson, in *First International Workshop on Arithmetic of Finite Fields (WAIFI 2007)*. Complexity reduction of constant matrix computations over the binary field, LNCS, vol. 4547 (Springer, 2007), pp. 103–115

5. R. Paško, P. Schaumont, V. Derudder, S. Vernalde, D. Ďuračková, A new algorithm for elimination of common subexpressions. IEEE Trans. Comput.-Aided Design **18**(1), 58–68 (1999)
6. C. Zeng, N. Wu, X.Q. Zhang, The optimization of AES S-box circuit design based on multiple-term CSE algorithm. Acta Electronica Sinica (Chinese Edition), **42**(3), (2014)
7. X. Zhang, K.K. Parhi, On the optimum constructions of composite field for the AES algorithm. IEEE Trans. Circuits Syst. II Express Briefs **53**(10), 1153–1157 (2006)

Chapter 12
Accuracy Enhancement of a SHM System by Light Scanning Sensor Improvement

Wendy Flores-Fuentes, Moises Rivas-López, Oleg Sergiyenko, Julio C. Rodriguez-Quiñonez, Daniel Hernandez-Balbuena and Javier Rivera-Castillo

Abstract This paper describes an analog signal processing technique, for the development of a novel electronic circuit to be embedded in a photodiode sensor, as an integrated circuit board for electronic signal processing, to detect the energy centre of an optical signal, which represents the most accurate position measurement from a light emitter source mounted on a structure (like buildings, bridges, and mines). The Optical Scanning Sensor for Structural Health Monitoring (SHM) is proposed due to the fact that the signal processing stage is embedded in the sensor and does not require additional software processing, reducing the time and memory spacing requirements for information recording. The theoretical principle of operation, technological and experimental aspects of design, development and validation are presented.

Keywords Energy signal centre · Estimation error · Optical scanning · Optical signal processing · Structural health monitoring · Support vector machine

W. Flores-Fuentes (✉) · M. Rivas-López · O. Sergiyenko · J. C. Rodriguez-Quiñonez ·
D. Hernandez-Balbuena · J. Rivera-Castillo
UABC, Mexicali, Mexico
e-mail: wendy.flores6@uabc.edu.mx

M. Rivas-López
e-mail: mrivas@iing.mxl.uabc.mx

O. Sergiyenko
e-mail: srgnk@mail.ru

J. C. Rodriguez-Quiñonez
e-mail: julio.rodriguez37@uabc.edu.mx

D. Hernandez-Balbuena
e-mail: dhernan@uabc.edu.mx

J. Rivera-Castillo
e-mail: javier.rivera.castillo@uabc.edu.mx

H. K. Kim et al. (eds.), *Transactions on Engineering Technologies*,
DOI: 10.1007/978-94-017-9115-1_12,

1 Introduction

Structures experience through its time of service deterioration and damage due to environmental conditions and excessive load conditions such as humidity, corrosion, earthquakes, gust waves, and traffic, among others. This results in structures deformation, cracking, dislocation and even collapses. Structures play in important role in safety and economy. Thus, the increasing demand for safest and functional structures has driven the SHM research of data acquisition and its analysis to obtain indicators of the structure health.

Nowadays, there are SHM systems with sensors based on several technologies like optical fiber, video cameras, and optical scanner sensors. Each of them with their advantages and disadvantages regarding the type of structure and the variables to monitor and analyze according to the kind of potential damage that could suffer and is intended to prevent.

This work introduces a position measurement method based on an optical scanning sensor that is composed of a light emitter source mounted on a structure, and an optical scanning aperture. This emits the light beam through a lens to a photodiode to convert it into a voltage output signal. Such output signal is electronically processed to detect the light energy centre which represents its most accurate position measurement to finally determine if any displacement has occurred [1, 2].

The energy signal centre concept has been used before in different applications with signals from various types of transducers and numerous mathematical methods have been developed to post process it. Some of these mathematical methods have been evaluated for optical scanning sensing; some of them are the Geometric Centroid, Power Spectrum Centroid and Peak Detection, but all of them require digital signal post processing time and memory storage [3]. Instead of a digital signal post processing, a real time electronic signal processing is developed embedded in the photodiode sensor.

2 Theoretical Principle of Operation

The light scanning sensor presented for SHM is a position measurement system. It is composed of a light emitter source installed on the structure under monitoring and an optical scanning aperture that is scanning the structure in search of the light emitter source to determine if it suffered a displacement.

2.1 Position Indicator

A light emitter source is used as a position indicator mounted on the structure under monitoring; it could be a coherent light emitter source such as a laser or an incoherent light source such as a bulb like the ones used in vehicles. Assuming that for any light emitter source there is only one energy centre that represents its punctual position.

2.2 Optical Scanning Aperture

The rotating optical aperture is designed as a 45° sloping mirror surface on a cylindrical rod for light beam deviation to a double convex lens with an interference filter and a photodiode. While the cylindrical rod mounted, on a dc electrical motor shaft, is rotating, an electronic signal is generated. Figure 1 illustrates a diagram with the main elements of the optical scanning aperture. When the mirror starts to spin, the sensor "s" is synchronized as the origin generates a pulse that indicates the 0° position and the starting of a cycle of 360° that finishes immediately before the "s" sensor generates the next beginning pulse. These pulses are used to calculate the scanning frequency and the zero reference to measure the angle where the light emitter source is found [4].

Figure 2 shows a signal timing diagram that exemplifies the starting pulse and the optoelectronic signal relation to calculate the light emitter energy signal centre position as described in the equations below. The interval $T_{2\pi}$ is equal to the time between m_1 and m_1 as in Eq. (1), that are expressed by the code $I_{2\pi}$ as defined in Eq. (2).

$$T_{2\pi} = \frac{2\pi}{\omega} = \frac{2\pi}{2\pi f} = \frac{1}{f} \tag{1}$$

$$I_{2\pi} = T_{2\pi} f \tag{2}$$

The time $t\alpha$ is equal to the interval between m_1 and m_2, could be expressed by the code $I_{\propto}$ as defined in Eq. (3). And finally the angle under measurement is calculated by Eq. (4) [5].

$$I_{\propto} = t_{\propto} f \tag{3}$$

$$\propto = \frac{2\pi I_{\propto}}{I_{2\pi}} = \frac{2\pi t_{\propto} f}{T_{2\pi} f} = \frac{2\pi t_{\propto} f}{\frac{2\pi}{\omega} f} = t_{\propto}\omega = t_{\propto} 2\pi f \tag{4}$$

where:

$T_{2\pi}$ is the interval of one cycle, from m_1 to $m_{1.}$

ω is the angular speed.

f is the scanning frequency (cycles in a second).

$I_{2\pi}$ is the interval code of one cycle, from m_1 to $m_{1.}$

$I_{\propto}$ is the interval code from starting cycle to energy signal centre.

$t_{\propto}$ is the interval from starting signal to energy signal centre, from m_1 to $m_{2.}$

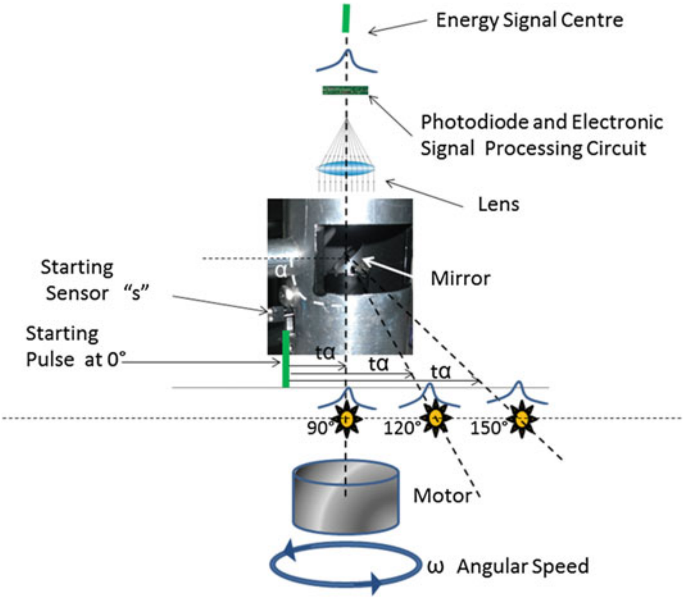


Fig. 1 Optical scanning aperture

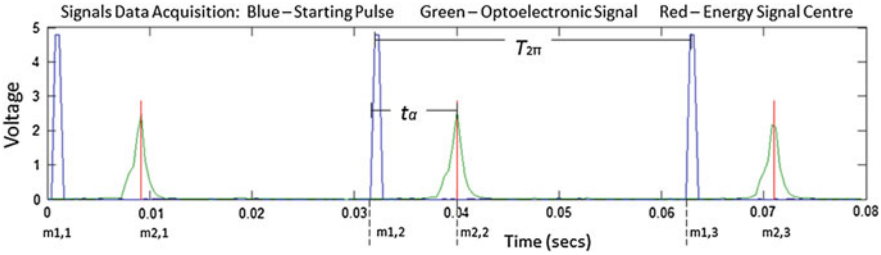


Fig. 2 Optical scanning aperture timing diagram

2.3 The Energy Signal Centre

The optoelectronic signal generated is a Gaussian-like shape. The photodiode converts the light input to voltage while the mirror is rotating. This is mainly observed at long distance, at near distance the light emitter source looks like a punctual source, but at distance the source expands its radius as a cone-like or an even more complex shape depending on the properties of the medium through which the light is travelling. To reduce errors in position measurements, the best solution is taking the measurement in the energy centre of the optoelectronic signal.

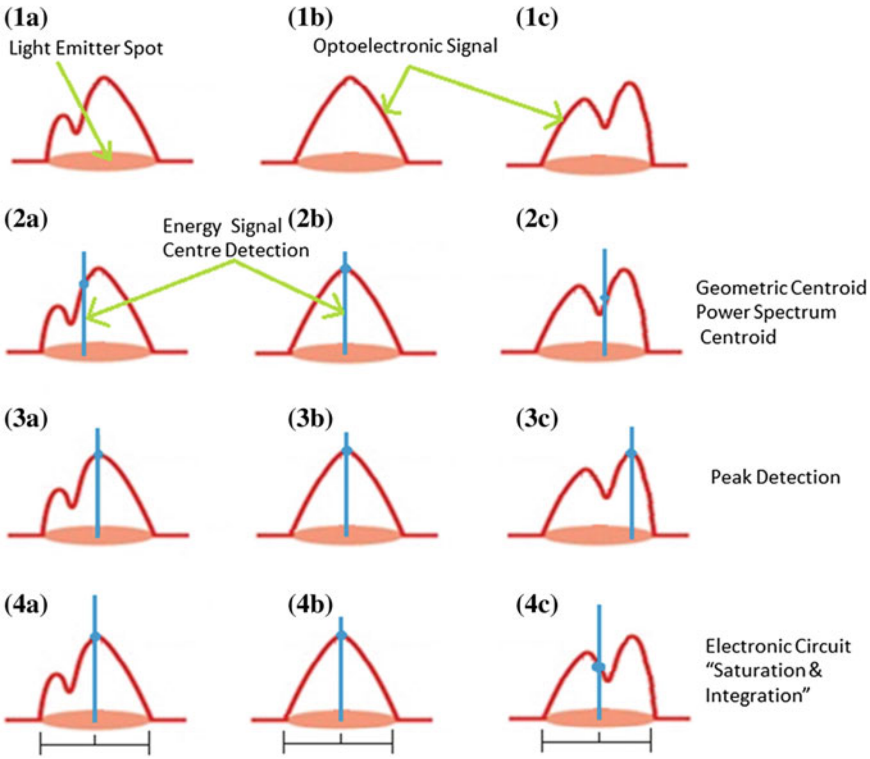


Fig. 3 Optoelectronic energy signal centre detection methods

Some digital methods for energy signal centre detection are: 2.3.1. Geometric Centroid and Power Spectrum Centroid, 2.3.2. Peak Detection and our proposed 2.3.3. Electronic Circuit method called "Saturation and Integration" [4].

In Fig. 3 a graphical representation of each method is observed. Centroid calculation and peak calculations give excellent results when the optoelectronic signal is almost a perfect Gaussian-like shape. But most of the times are observed that the optoelectronic signals are asymmetrical with different shapes in function of the scanning frequency, light emitter source distance to optical scanning aperture, the $t_{\propto}$ length (interval from starting signal to energy signal centre) and the light emitter source tilt over the surface, due to motor eccentricity at low speed scanning, light emitter montage and other optical phenomena such as reflection, diffraction, absorption and refraction. The luminous flux loss (energy loss) due to these phenomena could results in signals with asymmetrical shape (signals with more than one peak, with peak displaced, or even more strange shapes, etc.) instead of a symmetric Gaussian-like shape. Although it looks like two functions it is not a piecewise function, it still is just one continuous function. As consequence, when the signal is asymmetrical the results tend to the highest light density side, and the electronic circuit "Saturation and Integration" method is not affected by the signal deformation, due once the optoelectronic signal reaches the level reference voltage reference level the measurement is performed at its half time interval.

2.3.1 Geometric Centroid and Power Spectrum Centroid

These methods make a correlation between the energy signal centre with "the centre of mass" of the light incident in the surface of the optical scanner mirror. The Geometric Centroid for continuous 1-D light intensity distribution is given by Eq. (5) where *f(x)* is the irradiance distribution at the position x on the1-D light intensity distribution (a voltage function of the signal shape generated by the scanner).

$$x = \frac{\int_{-\infty}^{\infty} x f(x) dx}{\int_{-\infty}^{\infty} f(x) dx} \tag{5}$$

The Power Spectrum Centroid is calculated by Eq. (6) after processing the signal through Fourier Transformation to obtain the Power Spectrum which represents the signal energy by frequency to posterior correlate the power spectrum centroid calculated in the frequency domain to the corresponding centroid in time domain where *SC* is the power spectrum centroid in frequency (Hertz), $X^d[k]$ is the magnitude corresponding to frequency bin k, k is the frequency bin (*fs/N*) in hertz and *fs* is the frequency sample, and *N* is the length of the Fourier transformation series.

$$SC_{Hz} = \frac{\sum_{k=1}^{N-1} k * X^d[k]}{\sum_{k=1}^{N-1} X^d[k]} \tag{6}$$

2.3.2 Peak Detection

Peak Signal Algorithms are simple statistic algorithms for non-normally distributed data series to find the peak signal through threshold criteria. The algorithms which identify peaks in a given normally distributed time-series are selected to be applied in a power distribution data, whose peaks indicate high demands, and the highest corresponds to the energy centre. Each different algorithm is based on specific formalization of the notion of a peak according to the characteristics of the optical signal, as Eq. (7) for this kind of optoelectronic signals.

$$S(k,\ i,\ x_i, T) = \frac{\frac{\sum_{i=1}^{k} x_i - x_{i-1}}{k} + \frac{\sum_{i=1}^{k} x_i - x_{i+1}}{k}}{2} \tag{7}$$

where T is an optoelectronic signal containing N values. x_i be a given *i*th point in T. $k > 0$ is a given integer of k temporal neighbours of x_i (around the *i*th point), and S be a given peak function, S(i, x_i, T) with *i*th element xi of the given time-series T. A given point xi in T is a peak if S(i, x_i, T) $> \theta$, where θ is a user-specified (or suitably calculated) threshold value.

2.3.3 Electronic Circuit "Saturation and Integration"

In this method the optoelectronic signal is processed by means of an electronic circuit. The signal captured by the photodiode is processed through a circuit. It sets a threshold to the optoelectronic signal produced by the light emitter spot to measure the time it reaches the threshold and calculates its half time interval, which corresponds to the energy signal centre.

In base of this previous analysis we assume there is only one energy centre in a light source representing the position under monitoring. This approaches comes from the analysis of signal energy centre thru several methods just mentioned, where we could find that even when the optoelectronic signal could have more than one peak on the signal (light emitter source energy peaks), it could be found the energy signal centre by the centroid calculation on both time and frequency domains by geometric centroid and power spectrum centroid calculation respectively.

However although the power spectrum centroid and the geometric centroid results well coincides in our experiments, it has been detected that when the signal is affected by noise in any of its sides it affects the centroid position displacing it to the respective side with noise. And the improvement has been developed, in the proposed method to saturate the signal and establish a threshold from which noise is left out to calculate its half time interval, which correspond to the energy signal centre concept we are looking for to correlate with the position under monitoring.

3 Technological and Experimental Design

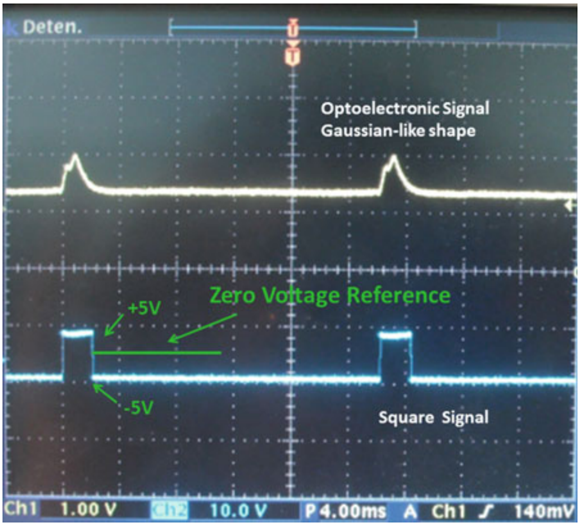
3.1 Measurement

The procedure below describes the steps to get the energy signal centre through the electronic circuit "Saturation and Integration" method:

3.1.1 Level Detector

A JFET operational amplifier is used as a voltage level detector to convert the optoelectronic Gaussian-like signal to an square signal, when the optoelectronic signal is lower than the reference voltage the JFET gets down to the negative

Fig. 4 Optoelectronic Signal and Square Signal from Level Detector Output



saturation voltage, and when the optoelectronic signal reaches the voltage reference the JFET gets up to the positive saturation voltage obtaining a signal as shown on Fig. 4.

3.1.2 Integrator

A JFET operational amplifier is used as an integrator to convert the square signal to a ramp signal to make the energy signal centre cross the zero reference as shown on Fig. 5.

3.1.3 Zero-Crossing Detector

A low input current voltage comparator with zero-crossing configuration is used to determine, by a rising edge pulse, where the ramp signal crosses the zero voltage reference, corresponding to the energy signal centre as shown in Fig. 6.

3.1.4 CR Circuit

A capacitor and a resistance are set in series at the zero-crossing detector output to obtain a pulse signal in the edges. Then a diode is used to conserve only the rising edge which corresponds to the energy signal centre, as shown in Fig. 7.

Figure 8 shows a correlation between the optoelectronic signal, the square signal, the ramp signal and the diode output representing the energy signal centre.

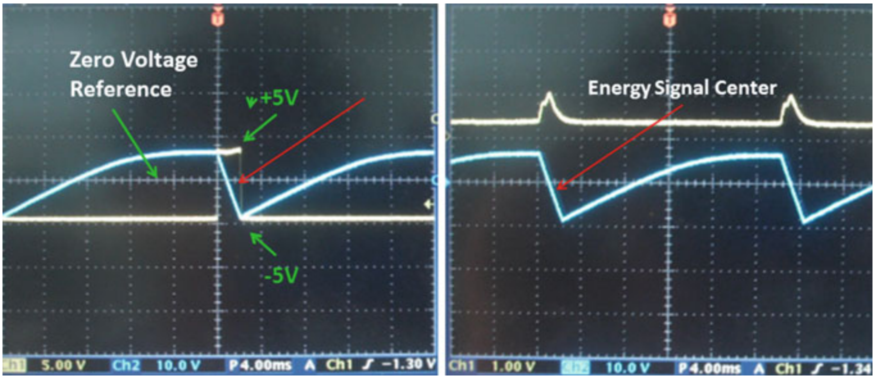


Fig. 5 Square signal from level detector output and ramp signal from integrator

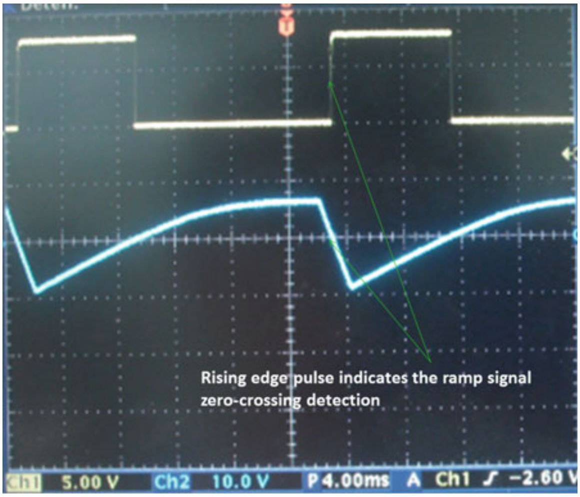


Fig. 6 Zero-crossing detector signal and ramp signal from integrator output

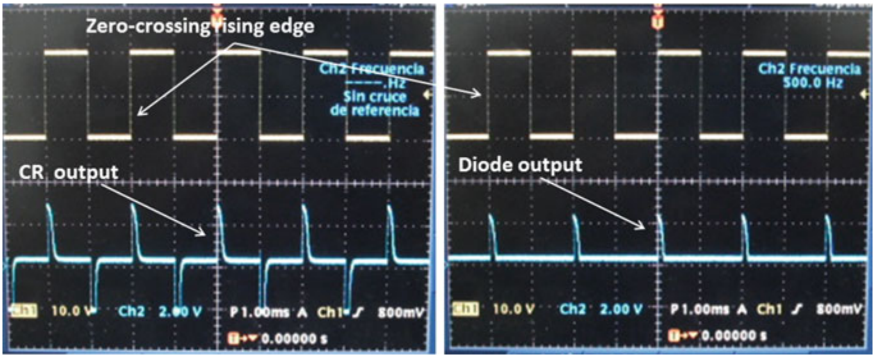


Fig. 7 CR output and final diode output (energy signal centre)

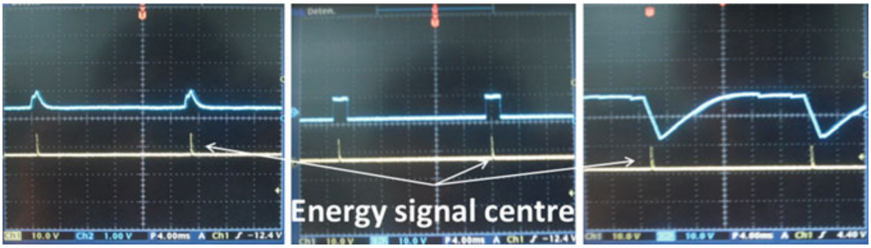


Fig. 8 Electronic optoelectronic signal processing steps

Table 1 Artificial intelligence algorithms MSE assesment results

Artificial intelligence algorithm	Training MSE results	Test MSE results	Best order result
Neuronal network	0.0907	0.0997	2
Support vector machine regression	0.0166	0.0205	1
Principal component regression	0.0987	0.0999	3
Partial least square regression	0.1010	0.1047	4
Ridge regression	0.1510	0.1568	6
Least absolutely shrinkage and selection operator	0.1033	0.1097	5
Generalized linear model fitting	0.0987	0.0999	3

The time measurement between m_1 and m_2 as described in Fig. 2 is used to calculate the angle by Eq. (4) without any requirement of digital signal processing.

3.2 Measurement Correction

During experimentation it was seen that the optoelectronic Gaussian-like shape signal experiences some deformation due to some internal and external error sources.

It is necessary to recognize that each set of measurements could be affected by different error sources generated due to environmental conditions or even errors due to the mechanism by itself. Hence, systematic and random errors do not follow a linear function, since their behavior is by the position—i.e. angle and distance, scanning frequency. For this reason, a digital rectification by a linear function is not suitable to the task at hand and seven artificial intelligence algorithms were assessed to select the one that best suit the data model as shown in Table 1, [6–9] with the goal to predict the value of the outcome error measure based on a number of input measures to compensate the error measurement caused by the few disadvantage on rotatory mirror scanners.

Therefore, in this work it is used error approximation functions to perform the digital rectification by using a well know machine learning regression model, the

Support Vector Machine (SVM). It is an intelligent algorithm to be trained with historical system behavior aimed to provide error predictions into the correction process at each measurement.

The measurement correction to minimize random and systematic error is performed by Eq. (8) by way of a (SVMR) Support Vector Machine Regression to predict the error based on system historical behaviour.

$$\alpha_{MC} = \alpha_M + E_p \tag{8}$$

where:

α_{MC} is the angle measurement by the system corrected.
α_M is the angle measured by the system.
E_p is the predicted error by SVMR.

4 Development and Validation

4.1 Experimental Work

With the Energy Signal Centre measurements and the real value, the measurement error was calculated by Eq. (9).

$$E = |\alpha_R - \alpha_M| \tag{9}$$

where:

E is the measurement error, representing how far the measurement is from the real value.
α_R is the target angle.
α_M is the angle measured by the system.

Measurements were performed by scanning from 45–135°, each 5°, at ten different positions by angle (10 measurements were taken at each point (angle, distance)). Obtaining a dataset with 1900 measurements.

All the data sets values were linear scaling to the range [−1, 1] and were separated in training data set and test data set. *X* training regression data set was composed of 1266 objects with 3 attributes (Angle, Distance, and Frequency). *Y* training regression data set was composed of 1266 targets (Error measurement). *X* test regression data set was composed of 634 objects with 3 attributes (Angle, Distance, and Frequency). Y test regression data set predicted composed of 634 targets (Error predicted). The SVMR was performed with the radial basis function (RBF) kernel described by Eq. (10), with the following settings:

NU Fraction of objects outside the ‘data tube’ = [2E^{-6}]

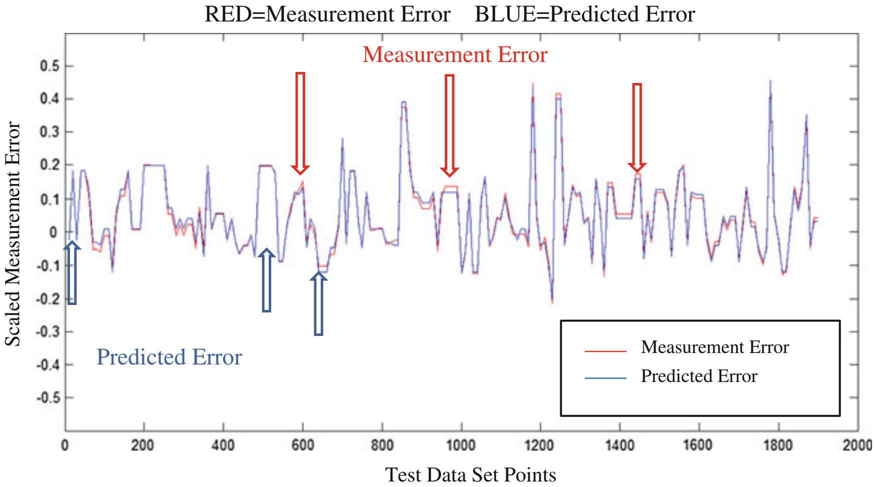


Fig. 9 Real versus predicted scaled measurement error

KPAR Parameter = [2]
EP Epsilon (ε), with of the 'data tube' = default value.

$$K\left(x_i x_j\right) = exp\left(-\gamma x_i - x_j\right)^2, \gamma > 0 \tag{10}$$

4.2 Results

Measurements show that the error distribution depends on the angle under measurement and the distance from the light emitter source to the aperture sensor.

SVMR was used to predict the error with successful results as shown in Fig. 9 where the red signal corresponds to the real measurement error and the blue signal corresponds to the predicted error. Linear regression was performed obtaining an R SVM coefficient value of 0.99556 where data and fit results tend to the plot regression target. Finally, the angle measurement error was with a Poisson distribution as shown in Fig. 10 with the probability of getting a measurement error $\geq \pm$ 4 of 0.5 % by Eq. (11), resulting in a 99.5 % of probability on getting measurements with an error less than 4°

$$p(x, \lambda) = \frac{\lambda^x \varepsilon^{-\lambda}}{x!} \tag{11}$$

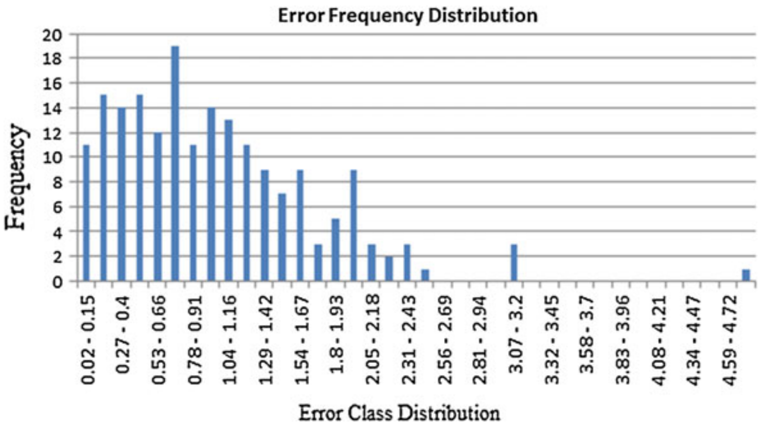


Fig. 10 Poisson distribution

5 Conclusion and Future Work

This paper proposes a novel electronic circuit design to be embedded in a photodiode sensor integrated circuit for optical signal processing to detect its energy centre which represents the most accuracy position measurement from a light emitter source mounted on a structure to monitor its displacements, without the signal post processing request. Measurement errors were corrected with successful results, enhancing the scanning system accuracy for all the angles and distances under measurement. Even more, other improvements have been visualized for further research on both the mechanical-electrical optical system, as in the energy signal centre processing and its correction by the analysis of the motor rotation frequency effects. Further research will continue to increase measurement accuracy for Optical Scanning System.

References

1. W. Flores-Fuentes, M. Rivas-Lopez, O. Sergiyenko, J. Rivera-Castillo, D. Hernandez-Balbuena, Analog signal processing in light scanning sensors for structural health monitoring accuracy enhancement, in *Proceedings of the World Congress on Engineering and Computer Science, WCECS 2013*. Lecture Notes in Engineering and Computer Science, vol. 2. San Francisco, 23–25 Oct 2013, pp. 655–661
2. M. Rivas Lopez, O.Y. Sergiyenko, V. Tyrsa, W. Hernandez Perdomo, L. Devia Cruz, D. Hernandez Balbuena, L. Burtseva, J. Nieto Hipolito, Optoelectronic method for structural health monitoring, Structural Health Monitoring **9**(2), 105–120 (2010). doi:10.1177/1475921709340975 (Online available: http://shm.sagepub.com/cgi/doi/10.1177/1475921709340975)
3. W. Flores Fuentes, M. Rivas Lopez, O. Sergiyenko, J. Rivera Castillo, Comparison of signal peak detection algorithms in the search of the signal energy center for measuring with optical scanning, in *Proceedings of the IEEE Section Mexico, IEEE ROC&C2011:XXII autumn*

international conference on communications, computer, electronics, automation, robotics and industrial exposition (CP10,PON15), Guerrero, Jan 2011

4. M. Rivas Lopez, W. Flores Fuentes, J. Rivera Castillo, O. Sergiyenko, D. Hernandez Balbuena, A Method and Electronic Device to Detect the Optoelectronic Scanning Signal Energy Centre, ed. by S. Pyshkin. Optoelectronics: Advanced Materials and Devices. ISBN: 978-953-51-0922-8, InTech, doi:10.5772/51993 (Online available from: http://www.intechopen.com/books/optoelectronics-advanced-materials-and-devices/a-method-and-electronic-device-to-detect-the-optoelectronic-scanning-signal-energy-centre)
5. M. Rivas, O. Sergiyenko, M. Aguirre, L. Devia, V. Tyrsa, I. Rendon, Spatial data acquisition by laser scanning for robot or SHM task. in *IEEE*, Jun 2008, pp. 1458–1462. doi:10.1109/ISIE.2008.4676974. (Online available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4676974)
6. A.J. Smola, B. Schlkopf, A tutorial on support vector regression. Stat. Comput. **14**, 199–222 (2004)
7. J. Gascón-Moreno, E. Ortiz-García, S. Salcedo-Sanz, A. Paniagua-Tineo, B. Saavedra-Moreno, J. Portilla-Figueras, in *Advances in Computational Intelligence*. Multi-parametric Gaussian Kernel function optimization for ε-SVMr using a genetic algorithm, Lecture Notes in Computer Science, vol. 6692, Chapter 15 (Springer, Berlin, 2011), pp. 113–120. doi:10.1007/978-3-642-21498-1_15. (Online available: http://www.springerlink.com/index/10.1007/978-3-642-21498-1_15)
8. Z.Z. Zhang, Some recent progresses in network error correction coding theory, in *IEEE*, Jan 2008, pp. 1–5 (Online available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4476186)
9. J.K. Choi, S.H. Park, D.J. Cho, K.Y. Seo, Correction error generation algorithm for differential positioning performance analysis of navigation equipment, in *IEEE*, Oct 2008, pp. 1099–1103 (Online available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4694316)

Chapter 13
A Metric Suite for Predicting Software Maintainability in Data Intensive Applications

Ruchika Malhotra and Anuradha Chug

Abstract Software maintainability is the vital aspect of software quality and defined as the ease with which modifications can be made once the software is delivered. Tracking the maintenance behaviour of a software product is very complex that is widely acknowledged by the researchers. Many research studies have empirically validated that the prediction of object oriented software maintainability can be achieved before actual operation of the software using design metrics proposed by Chidamber and Kemerer (C&K). However, the framework and reference architecture in which the software systems are being currently developed have changed dramatically in recent times due to the emergence of data warehouse and data mining field. In the prevailing scenario, certain deficiencies were discovered when C&K metric suite was evaluated for data intensive applications. In this study, we propose a new metric suite to overcome these deficiencies and redefine the relationship between design metrics with maintainability. The proposed metric suite is evaluated, analyzed and empirically validated using five proprietary software systems. The results show that the proposed metric suite is very effective for maintainability prediction of all software systems in general and for data intensive software systems in particular. The proposed metric suite may be significantly helpful to the developers in analyzing the maintainability of data intensive software systems before deploying them.

Keywords Data intensive applications · Empirical validation · Machine learning · Prediction models · Software design metric · Software maintainability

R. Malhotra
Department of Software Engineering, Delhi Technological University, Delhi, India
e-mail: ruchikamalhotra2004@yahoo.com

A. Chug (✉)
University School of Information and Communication Technology, Guru Gobind Singh Indraprastha University, Dwarka, New Delhi, India
e-mail: a_chug@yahoo.co.in

H. K. Kim et al. (eds.), *Transactions on Engineering Technologies*,
DOI: 10.1007/978-94-017-9115-1_13,

1 Introduction

Producing software which does not need change is not only impractical but also very uneconomical. The process of making changes in the software once it has been delivered to the customer is called software maintenance [1] and the ease with which it could be done is called as software maintainability [2]. The amount of resource, effort and time spent on it is much more than what is being spent on development. Thus, producing software that is easy to maintain may potentially save large costs [3]. Practitioners have suggested many ways to control the maintenance cost and one of them is to utilize the software design metrics and predict maintainability in the early phases of project development [3, 4]. Maintenance cost can be kept under control by accurate prediction of software maintainability due to many reasons such as:

(a) Productivity cost among projects can be compared.
(b) More effective planning of valuable resources can be done in advance.
(c) Major decision regarding staff allocation can be timely made.
(d) The threshold values of various metrics can be checked.
(e) Determinants of software quality can be enhanced.
(f) Practitioners are able to achieve optimized maintenance costs.

Various metrics have been proposed in the literatures which have significant impact on software maintainability. The main purpose of this study is 2-fold, firstly to review the role of C&K metrics suite for the prediction of software maintainability and secondly to propose a new suite of metrics with the induction of two new metrics which have larger impact on maintainability in highly data intensive applications. In order to achieve the goal, the data was collected from five proprietary software systems developed in Microsoft Visual Studio using C# language and based on object oriented (OO) methodologies with heavy use of databases for processing of each query. To measure the features of OO paradigm, C&K metric suite proposed by Chidamber and Kemerer [5] has been found to be a significant indicator of maintainability predictions in large number of studies [6–15]. We rely on the outcome of these studies and use C&K metric suite to capture the OO characteristics. There were two deficiencies found in this metric suite. First observation was the same as noted by Li et al. [16] that it does not take into account the structural complexity of the software. To overcome this deficiency we added two metrics i.e. Maintainability Index (MI) proposed by Oman et al. [17, 18] and Cyclomatic Complexity (CC) proposed by McCabe [19]. The second and main deficiency found in the metric suite is on account of the amount of database handling. To overcome this deficiency, two new metrics were proposed and validated for the applications which heavily use databases. The proposed metrics are Number of Data Base Connections (NODBC) made each time for query processing and the Schema Complexity to Comment Ratio (SCCR) to measure the understandability of the databases. The metrics definition and collection is discussed in Sect. 3. Overall a set of ten metrics were considered as independent variables in this study which included 06 from C&K metric suite and CC, MI, NODBC and SCCR while the dependent variable was the number of the changes made in the lines of source code. Two versions of each of the software systems were taken and analyzed to count the changes made in the new version with respect to the older version. Four different versions of Artificial Neural Network (ANN) i.e. Back Propagation Network (BPN), Kohonen Network (KN), Feed Forward Neural Network (FFNN) and General Regression Neural Networks (GRNN) are used for making the prediction model. Data analysis was performed using correlation coefficient to verify the findings. We found that the new proposed metrics suite is significantly related with dependent variable. It is also observed that maintainability predictions for the applications which heavily use databases were more precise and accurate using new metric suite. Univariate as well as multivariate analysis further confirmed the results and proved the significance of proposed metrics suite. By using the new proposed metrics suite software practitioners can considerably take decisions whether the developed application is maintainable or not, which would save the time and money for the organizations responsible for developing and deploying the customized software's for the customers to gain their better satisfaction in the industry. The rest of the chapter is organized as follows: Sect. 2 presents the related work and Sect. 3 introduces proposed metrics. Section 4 describes independent, dependent variables and data analysis. Section 5 describes the machine learning methods used in the predictions process. Section 6 presents results and analysis, Sect. 7 discusses threats to validity and finally Sect. 8 concludes the paper with future directions.

2 Related Work

The problem of predicting the maintainability of the software is widely acknowledged in the industry due to the subjectivity involved while trying to quantify it. Jorgensen [20] suggested that we can measure maintainability by measuring the change efforts during operations. Many empirical studies have been conducted to predict the software maintainability using various tools and processes at the time of designing an application [6–15]. Multiple Linear Regression (MLR) Model was used by Li and Henry [6] to predict maintenance effort and to ear marked those metrics which have strong impact on maintainability. Muthanna et al. [21] also used polynomial regression to establish the relationship between design level metrics using industrial software. The results using graph plots have shown that predicted values were quite close to the actual values. Dagpinar and Jahnke [9] also carried empirical study and recorded significant impact of direct coupling metrics and size on maintainability. Fioravanti and Nesi [22] presented a metric analysis to identify which metrics would be better ranked for its impact on prediction of adaptive maintenance using MLR for OO systems. The validation has identified that several metrics can be profitably employed for the prediction of software maintainability. Misra [23] used linear regression in his study which was based on intuitive and experimental analyses using twenty design and code measures to obtain their indications on software maintainability. In the last decade some machine learning algorithms have also been proposed, evaluated and verified that they can predict maintainability more accurately and precisely. Thwin and Quah [10] used Artificial Neural Network (ANN), Koten and Gray [11] applied Bayesian Belief Network (BBN), Elish and Elish [12] applied Tree Nets in maintainability prediction modeling for OO systems. Kaur et al. [14] have verified the use of soft computing approaches for maintainability prediction to achieve more accuracy. Recently many nature inspired algorithm are successfully applied such as evolutionary programming for open source software systems by Banker et al. [24], ant colony optimization used by Sun and Wang [25] for optimizing preventive maintenance and genetic algorithms by Vivanco et al. [26].

3 Proposed Metrics

It's important to give equal attention to the database accesses with the enhancement in data base usage now days. With the increase in the use of mobile and mobile based applications, data that once might have been accessed a couple of times a week now might be accessed multiple times per hour. As the software systems heavily use data bases; hence we observed that C&K metric suite would not be adequate as it does not capture the database handling aspects of the applications. We proposed two more metrics namely SCCR and NODBC as presented in Table 1, to remove these deficiencies and we claim that two proposed metrics carries more impact on software maintainability in database intensive applications.

NODBC is measured by counting the number of times database connections were made using the function call 'Open()'. To count the SCCR, ratio of the numbers of field in the schema to the number of comment lines was considered. Authors are of the strong opinion that understandability of the schema of database is equally important in maintaining any application.

4 Research Background

Independent and dependent variable: To validate the effectiveness of proposed metric suite, 10 independent variables have been considered as compiled in Table 2. C&K metric suite is used to measure OO characteristics and MI as well as CC were used to capture the structural complexity of the code. Inspired by the results of Malhotra and Chug [27], NODBC and SCCR also added to measure the data base handling aspect.

Empirical data collection: Five proprietary systems were considered as presented in Table 3. To calculate the values of all independent variables, following

Table 1 Proposed metrics

Metric name	Description
Scheme complexity to comment ratio (SCCR)	It calculates the ratio of number of comments lines to the number of field in the schema of data base
Number of data base connections (NODBC)	Number of data base connection is a measure to count number of times database connections were made

Table 2 Set of independent variables

Metric name	Description
Weighted methods per class (WMC)	It is the sum of McCabes's cyclomatic complexities of all local methods in a class
Depth of inheritance tree (DIT)	It measures the depth of a class in terms of the distance from root class; the value of DIT for root class is zero
Number of children(NOC)	It measures the number of child classes of a class as it counts number of immediate sub classes
Coupling between object (CBO)	It count the number of other classes to which the given class is coupled
Response for a class (RFC)	It counts the number of functions executed in response to the received message
Lack of cohesion of methods (LCOM)	It counts the number of disjoint sets of local methods
Scheme complexity to comment ratio(SCCR)	It calculates the ratio of number of comments lines to number of field in the schema of data base
Number of data base connections (NODBC)	Number of data base connection is a measure to count number of times database connections were made
Maintainability index (MI)	Calculates an index value between 0 and 100 that represents the relative ease of maintaining the code
Cyclomatic complexity (CC)	Measures the structural complexity of the code

Table 3 Brief description of the proprietary software systems used in the empirical study

System name	Data points	Description
File letter monitoring (FLM system)	233	Customize software to handle the movement of files and documents in an office
EASY system	292	Web portal for an educational institute to provide study material and online evaluations
Student management system (SMS system)	129	Maintains the record of students and teacher for private educational institute
Inventory management system (IM system)	96	Maintains inventory of the company at different branch offices in different cities
Angel bill printing (ABP system)	114	Maintains fully editable items list by client itself with generation of a common bill format

strategy is used. Five metrics namely MI, CC, DIT, CBO and NOC were retrieved from the Visual Studio wherein the metrics mentioned were calculated from the intermediate language code generated while compilation. Three metrics namely

Table 4 Descriptive statistics of FLM system and EASY system

Metric	Max	Min	Mean	Med	SD	Max	Min	Mean	Med	SD
WMC	16	1	6.276	5	4.97	23	1	10.5	9.5	8.57
DIT	7	1	4.379	5	1.32	5	1	3.6	4	2.50
NOC	7	0	3.1	3	1.67	8	0	4.23	3	2.91
CBO	50	3	26.14	30	13.85	54	0	33.5	38.5	21.58
RFC	67	12	25.16	18	7.89	78	21	37.73	27	4.89
LCOM	0	0	0	0	0	0	0	0	0	0
SCCR	5	2	3.276	3	2.97	7	3	4.57	5	5.57
NODBC	12	0	2.483	0	3.53	7	0	2.79	0.5	3.43
MI	91	40	61.14	56	18.04	94	43	64.1	56.5	17.91
CC	29	1	19.31	16	13.76	22	1	20.6	19	14.26
Change	95	5	41.98	67	45.67	87	9	52.52	63	43.23

Table 5 Descriptive statistics of SMS system and IMS system

Metric	Max	Min	Mean	Med	SD	Max	Min	Mean	Med	SD
WMC	29	2	16.63	17.5	9.17	12	0	3.147	3	2.57
DIT	6	1	3.25	4	2.12	5	4	4.029	4	0.17
NOC	11	0	4.85	4	2.67	7	0	2.81	3	1.91
CBO	59	3	45.38	52.5	18.66	30	2	13	13.5	8.09
RFC	83	19	37.09	31	5.87	43	18	21.09	27	5.07
LCOM	0	0	0	0	0	3	0	0.147	0	0.55
SCCR	6	2	4.625	16.5	9.17	12	0	3.147	3	2.57
NODBC	6	0	3.89	3	2.50	5	0	2.118	1	3.85
MI	81	49	55.25	52	10.56	100	48	71.79	67	17.84
CC	27	1	21.5	19.5	19.61	13	2	10.79	7	12.38
Change	79	13	67.89	47	32.43	213	18	79.87	103	67.93

WMC, LCOM and RFC were calculated with the help of CCCC tool [28]. Remaining two metrics as proposed in this study SCCR and NODBC were collected through tool we created in the first phase of our research plan. We observed the software over a period of 3 years since it has been delivered. Original as well as modified versions were compared manually to count the CHANGE i.e. dependent variable. Any line of source code added or deleted is counted as one whereas modification counted as two changes. The value of change for each class was compiled and combined with respective values of independent variables to generate the data points. Same methodology was adopted in Zhou et al. [8] and Malhotra et al. [29]. We found 233, 292, 129, 96 and 114 data points for FLM, EASY, SMS, IMS and ABP System respectively.

Descriptive statistics such as Max, Min, Mean, and Median (Med) and Std Dev(SD) were calculated for FLM and EASY systems and presented in Table 4, SMS and IMS system presented in Table 5 whereas ABP system presented in Table 6. From the table it can be observed that the Max value of LCOM for FLM, EASY, SMS, IMS and ABP are 0, 0, 0, 3 and 6 respectively which represents that

Table 6 Descriptive statistics of ABP system

Metric	Max	Min	Mean	Med	SD
WMC	11	1	2.483	2	1.84
DIT	6	3	4.017	4	0.13
NOC	9	0	5.25	5	1.09
CBO	29	4	14.93	17	8.56
RFC	49	21	26.83	31	9.89
LCOM	6	0	0.155	0	0.81
SCCR	11	1	2.483	2	1.84
NODBC	8	0	4.931	1	1.04
MI	100	40	69.5	61	21.03
CC	14	2	10.33	8.5	8.88
Change	189	19	91.23	78	45.63

Table 7 Pearson correlation coefficient at 0.01 level of significance (two tailed)

Metrics	FLM change	EASY change	SMS change	IMS change	ABP change
WMC	**0.73**	**0.66**	**0.54**	**0.61**	**0.59**
DIT	−0.38	−0.42	−0.36	−0.42	−0.44
NOC	0.29	0.48	0.33	0.41	0.45
CBO	0.46	**0.61**	0.49	**0.58**	0.51
RFC	**0.64**	0.49	**0.50**	**0.51**	0.47
LCOM	0.48	0.42	0.41	**0.68**	**0.71**
SCCR	**0.56**	**0.55**	**0.66**	**0.69**	**0.73**
NODBC	**0.71**	**0.65**	**0.58**	**0.79**	**0.81**
MI	**0.61**	0.49	0.36	0.47	**0.58**
CC	**0.59**	**0.62**	0.39	0.41	**0.55**

classes are quite cohesive in first three applications. Values of DIT for FLM, EASY, SMS, IMS and ABP are 7, 5, 6, 5 and 6 which represents that inheritance is properly exploited in all systems. SCCR is medium in FLM, EASY and SMS and High in IMS and ABP which means IMS and ABP would be easier to understand in maintenance phase. A value of NODBC is more than 8 in FLM and ABP systems and less than 7 in EASY, SMS and IMS systems.

Correlation Analysis: Correlation Analysis provides important information about the interdependence between two variables. We calculated the Pearson's correlation coefficient represented as 'r' to measures the linear relationship between independent variables versus change and presented in Table 7. Value of 'r' represents the amount of correlation exists between the two variables and lies between +1 to −1. Values in the range of ±0.5–1 represent high correlation; ±0.3–0.5 represents medium correlation whereas less than ±0.3 represents very low correlation. In the Table 7, all entries above 50 % are marked as bold. It is inferred that NODBC metric as well as SCCR metric is significantly related to change metric for all the systems. The value of 'r' for new proposed metric is quite competitive as compared to other metrics. For IMS and ABP systems, more than

75 % correlation was observed whereas for FLM, EASY and SMS systems it was in the range of 58–75 % which is quite significant. SCCR is also found to be significantly correlated with change metric for all systems. When compared with other metrics it was found that although DIT is comparatively less correlated with the change however MI and CC are reasonably well correlated. Among the C&K metric suite, WMC is found to be most significantly related as for all systems, value of 'r' is found to be more than 54 % for all systems. RFC is significantly correlated with change for FLM, SMS and IMS systems. CBO found to be significantly correlated with change in EASY and IMS systems.

5 Research Methodology

In this section, we explain the various Machine Leaning (ML) methods used for making the prediction models as well as to ascertain the relationship of design metrics with maintainability. Recent research activities carried by authors [7, 15, 27, 29] have revealed that ANN is very powerful in classifying and recognizing the data patterns, so they are well suited for prediction problems as in such cases although the required knowledge is difficult to specify but enough data for observations are available to learn. They are originally developed to mimic basic biological neural systems particularly the neurons present in the human brain. Four different versions of ANN models have been selected in the current study as mentioned below.

(a) Back Propagation Network (BPN): Although BPN is originally invented by Hu [30] in 1964 however it came into use only in 1986 by Rumelhart et al. [31] when it was used as supervised learning technique. Training data in BPN consists of pair of vector (input vector and target vector). During the training process, an input vector is presented to the network for the learning process. Output vector is generated from these learning and compared with the actual target vector. If there is any difference in the values, the weights of the network are re-adjusted to reduce this error and the process is repeated until the desired output is produced.

(b) Kohonen Network (KN): Proposed by Kohonen [32], KN is best known as self organizing networks as they learn to create maps of the input space in a self-organizing way. Although, KN is invented to provide a way of representing multidimensional data in much lower dimensional spaces, a network is created that learn the information such that any topological relationships within the training set are maintained without supervision.

(c) Feed Forward Neural Network (FFNN): In FFNN [33, 34], information moves in only one direction i.e. forward from input nodes to output nodes through hidden nodes and there are no loops in the network. The number of hidden neuron selected as 10 for the sample data collected from these five real life applications.

(d) General Regression Neural Networks (GRNN): Proposed by Specht [35], it is very powerful network as it needs only a fraction of the training samples during learning process and finishes the learning process in single pass. Due to the highly parallel structure, it performs well even in case of noisy and sparse data and the over fitting problem does not arise as neither do they set the training parameters during the commencement of learning process, nor they define the momentum. Once the network finished the training process, only smoothing factor is applied to determine how tightly the network matches its prediction [10].

6 Results and Discussion

Ten independent and one dependent variable were selected in this study. Total 864 classes were collected and combined with respective changes made in each class. Univariate and Multivariate analysis was performed to find the significance of each metric individually and cumulatively on change.

Univariate Analysis using linear regression was performed to find the individual effect of NODBC and SCCR on change using SPSS and the results are presented in Table 8. Four columns represent estimated coefficient, standard error, the t-ratio and p-value. The value of Sig (p-value) represents amount of significance of these metrics on change. As evident from the outcome, both variables received the p-value as 0.000 which means they are significantly correlated with change.

Multivariate Linear Regression (MLR) was also performed using stepwise linear regression model in order to identify the most significant metrics for each system. MLR is the most commonly used technique for fitting a linear equation on observed data [8]. There are three methods used for identifying and picking the subset of important metrics from the set of independent variables i.e. forward selection, backward selection and stepwise selection. In this study, stepwise selection method is used as it guarantees to provide optimum and most significant subset of independent variables. At each step either the certain variables are added or deleted to identify the final most optimized regression model. Unstandardized Coefficient, Std Error, t-ratio and p-value (sig) to three decimal places are presented in Table 9.

Results show that two proposed metrics were found to be statistically significant for all systems as almost all p-value are less than .050. Unstandardized Coefficients represents the value when the dependent and independent (predictor) variables were all transformed to standard scores before running the regression and used to compare the relative strength of the various predictors. NODBC has the largest coefficient and one standard deviation increase in NODBC leads to a 0.915 decrease in change for IMS system. SCCR is also found to be quite competitive as one standard deviation increase in SCCR in turn leads to 0.858 standard deviation

Table 8 Results of univariate analysis

Metric	Unstandardized coefficient	Std. error	t-ratio	Sig (p-value)
NODBC	0.307	0.020	5.101	0.000
SCCR	0.245	0.046	2.826	0.000

Table 9 Results of multivariate analysis

Software system	Most significant metrics identified	Unstandardized coefficient	Std. error	t-ratio	Sig (p-value)
FLM system	Intercept	3.779	1.054	3.586	0.000
	WMC	0.009	0.004	2.532	0.012
	MI	0.055	0.067	2.489	0.014
	SCCR	0.485	0.015	2.203	0.000
EASY system	Intercept	0.542	0.191	2.838	0.005
	WMC	0.023	0.226	1.882	0.018
	SCCR	0.378	0.773	2.563	0.000
	NODBC	0.584	0.798	2.066	0.000
SMS system	Intercept	0.697	0.854	1.806	0.000
	NODBC	0.860	0.211	0.055	0.002
	MI	0.707	2.876	0.013	0.004
	SCCR	0.858	0.463	0.019	0.001
IMS system	Intercept	0.912	0.687	0.258	0.000
	SCCR	0.345	0.605	0.069	0.004
	NODBC	0.915	2.463	0.150	0.002
	RFC	0.301	4.501	0.663	0.000
ABP system	Intercept	0.032	1.268	0.757	0.003
	WMC	0.817	3.412	0.681	0.010
	NODBC	0.476	3.406	0.146	0.004
	MI	0.817	3.412	0.681	0.010

increase in change for SMS system. Apart from two reported metrics WMC and MI were also found to be most significant predictor of change.

Maintainability Prediction: Two types of prediction models were constructed for each system. **Model-1** is constructed using metrics suite presented by C&K [5] and **Model-2** is constructed by adding four more metrics MI, CC, NODBC and SCCR to the existing C&K metrics suite resulting in the set of 10 metrics in all. MLR, BPNN, KN, FFNN and GRNN were employed for software maintainability prediction by dividing the data into three parts i.e. 70 % for training and 30 % for testing as it is the commonly accepted proportion used by many practitioners [6–15]. Three prediction accuracy measures proposed by Kitchenham et al. [36] as presented in Table 10 are used to compare the performance of Model-1 and Model-2. Detailed method for their calculations are available in Malhotra et al. [15].

Results are presented in Table 11 where three rows for each software system represent the values of accuracy measures when MLR as well as ML models were applied with metric suite Model-1 (M-1) and metric suite Model-2 (M-2). For

Table 10 Prediction accuracy comparison proposed by Kitchenham et al. [36]

Name	Formula	Definition
MRE (Magnitude of relative error)	$\frac{\lvert Actual-Predicted\ Value\rvert}{Actual\ Value}$	Normalized measure of the discrepancy between actual and predicted value
MMRE (Mean magnitude of relative error)	$\sum_{i=1}^{N} MRE_i$	Average relative discrepancy
Pred(q)	$Pred(q)=\frac{K}{N}$	What proportion of the predicted values have MRE less than or equal to specified value

Table 11 Prediction accuracies of model-1(M-1) and model-2 (M-2) for all data sets

Software system	Accuracy measures	MLR model		BPNN model		KN model		FFNN model		GRNN model	
		M-1	M-2	M-1	M-2	M-1	M-2	M-1	M-2	M-1	M-2
FLM system	Max MRE	2.14	1.53	1.98	1.20	1.87	1.09	1.332	0.98	1.78	1.11
	MMRE	0.51	0.48	0.49	0.47	0.45	0.42	0.41	0.39	0.48	0.41
	Pred(0.25)	0.64	0.78	0.57	0.69	0.68	0.78	0.66	0.71	0.68	0.76
Easy system	Max MRE	2.09	1.82	1.24	1.65	0.99	0.98	1.090	1.02	1.47	1.32
	MMRE	0.66	0.57	0.51	0.46	0.46	0.36	0.43	0.37	0.49	0.42
	Pred(0.25)	0.58	0.63	0.59	0.63	0.69	0.74	0.70	0.77	0.68	0.69
SMS system	Max MRE	1.98	1.47	1.30	1.22	1.11	0.97	2.786	1.70	1.98	1.8
	MMRE	0.68	0.56	0.44	0.40	0.43	0.41	0.46	0.38	0.42	0.39
	Pred(0.25)	0.60	0.66	0.52	0.69	0.63	0.77	0.52	0.69	0.60	0.71
IMS system	Max MRE	2.34	1.71	1.52	1.43	2.33	1.10	1.803	1.22	1.88	1.63
	MMRE	0.71	0.59	0.40	0.35	0.43	0.32	0.39	0.30	0.38	0.29
	Pred(0.25)	0.54	0.63	0.59	0.71	0.68	0.70	0.59	0.63	0.58	0.69
ABP system	Max MRE	1.78	1.37	1.29	1.33	1.89	0.33	2.092	1.66	1.67	1.45
	MMRE	0.49	0.41	0.37	0.29	0.43	0.32	0.48	0.37	0.58	0.40
	Pred(0.25)	0.69	0.76	0.58	0.67	0.62	0.72	0.67	0.66	0.71	0.74

example first three rows belong to the results received using FLM system as MLR, BPNN, KN, FFNN and GRNN models were applied for each prediction algorithms with two different data sets i.e. Model-1 (M-1) and Model-2 (M-2).

From the results it is quite evident that overall improvement in the prediction accuracy is observed with new proposed metric suite for all systems. To further analyze the results we further sorted the systems in ascending order on the values of NODBC and SCCR. We observed that more improvement in prediction accuracy was achieved for those systems which have high values of NODBC and SCCR. ABP system has maximum SCCR and NODBC as compared to other systems. For ABP system maximum improvement in prediction accuracy is observed i.e. 23 % in the for MMRE whereas other systems such as FLM, EASY, SMS and IMS observed 7, 14, 11 and 19 % improvement in MMRE respectively. MaxMRE was improved by 39, 1, 21, 28 and 29 % for FLM, EASY, SMS, IMS and ABP System respectively. Lowest improvement for Easy systems was noticed

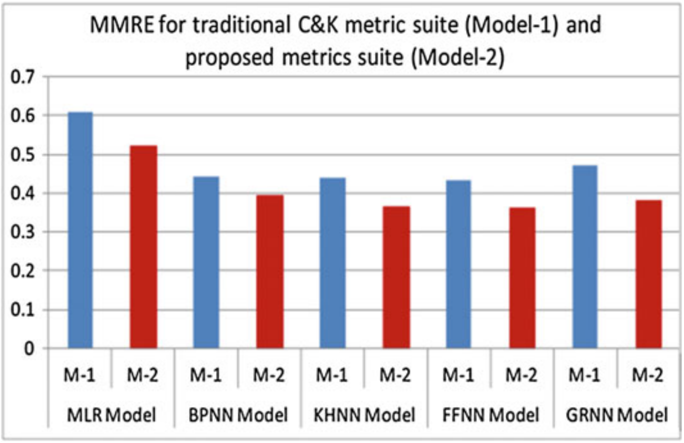


Fig. 1 MMRE for Model-1 and Model-2

which also has lowest SCCR as well as NODBC among all systems. Prediction accuracies achieved by all models were also compared and observed that the performance of ML models is better than MLR in general. When we compared the MMRE values for Model-2, it is found to be 0.94, 0.82, 0.66, 0.79 and 0.86 for MLR, BPNN, KN, FFNN and GRNN respectively. That means KN performance is best among all ML models. Graphs were also plotted to observe improvement in prediction accuracies from Model-1 to Model-2 w.r.t. MMRE and Pred(0.25) in Figs. 1, 2 respectively. It is quite evident from Fig. 1 that MMRE was significantly reduced from Model-1 to Model2 for all prediction techniques. Figure 2 represents the comparison of prediction accuracies achieved at 25 %. It is quite visible from the graph that pred(0.25) is improved from Model-1 to Model-2 for all techniques.

7 Threats to Validity

Whenever any empirical data is collected from proprietary software system, it has got few specific characteristics and their generalization always carries few threats to its validity. Also in this study, OO characteristics were measured using internal quality metrics suite proposed by C&K. However, software maintainability also depends upon external quality attributes such as competency of developers, familiarity with the code etc. They were intentionally avoided due to the subjectivity involved in their measurement. We also cannot assure if the proposed metrics suite is universally applicable for different programming languages and environment. In order to capture the cause-effect relationship between particular metric and maintainability, we need to perform controlled experiment where one metric is kept constant and others varied. This threat also exists in our study as carrying such experiments is extremely difficult.

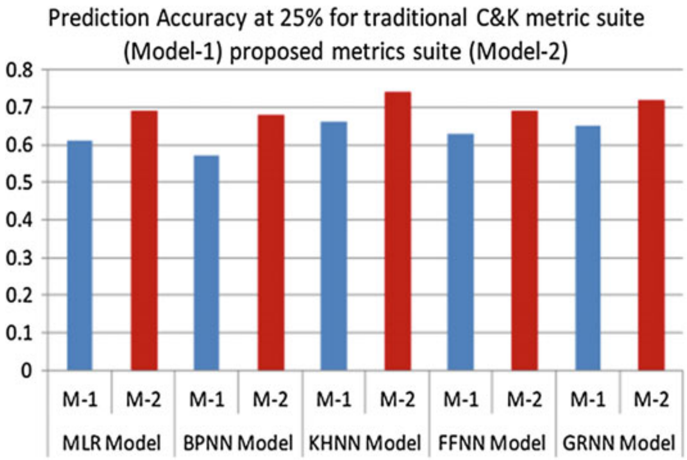


Fig. 2 Pred(0.25) for Model-1 and Model-2

8 Conclusion and Future Work

The goal of our research was to empirically examine the effectiveness of new proposed metric suite for predicting software maintainability for data intensive applications as it's important to give equal attention to the database accesses with the increase in data as well as the number of times data get accessed. We employed MLR, BPNN, FFNN, KN, and GRNN techniques for making software maintainability prediction model. Observing five proprietary software over a period of 3 years, we analyzed the performance of proposed metric suite using prediction accuracy measures such as MRE, MMRE and pred(0.25). Four more metrics were added (MI and CC for measuring the structural complexity and NODBC and SCCR for measuring the database aspect) to the traditional C&K metrics suite. Main results of the current study are summarized as follows:

- The predicted results indicate that proposed metric suite is significant indicator of software maintainability, as improvements in all five datasets were observed when four more metrics added to the C&K metric suite.
- The results received from pearson's correlation coefficient safely suggest that proposed metrics were significantly correlated with change.
- The predicted results indicate that we can use KN in building maintainability prediction models in data intensive applications.
- Multivariate analysis using stepwise linear regression identified NODBC and SCCR as good indicator of software maintainability in data intensive applications.

Result of this study helps practitioners in using new metric suite for developing maintainability prediction models. The results help us in identification of those classes which require big share of maintenance resources and the limited resources can be planned accordingly. The results of our study are valid for medium systems developed in C#. In future, we plan to replicate our studies on data sets having different characteristics such as datasets with different programming languages and environments.

References

1. IEEE Standard 1219–1993. IEEE Standard for Software Maintenance. INSPEC Accession Number: 4493167 doi: 10.1109/IEEESTD.1993.115570 June, 1993
2. Software Engineering Standards Committee of the IEEE Computer Society, IEEE Std. 828–1998 IEEE Standard for Software Configuration Management Plans, http://standards.ieee.org/findstds/standard/828–1998.html
3. K.K. Aggarwal, Y. Singh, A. Kaur, R. Malhotra, Analysis of object-oriented metrics, in *International Workshop on Software Measurement (IWSM)*, 2005
4. R. Bandi, Predicting maintenance performance using object-oriented design complexity metrics. IEEE Trans. Softw. Eng. **29**(1), 77–87 (2003)
5. S. Chidamber, C. Kemerer, A metrics suite for object oriented design. IEEE Trans. Softw. Eng. **20**(6), 476–493 (1994)
6. W. Li, S. Henry, Object-oriented metrics that predict maintainability. J. Syst. Softw. **23**, 111–122 (1993)
7. K.K. Aggarwal, Y. Singh, A. Kaur, R. Malhotra, Application of artificial neural network for predicting maintainability using object oriented metrics. Proc. World Acad. Sci. Eng. Technol. **15**, 285–289 (2006)
8. Y. Zhou, H. Leung, Predicting object-oriented software maintainability using multivariate adaptive regression splines. J. Syst. Softw. **80**(8), 1349–1361 (2007)
9. M. Dagpinar, J.H. Jahnke, Predicting maintainability with object-oriented metrics: an empirical comparison, in *Proceedings of the 10th Working Conference on Reverse Engineering (WCRE ‘03)*, IEEE Computer Society, Washington, 2003
10. M. Thwin, T. Quah, Application of neural networks for software quality prediction using object oriented metrics. J. Syst. Softw. **76**(2), 147–156 (2005)
11. C.V. Koten, A.R. Gray, An application of Bayesian network for predicting object-oriented software maintainability. Inf. Softw. Technol. **48**(1), 59–67 (2006)
12. M.O. Elish, K.O. Elish, Application of TreeNet in predicting object-oriented software maintainability: a comparative study, in *Proceedings of European Conference on Software Maintenance and Reengineering*, 2009
13. C. Jin, J.A. Liu, Applications of support vector machine and unsupervised learning for predicting maintainability using object-oriented metrics. in *Proceedings of the 2nd International Conference on Multi Media and Information Technology*, 2010
14. A. Kaur, K. Kaur, R. Malhotra, Soft computing approaches for prediction of software maintenance effort. Int. J. Comput. Appl. **1**(16), 975–988
15. R. Malhotra, A. Chug, Software maintainability prediction using machine learning algorithms. Softw. Eng. Int. J. **2**(2), 19–36 (2012)
16. W. Li, Another metric suite for object-oriented programming. J. Syst. Softw. **44**(2), 155–162 (1998). doi:10.1016/O164-1212(98)10052-3
17. P. Oman, J. Hagemeister, Metrics for assessing a software system’s maintainability, conference on software maintenance (IEEE Computer Society Press, Los Alamitos, 1992), pp. 337–344
18. P. Oman, J. Hagemeister, Construction and testing of polynomials predicting software maintainability. J. Syst. Softw. **24**, 251–266 (1994)
19. T. McCabe, A complexity measure. IEEE Trans. Softw. Eng. **SE-2**(4), 308–320 (1976)

20. M. Jorgensen, Experience with the accuracy of software maintenance task effort prediction models, IEEE Transc. Softw. Eng. **21**(8), 674–681 (1995)
21. S. Muthanna, K. Kontogiannis, B. Ponnambalam, A. Stacey, Maintainability model for industrial software system using design level metrics, in *Proceedings of 7th Working Conference on Reverse Engineering*, 2000, pp. 248–256
22. F. Fioravanti, P. Nesi, Estimation and prediction metrics for adaptive maintenance effort of object-oriented system. IEEE Trans. Softw. Eng. **27**(12), 1062–84 (2001)
23. S.C. Misra, Modeling design/coding factors that drive maintainability of software systems. Softw. Qual. J. **13**(3), 297–320 (2005)
24. A.D. Banker, A.B. Sultan, H. Zulzalil, J. Din, Applying evolution programming search based software engineering (SBSE), in *Proceedings of Selecting the Best Open Source Maintainability Metrics, International Symposium, ISCAIE, 2012*
25. P. Sun, A. Wang, Application of ant colony optimization in preventive software maintenance policy, in *Proceedings of IEEE international Conference on Information Science and Technology*, China, Mar 2012
26. R. Vivanco, N. Pizzi, Finding effective software metrics to classify maintainability using a parallel genetic algorithm. Genetic Evol. Comput. (Lecture Notes in Computer Science) **30**(13), 1388–1399 (2004)
27. R. Malhotra, A. Chug, An empirical study to redefine the relationship between software design metrics and maintainability in high data intensive applications, in *Proceedings of the World Congress on Engineering and Computer Science 2013, WCECS 2013*. Lecture Notes in Engineering and Computer Science, San Francisco, 23–25 Oct, 2013, pp. 61–66
28. C and C++ Code Counter to Compute OO Metrics and the McCabe Cyclomatic Complexity Metrics from Source Code. http://cccc.sourceforge.net/
29. R. Malhotra, A. Chug, An empirical validation of group method of data handling on software maintainability prediction using object oriented systems, in *Proceedings of International Conference on Quality Reliability InfoCom Technology and Industrial Technology, ICQRITM 2012*, New Delhi, pp. 49–57
30. M.C.J. Hu, Application of the adaline system to weather forecasting. Master thesis, technical report 6775-i, Stanford Electronics Laboratories, 1964
31. D.E. Rumelhart, G.E. Hinton, R.J. Williams, *Learning Internal Presentation by Back-Propagating Errors, the PDP Research Group, Parallel Distributing Processing: Exploration in the Microstructure of Cognition* (MIT Press, MA, 1994)
32. T. Kohonen, *Self-Organization and Associative Memory* (Springer, Berlin, 1989)
33. A.E. Bryson, Y.C. Ho, *Applied Optimal Control: Optimization, Estimation, and Control* (Blaisdell Publishing Company, New York, 1969), p. 481
34. S. Russell, P. Norvig, *Artificial Intelligence: A Modern Approach*, 2nd edn. (Prentice Hall, Upper Saddle River, 2003)
35. D.F. Specht, A general regression neural network. IEEE Trans. Neural Netw. **2**(6), 568–576 (1991)
36. B.A. Kitchenham, L.M. Pickard, S.G. MacDonell, M.J. Shepperd, What accuracy statistics really measure. IEE Proc. Softw. **148**(3), 81–85 (2001)

Chapter 14
High-Assurance Cryptography for Web-Based Enterprises

William R. Simpson and Coimbatore Chandersekaran

Abstract Each web service and each infrastructure service has a need for symmetric and asymmetric encryption, as well as signature processing and other cryptographic processes. A number of specialized cryptographic functions have been developed for hardware and network operations. Their use is appropriate for network level operations. For purposes of this chapter, the discussion is limited to IP enabled communications and similar algorithms. Cryptography is used by most of the services in an enterprise. Asymmetric encryption is performed in suitably security hardened stores and symmetric encryption is performed in most bi-lateral operations. Signatures for integrity and trust use are pervasive. Key management is required throughout the enterprise. The crypto services may be used through all of the Open Systems Interconnection (OSI) model layers, however, this document concentrates on layers 4 and above. The pace of development of computer systems has led to a need for greater and greater key lengths to insure that keys used for encryption cannot be easily discovered. The chapter reviews many of the cryptographic algorithms in use and recommends those that will provide high assurance systems with adequate protection. The chapter also reviews the computational losses in bit effectiveness and provides an algorithm for computing the bits required for levels of protection.

Keywords Cryptographic hash · Cryptography · Electronic signature · Key discovery · Parallel processing · Public key infrastructure · Transport layer security · Web services

W. R. Simpson (✉) · C. Chandersekaran
Institute for Defense Analyses, 4850 Mark Center Drive, Alexandria, VA 22311, USA
e-mail: rsimpson@ida.org

C. Chandersekaran
e-mail: cchander@ida.org

H. K. Kim et al. (eds.), *Transactions on Engineering Technologies*,
DOI: 10.1007/978-94-017-9115-1_14,

1 Introduction

This paper is based in part on a paper published by WCECS [1]. Cryptographic services are pervasive in a high-assurance enterprise. Some terminology that is discussed follows.

- *Plaintext*—the material that is to undergo cryptographic operations.
- *Key*—The unique mathematical parameter used in the cryptographic operations to generate a ciphertext.
- *Cipher*—The algorithm that is used to apply a key and a cryptographic operation.
- *Encryption*—The act of applying the cipher to the key and plaintext, resulting in ciphertext.
- *Ciphertext*—the results of encryption of a plaintext.
- *Decryption*—The act of applying the cipher to the key and the ciphertext, resulting in plaintext.

Cryptographic services and functions need to be standardized to ensure interoperability. These services provide security protection for data and messages and support security properties such as confidentiality, integrity, non-repudiation, authentication and authorization. Functions to sign, validate, and timestamp messages are provided via mechanisms approved for enterprise usage. Authentication and signature cryptographic functions employed in the enterprise are based on the use of the Public Key Infrastructure (PKI) public/private key pairs. The strength of the protection mechanisms for these functions relies on the private key portion remaining secret. For services and devices, private keys are kept in a secure store. Cryptographic Functions in WS Security—Protection for service calls follow the WS Security standards [2] for protecting messages. Several supporting services are required; Extensible Markup Language XML Canonicalization [3], Hashing function. Secure Hash Algorithm (SHA)-512 [4], and strong random number generation (Strong implies uniform distribution of values over the appropriate range). Java provides many cryptographic services through the Java Cryptography Architecture (JCA) framework. The libraries used to do cryptographic operation must be certified [5]. Data at Rest—Data at Rest may be encrypted to protect the confidentiality and integrity of the data. If encrypted, the keys are derived from the primary credentials of the owner of the data. Data in Motion—All data in motion are encrypted and transported using Transport Layer Security (TLS) with mutual authentication [6]. End-to-end encryption is used for all active entity communications.

This chapter is divided into the following sub-paragraphs for the handling of dependent functions; Cryptographic Keys, Encryption, Decryption, Hash Function, Signatures and Key Lengths.

2 Cryptographic Keys and Key Management

Each active entity has an asymmetric key pair issued as part of their Enterprise X.509 [7] PKI certificate. These keys are RSA 2048 bit asymmetric keys. Symmetric keys are generated for Transport Layer Security and other forms of communication because their computation is more efficient.

2.1 Asymmetric Key Pairs

Asymmetric key pairs are generated in a secure process, and PKI certificates are issued for all active entities. All current enterprise PKI certificates use RSA 2048 bit asymmetric key generation. Users are issued Smart Cards through the normal identification issuance process. All other active entities (machines, devices, servers, web services, and other OSI level-5 + software) are issued enterprise X.509 PKI certificates.

2.1.1 RSA Key Generation

The keys for the RSA algorithm are generated by choosing two distinct prime numbers p and q. For security purposes, the integers p and q should be chosen at random, and should be of similar bit-length.

$$\mathbf{Compute\,n = pq;} \tag{1}$$

n is used as the modulus for both the public and private keys

$$\mathbf{Compute\,\varphi(n) = (p - 1)(q - 1);} \tag{2}$$

where φ is Euler's totient function. Choose an integer e such that $1 < e < φ(n)$ and gcd (e, φ(n)) = 1, i.e. e and φ(n) are coprime. e is released as the public key exponent. The bit length for the enterprise cryptographic operation is 2,048.

$$\mathbf{Determine\,d = e{-}1\,mod\,\varphi(n);} \tag{3}$$

i.e. d is the multiplicative inverse of e mod φ(n). This is more clearly stated as solve for d given (d*e) mod φ(n) = 1. d is kept as the private key exponent.

The public key consists of the modulus n and the public (or encryption) exponent e. The private key consists of the private (or decryption) exponent d which must be kept secret.

2.2 Generating Symmetric Keys

Symmetric keys are developed for a number uses. One of main reasons to do so is the algorithms for encrypt and decrypt are faster using symmetric keys than similar algorithms for asymmetric keys.

2.2.1 TLS Mutual Authentication Key Production

TLS Mutual Authentication requires that the TLS client-side holds a certificate (all active entities are required to hold Enterprise X.509 certificates). TLS involves three basic phases; peer negotiation for algorithm support, key exchange and authentication, and symmetric cipher encryption and message authentication.

2.2.2 Other Key Production

Symmetric keys may be generated for a number of other reasons, including safe store of data during operations and encryption of memory for protection. Keys must be generated using a strong Random Number Generator and must be at least 512 bits in length. When a random number generator is used in the key generation process, all values are generated randomly or pseudo-randomly such that all possible combinations of bits and all possible values are equally likely to be generated. Intermediate key generation states and values are not accessible outside of the module in plaintext or otherwise unprotected form [8].

2.3 Store Keys

All asymmetric keys are generated in a security hardened store such as a hardware storage module, HIC or other store that is hardened against theft, tampering or destruction. Storing of session keys are usually kept in memory by the server and browser, and should be stored in a memory space that is unavailable to other programs.

2.4 Delete Keys

A cryptographic module of any web service in an enterprise utilizes an internal capability to zeroize all plaintext cryptographic keys and other unprotected critical security parameters within the module when the keys are no longer in use. Zeroization of cryptographic keys and other critical security parameters is not required if the keys and parameters are either encrypted or otherwise physically or logically protected (e.g., contained within an additional embedded FIPS 140-2 cryptographic module).

3 Encryption

In cryptography, encryption is the process of transforming information (plaintext) using an algorithm (cipher) to make it unreadable to anyone except those possessing special knowledge, usually referred to as a key. The result of the process is ciphertext. Encryption can be used to protect data "at rest", such as files on computers and storage devices, or to protect data in transit, for example data being transferred via networks. There have been numerous reports of data in transit being intercepted in recent years [9].

3.1 Symmetric Versus Asymmetric Encryption Algorithms

Symmetric encryption algorithms encrypt and decrypt with the same key. The main advantages of symmetric algorithms are its security and high speed. Asymmetric encryption algorithms encrypt and decrypt with different keys. Data is encrypted with a public key, and decrypted with a private key (or vice versa). Generally, symmetric encryption algorithms are much faster to execute than asymmetric ones. In practice they are often used together, so that a public-key algorithm is used to encrypt a randomly generated encryption key, and the random key is used to encrypt the actual message using a symmetric algorithm.

3.1.1 Asymmetric Encryption

There are a number of algorithms in use, for reference; Rivest, Shamir and Adleman—(RSA), Digital Signature Algorithm—(DSA), Pretty Good Privacy—(PGP). Keys that are RSA [10] compatible are generated for the enterprise X.509 certificates [11].

RSA Asymmetric Encryption

RSA uses public and private keys that are functions of a pair of large prime numbers based on the difficulty of factoring large integers. The keys used for encryption and decryption in RSA algorithm, are generated using random data. The key used for encryption is a public key. Public keys are stored anywhere publicly accessible. The sender of message encrypts the data using public key, and the receiver decrypts it using his/her own private key.

3.1.2 Symmetric Encryption

Symmetric Encryption uses the same secret key to encrypt and decrypt information or there is a simple transform between the two keys. A secret key can be a number, a word, or just a string of random letters. Symmetric algorithms require that both the sender and the receiver know the secret key.

Symmetric-key algorithms [11]—Symmetric-key algorithms can be divided into Stream algorithms (Stream ciphers) and Block algorithms (Block ciphers).

Stream Ciphers

Stream ciphers encrypt the bits of information one at a time—operate on 1 bit (or sometimes 1 byte) of data at a time (encrypt data bit-by-bit). Stream ciphers are faster and smaller to implement than block ciphers, however, they have an important security gap. If the same key stream is used, certain types of attacks may cause the information to be revealed. This document does not specify stream ciphers.

Block Ciphers

Block cipher (method for encrypting data in blocks) is a symmetric cipher which encrypts information by breaking it down into blocks and encrypting data in each block. A block cipher encrypts data in fixed sized blocks (commonly of 64 bits). The most used block ciphers are Triple DES and AES. The enterprise uses AES for block ciphers.

AES/Rijndael Encryption

AES stands for Advanced Encryption Standard. AES is a symmetric key encryption technique which replaces the commonly used Data Encryption Standard (DES). It was the result of a worldwide call for submissions of encryption algorithms issued by the US Government's National Institute of Standards and Technology (NIST) in 1997 and completed in 2000. The winning algorithm, Rijndael, was developed by two Belgian cryptologists, Vincent Rijmen and Joan Daemen. [12]. The AES algorithm uses three key sizes: a 128-, 192-, or 256-bit encryption key. Each encryption key size causes the algorithm to behave slightly differently, so the increasing key sizes not only offer a larger number of bits with which you can scramble the data, but also increase the complexity of the cipher algorithm.

Description of the AES Cipher

AES is based on a design principle known as a Substitution permutation network. It is fast in both software and hardware. Unlike its predecessor, DES, AES does not use a Feistel network. AES has a fixed block size of 128 bits and a key size of 128, 192, or 256 bits, whereas Rijndael can be specified with block and key sizes in any multiple of 32 bits, with a minimum of 128 bits. The blocksize has a maximum of 256 bits, but the keysize has no theoretical maximum. AES operates on a 4 × 4 matrix of bytes, termed the state (versions of Rijndael with a larger block size have additional columns in the state). The AES cipher is specified as a number of repetitions of transformation rounds that convert the input plaintext into the final output of ciphertext. Each round consists of several processing steps, including one that depends on the encryption key. The key cipher algorithm process involves matrix transformations [12]. Hardware for performing these functions exists, and may be used. Intel cores have added instruction sets to speed the algorithms.

Data Encryption Standard (DES)

Data Encryption Standard (DES) is a symmetric block cipher developed by IBM. The algorithm uses a 56-bit key to encipher/decipher a 64-bit block of data. The key is always presented as a 64-bit block, every 8th bit of which is ignored. The algorithm is best suited to implementation in hardware. DES is the most widely used symmetric algorithm in the world, despite claims that the key length is too short. Ever since DES was first announced, controversy has raged about whether 56 bits is long enough to guarantee security. NIST recommended that DES should be replaced by Triple DES, a modified version employing 112- or 168-bit keys. DES's versatility also was limited because it worked only in hardware, and the explosion of the Internet and e-commerce led to much greater use and versatility of software than could have been anticipated by DES's designers. As DES's vulnerabilities became apparent, NIST opened an international competition in 1997 to find a permanent replacement for DES (see AES cipher above).

Triple DES

Triple DES is a variation of DES. It uses a 64-bit key consisting of 56 effective key bits and 8 parity bits. The size of the block for Triple-DES is 8 bytes. Triple-DES encrypts the data in 8-byte chunks. The idea behind Triple DES is to improve the security of DES by applying DES encryption three times using three different keys.

4 Decryption

Decryption is the complement of encryption. In cryptography, decryption is the process of transforming information (ciphertext using an algorithm (cipher) to make it readable to an entity that possesses special knowledge (key). The result of the process is decrypted information (in cryptography, referred to as plaintext).

4.1 Asymmetric Decryption

Since we have specified that asymmetric encryption and decryption is by PKI properties of the enterprise certificate for active entities we use the RSA algorithm by inverting the encryption process and using the complementary key.

4.2 Symmetric Decryption

For symmetric encryption we have limited the possible algorithms to AES and Triple DES.

The AES decryption algorithm is applied as follows: Recall: The AES cipher is specified as a number of repetitions of transformation rounds that convert the input plaintext into the final output of ciphertext. Each round consists of several processing steps, including one that depends on the encryption key. A set of reverse rounds are applied to transform ciphertext back into the original plaintext using the same encryption key.

5 Hash Function

A hash function is any algorithm or subroutine that maps plaintext to a value, called keys. For example, a value can serve as an index to list of plaintext items. Hash functions are mostly used to accelerate table lookup or data comparison tasks such as finding items in a database, detecting duplicated or similar records in a large file, finding similar stretches in DNA sequences, and so on. In cryptographic services, a hash is used to nearly uniquely identify the content of a plaintext. Some hash functions may map two or more plaintexts to the same hash value, causing a collision. Cracking a hash code would be devising a set of rules by which one change can be compensated with other changes to make the hash the same. Such hash functions try to map the plaintext to the hash values as evenly as possible because, as Hash Tables fill up, collisions become more frequent.

5.1 Hash Function Algorithms

For most types of hashing functions the choice of the function depends strongly on the nature of the input data, and their probability distribution in the intended application, in this case we use plaintext which consists of SAML tokens, messages in SOAP envelopes, log records, etc. The hash is part of the almost unique identifier bound to a signature.

5.2 Hashing with Cryptographic Hash Function

Some cryptographic hash functions, such as SHA-3/SHA-512, have stronger uniformity guarantees than checksums or fingerprints, and thus can provide very good general-purpose hashing functions. A cryptographic hash function is a deterministic procedure that takes an arbitrary block of data (plaintext) and returns a fixed-size bit string; the (cryptographic) hash value, such that an accidental or intentional change to the data changes the hash value. The data to be encoded (plaintext) is often called the “message,” and the hash value is sometimes called the message digest, digest or hash. The ideal cryptographic hash function has four main or significant properties; it is easy to compute, but it is unlikely to: generate a message that has a given hash, change a message and not change the hash, or find different messages with the same hash. There are a large number of cryptographic hash algorithms but for the enterprise we restrict usage to two such constructs:

- MD-5 Message Digest 5
- SHA-512 Secure Hash Algorithm 512 bits.

5.2.1 MD-5

The MD5 Message-Digest Algorithm is a widely used cryptographic hash function that produces a 128-bit (16-byte) hash value. MD5 is commonly used to check data integrity. In an attack on MD5 published in December 2008, a group of researchers used this technique to fake SSL certificate validity. US-CERT has advised that MD5 “should be considered cryptographically broken and unsuitable for further use” and most U.S. enterprise applications now require the SHA-2 family of hash functions. For this reason, MD-5 is not to be used in standalone operations but may be used in HashMAC (HMAC) construction for TLS.

5.2.2 SHA-3 Defined SHA-512

In cryptography the Secure Hash Algorithm (SHA), SHA-2 is a standard that defines a set of cryptographic hash functions by bit strength (SHA-224, SHA-256, SHA-384, and SHA-512) designed by the National Security Agency (NSA) and published in 2001 by the NIST as a U.S. Federal Information Processing Standard. SHA stands for Secure Hash Algorithm. SHA-2 includes a significant number of changes from its predecessor, SHA-1. In October 2012, the National Institute of Standards and Technology (NIST) chose a new algorithm (Keccak) as the algorithm to be used in the SHA-3 standard [13]. A notable problem in moving from SHA-1 to SHA-2 is that they both used the same algorithmic approach (Merkle-Damgard), to process message text. This means that a full attack on SHA-1 becomes a potential threat for SHA-2. While no successful attacks against a full-round SHA-2 have been announced, there is no doubt that attack mechanisms are being developed in private. For the enterprise the goal is that SHA-3/SHA-512 be used for all signature data. The SHA-3 standard has yet to be published even though the algorithm has been chosen. However, because of the compromises of other hash algorithms, and the time required to implement new models, a transition to SHA-3 defined SHA-512 should be undertaken. It may be combined with MD-5 for HMAC applications in TLS.

6 Signatures

A digital signature is an information element that can be added to a document that can be used to authenticate the identity of the generator or the signer of a document, and ensure that the original content of the document that has been sent is unchanged. Digital signatures are easily transportable, cannot be imitated by someone else. The digital signature has specific content elements that include an encrypted hash of the document, hash details, a time stamp, and the X.509 certificate of the signer. An important feature of each signature is the ability to validate the signature. The attachment of the X.509 of the signer which must be verified and validated through a number of steps:

- The certificate is issued by a trusted certificate authority
- The certificate date is valid and the distinguished name of the signer matches the certificate.
- The certificate is not revoked

The hash must be validated by decrypting with the public key of the signer, and comparing it to an independently computed hash. For the enterprise we consider three basic signatures:

- XML Signatures
- S/MIME Signatures Secure/Multipurpose Internet Mail Extensions (S/MIME)
- E-Content Signatures

6.1 XML Signature

XML signatures are digital signatures designed for use in XML transactions. They are defined by a standard schema for capturing the result of a digital signature operation applied to arbitrary (but often XML) data. Like non-XML-aware digital signatures (e.g., PKCS), XML signatures add authentication, data integrity, and support for non-repudiation to the data that they sign. However, unlike non-XML digital signature standards, XML signature has been designed to both account for and take advantage of the Internet and XML. A fundamental feature of XML

Signature is the ability to sign only specific portions of the XML tree rather than the complete document. The entire return may be encrypted for security [14].

6.2 S/MIME Signature

S/MIME (Secure/Multipurpose Internet Mail Extensions) provides a consistent way to send and receive secure MIME data. Based on the popular Internet MIME standard, S/MIME provides the following cryptographic security services for electronic messaging applications: authentication, message integrity and non-repudiation of origin (using digital signatures), and data confidentiality (using encryption). S/MIME can be used by traditional mail user agents (MUAs) to add cryptographic security services to mail that is sent, and to interpret cryptographic security services in mail that is received. However, S/MIME is not restricted to mail; it can be used with any transport mechanism that transports MIME data, such as HTTP. S/MIME [15] provides one format for enveloped-only data, several formats for signed-only data, and several formats for signed and enveloped data.

6.3 E-Content Signature

The Digital Signature Standard (DSS) (FIPS 186-3) developed by NIST is used for general E-Content signing. This Standard specifies algorithms for applications requiring a digital signature, rather than a written signature. A digital signature is represented in a computer as a string of bits. A digital signature is computed using a set of rules and a set of parameters that allow the identity of the signatory and the integrity of the data to be verified. Digital signatures may be generated on both stored and transmitted data. Signature generation uses a private key to generate a digital signature; signature verification uses a public key that corresponds to, but is not the same as, the private key. Each signatory possesses a private and public key pair. Public keys may be known by the public; private keys are kept secret. Anyone can verify the signature by employing the signatory's public key. Only the user that possesses the private key can perform signature generation. A hash function is used in the signature generation process to obtain a condensed version of the data to be signed; the condensed version of the data is often called a message digest. The message digest is input to the digital signature algorithm to generate the digital signature. The hash functions to be used are specified in the Secure Hash Standard (SHS) [4]. FIPS approved digital signature algorithms are used with SHA-512. The digital signature is provided to the intended verifier along with the signed data. The verifying entity verifies the signature by using the claimed signatory's public key and the same hash function that was used to generate the signature. Similar procedures may be used to generate and verify signatures for both stored and transmitted data [4].

7 A Note on Cryptographic Key Lengths

For the high-assurance process all communications are encrypted using TLS 1.2 for confidentiality. Private keys are stored in Hardware Security Modules (HSMs). All Cryptography is done locally (no enterprise cryptographic services). Software Cryptographic Suites used are FIPS 140 approved. In order to develop an enterprise solution, all processes, including the cryptographic procedures must be published and openly available. As a result, adversaries know the processes and algorithms to be used.

7.1 Encryption Key Discovery

In a large scale enterprise, all crypto methods (algorithms—for example AES 256) and all message types must be published. This is needed to get all administrators, software developers and providers on the same path. Two approaches or lines of attack are possible. The first is to find a flaw in the algorithm or computer code. This latter requires very few computations to discovery. Assuming the code is clean and the algorithm is tight, we can try guessing the key. Brute Force Decryption Requires a Known Number of Compute Cycles.

- Pick a feasible key (not all keys may be feasible). Example-asymmetric keys must be prime!
- Apply decryption algorithm
- Examine output and look for recognizable material.
  - If recognize—Done apply key to all session Packets
  - If not recognize—go to 1

Continue until all feasible keys are tried or discovery occurs (Fig. 1).

7.2 The High-Performance Dilemma

The locality of cryptographic services makes scaling through high-performance multi-core thread management easier and eliminates the need for front-end load balancing. However, maintaining confidentiality with such computing capability available to an adversary is problematic. The encryption methodology must be standard and published so that all elements of the enterprise can obtain the proper software and hardware to perform the needed operations. Rogue agents (including insider threats) may be present and to the extent possible, we should be able to operate in their presence, although this does not exclude their ability to view and export some activity. Key extraction from encrypted data is inherently a parallel function. For example, all of the possible keys can be distributed to many cores

Key =xxxxxxxxxxxxx (n bits)
A compute cycle 1 Key =0000000000000 (n bits)
A compute cycle 1 Key =0000000000001 (n bits)
A compute cycle 1 Key =0000000000010 (n bits)
A compute cycle 1 Key =0000000000011 (n bits)
⋮

Normal Computing.	Fast Computing.	High Performance Computing.
Total guesses = 2^n compute cycles Expect to do 50% of guesses For 20 bits expect 2^{19} compute cycles 524288 compute cycles Assume 1 milli-sec/ compute cycle 20 bits = 524 seconds (8.73 minutes) 30 bits 149.1 hours = 6.2 days 32 bits = 24.9 days 256 bits 1.8358715e+66 years Etc.	Total guesses = 2^n compute cycles Expect to do 50% of guesses For 20 bits expect 2^{19} compute cycles 524288 compute cycles Assume 1 micro-sec/ compute cycle 20 bits = 0.5 seconds 30 bits = .1491 hours = 8.9 minutes 32 bits = 3.5 hours 256 bits 1.8358715e+63 years Etc.	Total guesses = 2^n compute cycles Expect to do 50% of guesses For 20 bits expect 2^{19} compute cycles 524288 compute cycles Assume 1 pico-sec/ compute cycle 20 bits = 0.0005 seconds 30 bits = .0001491 hours = 0.5 seconds 32 bits = 12.6 seconds 256 bits 1.8358715e+60 years Etc.

Time can be reduced by choosing only feasible keys. Finding a flaw in the computer algorithm can be very quick.

Fig. 1 Encryption key discovery by brute force

and each can spend a few cycles attempting to decrypt encrypted packets. When one recognizes the output then the key is discovered. This parallel decomposition is may be across cores in a parallel array.

7.3 The Mathematics of Parallel Decomposition of Key Discovery

For a given key length, m, the number of possible keys is given by 2 m and this represents the sequential computational burden for trying each possible key. Further, the first n bits may be distributed among cores (c) for processing at the scale of c = 2n. The remaining bits (m-n) may be sequentially processed with a total processing burden on each processor equal to 2 m-n. This results in a confidentiality effectiveness loss of n bits. Note that this loss is independent of the original key length, m.

If the key length is only 10 bits, then the number of possible keys is 1,024. If this is spread over 1,024 cores, then a single decrypt cycle (a decrypt cycle includes applying the decryption process and applying a recognition algorithm to the answer. This could be 10 machine cycles or less) reveals the key and provide an adversary a way to overcome confidentiality. If the key length is only 12 bits, then the possible keys that encrypted the packet is 4,096. If this is spread over 1,024 cores, then 4 decrypt cycles reveals the key and provide an adversary a way to overcome confidentiality. This latter is equivalent to weakening the cryptographic key by ten bits. Table 1 provides a summary of the discovery process.

From Table 1, it can be seen that high-performance computing over 500 k cores is equivalent to a loss of 19 bits in the strength of encryption, and a million cores essentially reduces the bit effectiveness by 20 bits. This leads to a race that

Table 1 Loss of Bit Effectiveness

Distributed key bits (*n*)	Values spread over cores (*c*)	Confidentiality effectiveness loss (bits)	Distributed key bits (*n*)	Values spread over cores (*c*)	Confidxentiality effectiveness loss (bits)
1	2	1	16	65,536	16
2	4	2	17	131,072	17
3	8	3	18	262,144	18
4	16	4	19	524,288	19
5	32	5	20	1048,576	20
6	64	6	21	2097,152	21
7	128	7	22	4194,304	22
8	256	8	23	8388,608	23
9	512	9	24	16777,216	24
10	1,024	10	25	33554,432	25
11	2,048	11	26	67108,864	26
12	4,096	12	27	134217,728	27
13	8,192	13	28	268435,456	28
14	16,384	14	29	536870,912	29
15	32,768	15	30	1073741,824	30

technology has delivered to us, and the response is to increase the bits in the encryption process. The number of bits one needs is related to the time nit takes for brute force key discovery and it is recommended that this is supplemented by 20 bits. The increased bit length adds computational burden to the normal computational process increasing the need for high-performance computing. This race can only be broken by developing a non-parallel decomposable decryption process, which is not currently available. The cryptographic key lengths in this document have been increased by 20 bits and rounded to the next power of 2. In certain cases, reports of compromise on crypto methods may result in a recommendation for a different algorithm or increased bit lengths.

8 Summary

We have reviewed a set of cryptographic processes for confidentiality, integrity and authentication. Web-based architectures require cryptographic services at the application layer. The high assurance architecture demands strong cryptographic services for data both in transit and at rest. The strength of cryptographic functions in today's environment is beyond previous recommendations due to the improvements in computing and the adaptability of the threat. The cryptographic process is part of comprehensive enterprise architecture for high assurance that is web-service based and driven by commercial standards. Portions of this architecture are described in references [16–18].

References

1. Coimbatore Chandersekaran, W.R. Simpson, Cryptography for a high-assurance web-based enterprise, lecture notes in engineering and computer science, in *Proceedings World Congress on Engineering and Computer Science*, WCECS2013, pp. 23–28, San Francisco (2013)
2. OASIS Open Set of Standards, WS-Security Specification 1.1, OASIS, Nov 2006
3. World Wide Web Consortium (W3C), Canonical XML version 1, March 2001
4. National Institute of Standards, FIPS PUB 180-3. Secure Hash Standard, Gaithersburg, Aug 2002
5. National Institute of Standards, FIPS PUB 140-2, Security Requirements for Cryptographic Modules, Gaithersburg, 25 May 2001
6. Internet Engineering Task Force (IETF) Standards, RFC 5246 The Transport Layer Security Protocol 1.2, 2008
7. Internet Engineering Task Force (IETF) Standards, RFC 2459 Internet X.509 Public Key Infrastructure Certificate and CRL Profile, Jan 1999
8. National Institute of Standards, FIPS PUB 186-3, Digital Signature Standard, Gaithersburg, June 2009
9. S.K. Miller, Fiber Optic Networks Vulnerable to Attack, Information Security Magazine, 15 Nov 2006
10. PKCS #1, RSA Cryptography Standard, http://www.rsa.com/rsalabs/node.asp?id=2125
11. Some material adapted from InfoBlox, http://www.encryptionanddecryption.com/encryption/asymmetric_encryption.html
12. National Institute of Standards, FIPS 197, Advanced Encryption Standard (AES), Gaithersburg, Nov 2001
13. J.R.C. Cruz, Dobb's, Keccak: The New SHA-3 Encryption Standard, The World of Software Development, http://www.drdobbs.com/security/keccak-the-new-sha-3-encryption-standard/240154037 see http://keccak.noekeon.org/. May 2013
14. World Wide Web Consortium (W3C), XML Encryption Syntax and Processing 10 Dec 2002
15. Internet Engineering Task Force (IETF) Standards, RFC 5751 Secure/Multipurpose Internet Mail Extensions (S/MIME) Version 3.2 Message Specification, July 2010
16. W.R. Simpson, Coimbatore Chandersekaran, A. Trice, A persona-based framework for flexible delegation and least privilege, Electronic Digest of the 2008 System and Software Technology Conference, Las Vegas, Nevada, May 2008
17. Coimbatore Chandersekaran, W.R. Simpson, The case for bi-lateral end-to-end strong authentication, World Wide Web Consortium (W3C) Workshop on Security Models for Device APIs, 4 pp, London, Dec 2008
18. W.R. Simpson, Coimbatore Chandersekaran, Information sharing and federation, in *The 2nd International Multi-Conference on Engineering and Technological Innovation,* IMETI2009, Orlando, vol. I, pp. 300–305, July 2009

Chapter 15
Algorithm Based Partial Reconfiguration with Application on Matrix Inverse Computations

Etienne Aubin Mbe Mbock

Abstract Partial reconfiguration of algorithms is becoming increasingly attractive in many computational applications. This research article covers two aspects of the reconfiguration approach: The first aspect shows that partial reconfiguration is capable of reconstructing computations. The second aspect will construct a theoretical hardware device that realises these computations. With this research article, we analyse the importance of partial reconfiguration for algorithms in one hand and in the second hand we use and apply this concept for the invention of a method that computes two matrices that are inverses of each other. In this paper we specify the computation of two inverse upper and lower matrices using the partial dynamic reconfigurability concept. We propose for this novel algorithm a pseudo code implementation and its hardware construction.

Keywords Algorithm · Analysis · Computations · Construction · FPGA · Hardware · Inverse matrix · Partial · Process · Reconfiguration

1 Introduction

Partial reconfiguration computing is nowadays a subject that is establishing itself as a basic scientific field. This concept is especially developed in computing sciences and computer technology [1–4]. It covers various classical subjects including Xilinx technologies, FPGA's construction, migration towards control theory with the implementation of the Kalman Filter on FPGA, algorithm analysis with the creation of the recursive dynamic process and engineering fields with algorithm optimization. Reconfiguration in hardware implies the dynamic

E. A. M. Mbock (✉)
Institut für Technische Informatik Heidelberg, Heidelberg, Germany
e-mail: E.Mbe_Mbock@stud.uni-heidelberg.de

H. K. Kim et al. (eds.), *Transactions on Engineering Technologies*,
DOI: 10.1007/978-94-017-9115-1_15,

modification of blocks of logic by downloading partial bit files while the remaining logic continues to operate without interruption. The Partial Reconfiguration technology allows designers to change functionalities of their hardware devices. As a consequence, computer technology designers will migrate to devices with lower hardware complexity. Because the benefits of this hardware definition are various including among other performance development, hardware complexity reduction from the technological point of view. It becomes interesting to analyse and apply these features in computing and algorithm creation. For this research article we suppose that we can partially reconfigure any algorithm. This reconfiguration is guaranteed by theorems on dynamic partial reconfiguration of algorithms. We will admit that the general recursive linear, specified by:

$$\left\{ \mathrm{q}_1,\ \mathrm{q}_\mathrm{j} = \sum_{\mathrm{i}=1}^{\mathrm{j}-1} \alpha_{\mathrm{ji}}\mathrm{q}_\mathrm{i}, \mathrm{j} \in \{2, 3, \cdots, \mathrm{N}\} \right\}$$

is given. All matrix operations are performed on two initial $n \times n$ matrices A, B with entries

$$\left\{ \mathrm{a}_{\mathrm{i,\,j}},\ 1 \leq \mathrm{i} \leq \mathrm{n},\ 1 \leq \mathrm{j} \leq \mathrm{n} \right\}$$

these matrices will be considered as n-length column vectors, represented by:

$$\begin{aligned} \mathrm{A} := [\mathrm{A}_1 \quad \mathrm{A}_2 \quad \cdots \quad \mathrm{A}_\mathrm{n}] = \left(\mathrm{A}_\mathrm{j},\ \mathrm{j} \in \{1, 2, \cdots,\ \mathrm{n}\}\right) \\ \mathrm{B} := [\mathrm{B}_1 \quad \mathrm{B}_2 \quad \cdots \quad \mathrm{B}_\mathrm{n}] = \left(\mathrm{B}_\mathrm{j},\ \mathrm{j} \in \{1, 2, \cdots,\ \mathrm{n}\}\right). \end{aligned}$$

This research article will provide the advantages of reconfiguration of algorithms, and with these features of reconfiguration we demonstrate the construction of an algorithm. This novel algorithm solves the matrix inversion construction problem. That is, the given two matrices A, B represented by its respective column entries

$$\left(\mathrm{A}_\mathrm{j},\ \mathrm{B}_\mathrm{j},\ \mathrm{j} \in \{1, 2, \cdots,\ \mathrm{n}\}\right),$$

construct with a partial reconfiguration two matrices that are inverse to each other. We state the theorem, demonstrate it and propose an algorithm for the matrix inverses. In addition to this description, we construct the related hardware to this algorithm for future FPGA implementations. This research article assumes some matrix analysis and computation basic results [5–11]. We suppose that the experimentation in this research article takes place in a n-dimensional vector space. The aim of this research article, is the construction of the matrix inverse algorithm and demonstrate that the algorithm performs correctly. The inverted matrix will have the following expansion

$R_{1,1}$	$R_{1,2}$	$\cdots$	$R_{1,k}$	$\cdots$	$R_{1,n}$
0	$R_{2,2}$	$\cdots$	$\cdots$	$\cdots$	$R_{2,n}$
$\vdots$	0	$\ddots$	$\ddots$	$\ddots$	$\vdots$
$\vdots$	$\vdots$	$\ddots$	$R_{k-1,k-1}$	$\cdots$	$R_{k-1,n}$
$\vdots$	$\vdots$	0	$\ddots$	$\ddots$	$\vdots$
0	0	$\cdots$	0	0	$R_{n,n}$

In addition to the proposed algorithm, a theoretical construction of the hardware for the proposed algorithm will be provided.

2 Advantages of Reconfiguration of Algorithms: A Software Point of View

Algorithms that are nowadays software are complex because in a system in which they are to be integrated are increasing, for the simple case of the Kalman Filter Software. If we consider a sequence of n estimations problems p_i, $i \in \{1, 2, \cdots, n\}$. The partial reconfiguration of the software will adapt the Kalman Filter algorithm to each p_i, $i \in \{1, 2, \cdots, n\}$. In this way the partial reconfiguration will not only optimize the Kalman Filter algorithm, but also it creates a new strategy for problem estimation using the software. The software will now have a new functionality depending upon the problem to be solved. This means that partial reconfiguration aims at algorithm optimization in terms of the complexity of the algorithm and its flexibility. This will still have a consequence on the hardware construction. If the algorithm is now optimized per reconfiguration then the constructed hardware will be cheaper for the designer, space, cost and power consumption reduction. These points are general when it comes to partial reconfiguration. In this research we want to give some concrete improvement base on our research:

1. The application of this concept in the special case of the Kalman Filter extends the method with the second condition for optimality, “Six Valid Error Estimation and the Stop Mod reconfiguration”, “Three Optimal Error Covariance Matrix and Stop Mod”, “Three Computation of the Error Covariance and the Stop Mod”
2. Linear Recursive Process Creation and Optimization
3. Matrix [Q, R]-Decomposition
4. Reconfiguration Speed Problem Resolution
5. Real Vectors Coding.

The improvements that we listed are more related to algorithms and computational optimization, which has also been applied in robotics [12, 13]. The idea of partial reconfiguration of algorithms aims at algorithms functionalities extension and this in turn still alters the true hardware complexity of the algorithm. The previous cited advantages of partial reconfiguration of algorithms constitutes the background we need to present the novel algorithm for matrix inverse computations.

3 Principles and Theorems

Theorem 1 *Under consideration of the existence of the recursive dynamic process, there exists two matrices that are upper and lower matrices and inverse to each other.*

Proof

1. Assuming that the recursive dynamic process exist, we can construct the recursive linear process algorithm.

2. Given a n × n matrix A. There is a n × n matrix B associated to the matrix A. The process computes the following entries $R_{i,\ j}$, the process being a linear recursive process will not compute the $R_{i,\ j}$. We then suppose that $R_{i,\ j} = 0$ for j ϵ {0, 1, 2,···, n − 1} a recursive vector will exist and we denote it by $V_j^{(k)}$ a vector of length n. The vector $V_i^{(k)}$ is related to the entries $R_{i,\ j}$ according to the following scheme:

$$V_j^{(k)} = V_j^{(k-1)} - R_{k,j} \cdot V_i^{(k)}$$

   for all $j \in \{1, 2, \cdots, n\}$

3. The matrice R with entries $R_{i,\ j}$ will be computed from the input column matrix A according to the following scheme $R_{k,j} = \begin{cases} V_j^{(k)} \cdot A_j & \text{if} \quad k \neq j \\ \left\| V_j^k \right\| & \text{if} \quad k = j \end{cases}$

4. The commutative product of the matrix R and the column vectors $V_j^{(k)}$ is the identity matrix and completes the proof.

3.1 Construction of the Algorithm and Computations

Assuming that the n × n matrices A are given, and applying Theorem 3, we then propose the following algorithm that will compute two matrices that are inverse to each other. The proof provided above summarized in four steps in which the values of the matrix V are computed according to the recursive formulae:

$$V_j^{(k)} = V_j^{(k-1)} - R_{k,j} \cdot V_j^{(k)}$$

and the entries of the matrix R are all the values $R_{i,j}$ computed in step three of the proof above. We will initialise the matrix R to zero.

Algorithm 2

```
function PROBLEM$(A,B)$
    $[m,n] \leftarrow size(A)$
    $[p,q] \leftarrow size(B)$
    $R \leftarrow zeros(n,n)$
    if $m = p$ or $q = n$ then
        $R \leftarrow zeros(n,n)$
        for $j = 2,3,\cdots,n$ do
            $V(:,j) = B(:,j)$
            for $i = 1 : j-1$ do
                $R(i,j) \leftarrow V(:,i)^t * A(:,j)$
                $V(:,j) \leftarrow$
                $V(:,j) - R(i,j) * V(:,i)$
            end for
            $R(j,j) \leftarrow \|V(:,j)\|$
            $V(:,j) = \frac{V(:,j)}{R(j,j)}$
        end for
    end if
end function
```

The proposed algorithm computes in the case of the Hilbert matrix of order 3 in the following way:

1. $R_{1,1} = \|B_1\| = 1$ and $V_1^{(1)} = (1 \quad 0 \quad 0)^t$
2. $R_{1,2} = V_{(1)}^{(1)} \cdot A_2$ and $V_2^{(1)} = V_2^{(0)} - R_{1,2} \cdot V_1^{(1)},\ R_{2,2} = \left\| V_2^{(1)} \right\|$
3. $V_2^{(2)} \cdot \left\| V_2^{(1)} \right\| = V_2^{(1)}$ and $R_{1,3} = V_{(1)}^{(1)} \cdot A_3$
4. $V_3^{(1)} = V_3^{(0)} - R_{1,3} \cdot V_1^{(1)}$ and $V_3^{(1)} \cdot \left\| V_3^{(1)} \right\| = V_3^{(1)},\ R_{2,3} = V_{(2)}^{(2)} \cdot A_3$
5. $V_3^{(2)} \cdot \left\| V_3^{(1)} \right\| = V_3^{(1)} - V_{(2)}^{(2)} \cdot A_3 \cdot V_2^{(2)},\ R_{3,3} = \left\| V_3^{(2)} \right\|$ and $V_3^{(3)} \cdot \left\| V_3^{(2)} \right\| = V_3^{(2)}$

4 Reconfigurations Analysis

The recursive linear process algorithm is given, we can provide as follows the reconfiguration analysis of this method. This will consider the following algorithm's parts:

1. Inputs parameters, m × n matrices Alpha and Beta.
2. The body of the linear recursive process.

The analysis of reconfiguration of the linear recursive process algorithm aims at optimizing the process towards the construction of two matrices that are inverse to each other. Considering the first part for reconfiguration, we postulate the following principles and facts.

The matrix Gama and Beta of the must be reconfigured into square matrices of the same size. In this paper reconfigure the matrix Beta to the identity matrix.

Principle 3 Parameters Reconfiguration

The starting index $j \in \{2, 3, \cdots, n\}$ of the recursive linear algorithm will reconfigure in $j \in \{2, 3 \cdots, n\}$, Gama_j will reconfigure to $V_j^{(0)}$, The index i will not reconfigure, $R_{i,j}$ will reconfigure to the scalar product $V_i \cdot A_j$, Gama_j will reconfigure to a normed vector

Principle 4 Body Reconfiguration of the Algorithm

The reconfiguration steps are finite, the algorithm converges and compute the expected matrix decomposition

Principle 5 Termination of the Algorithm

5 General Computations Presentation for the Inverses Matrix Computations

The algorithm can be performed in two ways, we propose with the following presentation a combination of top-down and up-down computation model. The second method will be recursive starting from $V_1^{(1)}$. The figures that will next follow give numerical values in the special case of the Hilbert matrix

1. Write $\left\{V_i^{(j)}, R_{j,i}\right\}$, for $j \in \{1, 2 \cdots, n-1\}$ and $i \in \{n, n-1, \cdots, 1\}$.
2. Compute $V_1^{(1)}$.
3. Compute $V_i^{(j)}, R_{j,i}$, for $j \in \{1, \cdots, i-1\}$ and $i \in \{1, 2, \cdots, n\}$.

Figure 1 describes the computational presentation of the algorithm. The numerical values of the algorithm in the special case of a Hilbert n × n matrix gives the values of the matrix V and R. The third column vector of the matrix V corresponds to V_3^3 its formula is given by the following vector

$$\begin{pmatrix} \frac{-a_{13}}{\left\|V_3^{(1)}\right\|\left\|V_3^{(2)}\right\|} - \frac{a_{12}^2 \cdot a_{13}}{\left\|V_1^{(1)}\right\|^2\left\|V_3^{(2)}\right\|} + \frac{a_{23} \cdot a_{12}}{\left\|V_2^{(1)}\right\|\left\|V_3^{(2)}\right\|\left\|V_1^{(1)}\right\|} \\ \frac{a_{12}^2 \cdot a_{13}}{\left\|V_2^{(1)}\right\|^2\left\|V_3^{(2)}\right\|} - \frac{a_{23}}{\left\|V_2^{(1)}\right\|^2\left\|V_3^{(2)}\right\|} \\ \frac{1}{\left\|V_3^{(1)}\right\|^2\left\|V_3^{(2)}\right\|} \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

The matrix entry R(3, 4) is given by the following formulae:

$$\begin{aligned} R_{3,4} = & \left(\frac{-a_{13} \cdot a_{14}}{\left\|V_1^{(3)}\right\| \cdot \left\|V_3^{(2)}\right\|} - \frac{a_{12}^2 \cdot a_{13} \cdot a_{14}}{\left\|V_2^{(1)}\right\|^2 \cdot \left\|V_3^{(2)}\right\|}\right) \\ & + \left(\frac{a_{23} \cdot a_{12} \cdot a_{14}}{\left\|V_2^{(1)}\right\|^2 \cdot \left\|V_3^{(2)}\right\|} + \frac{a_{12} \cdot a_{13} \cdot a_{24}}{\left\|V_2^{(1)}\right\|^2 \cdot \left\|V_3^{(2)}\right\|}\right) \\ & \left(-\frac{a_{23} \cdot a_{24}}{\left\|V_2^{(1)}\right\|^2 \cdot \left\|V_3^{(2)}\right\|} - \frac{a_{34}}{\left\|V_3^{(1)}\right\|\left\|V_3^{(2)}\right\|}\right) \end{aligned} \tag{1}$$

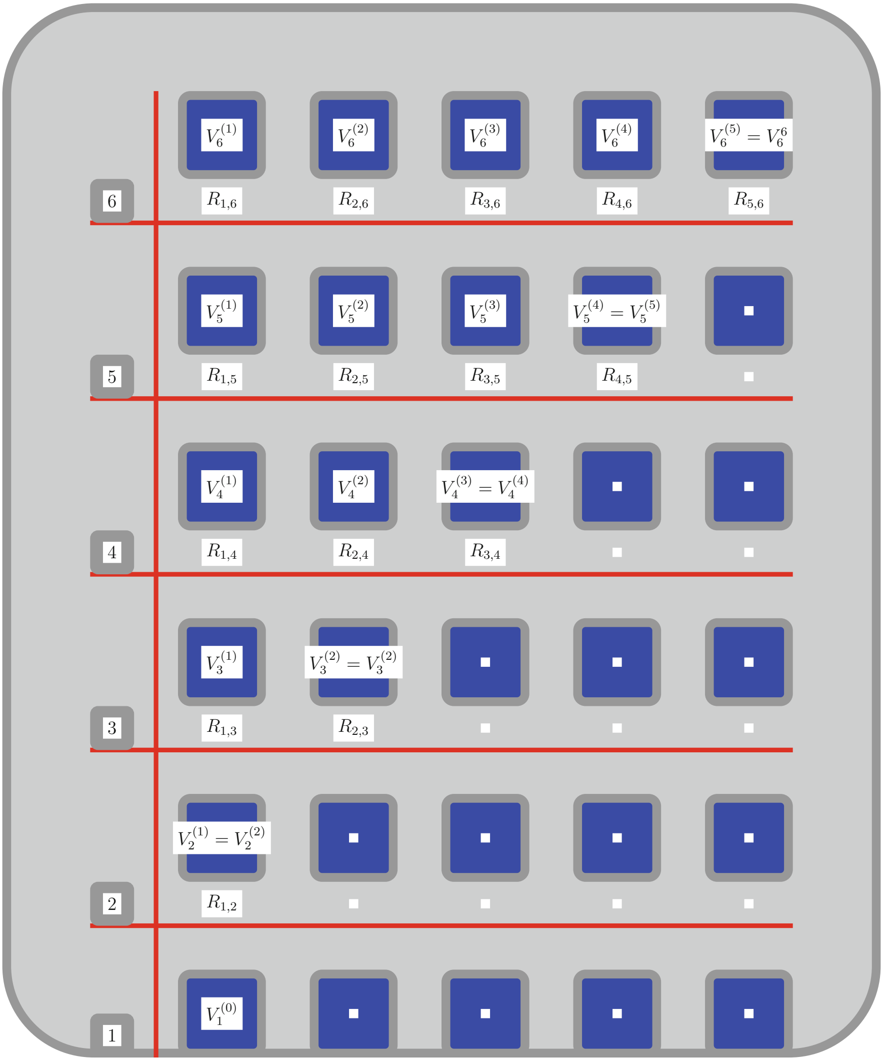
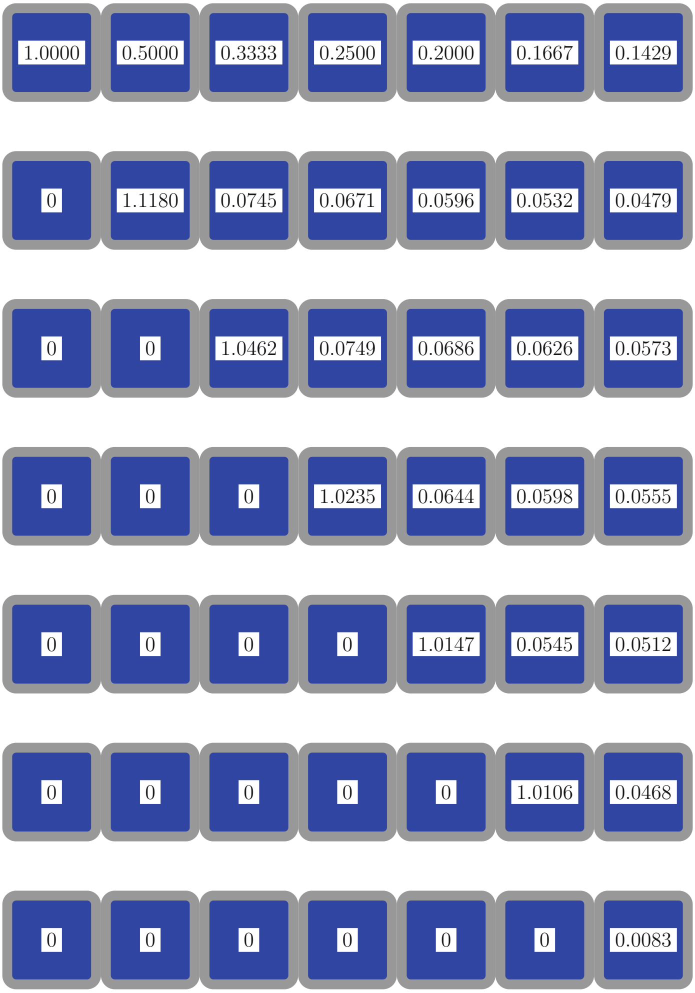
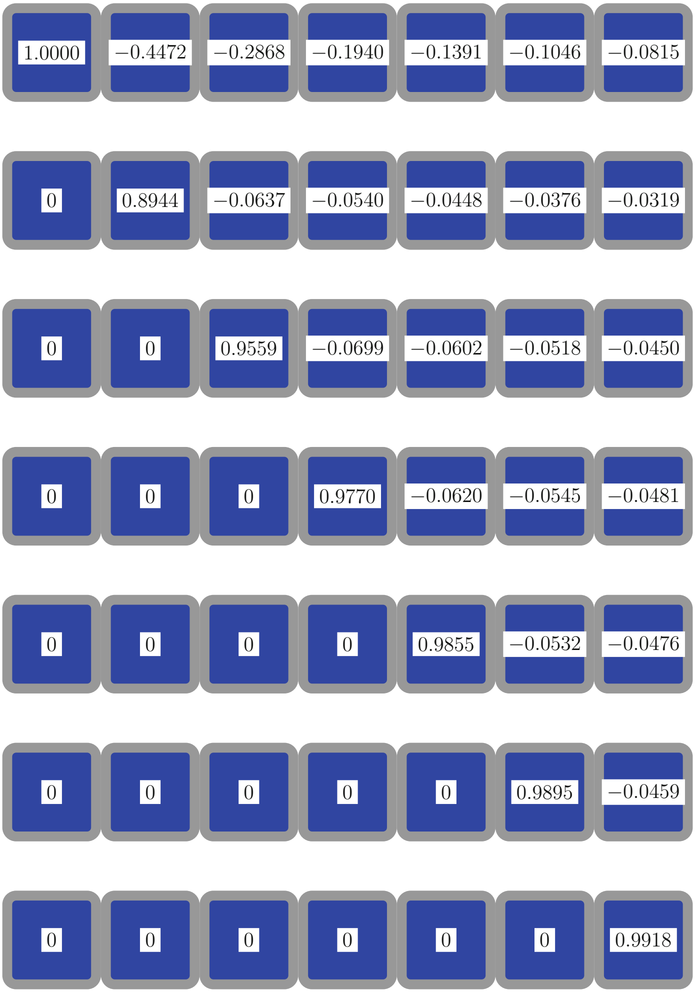


Fig. 1 Presentation of the matrix inverse computations



1.0000	0.5000	0.3333	0.2500	0.2000	0.1667	0.1429
0	1.1180	0.0745	0.0671	0.0596	0.0532	0.0479
0	0	1.0462	0.0749	0.0686	0.0626	0.0573
0	0	0	1.0235	0.0644	0.0598	0.0555
0	0	0	0	1.0147	0.0545	0.0512
0	0	0	0	0	1.0106	0.0468
0	0	0	0	0	0	0.0083

Fig. 2 Numerical values of the algorithm, computation of R



1.0000	−0.4472	−0.2868	−0.1940	−0.1391	−0.1046	−0.0815
0	0.8944	−0.0637	−0.0540	−0.0448	−0.0376	−0.0319
0	0	0.9559	−0.0699	−0.0602	−0.0518	−0.0450
0	0	0	0.9770	−0.0620	−0.0545	−0.0481
0	0	0	0	0.9855	−0.0532	−0.0476
0	0	0	0	0	0.9895	−0.0459
0	0	0	0	0	0	0.9918

Fig. 3 Numerical values of the algorithm, computations of V the two matrices that represent V and R are inverses of each other

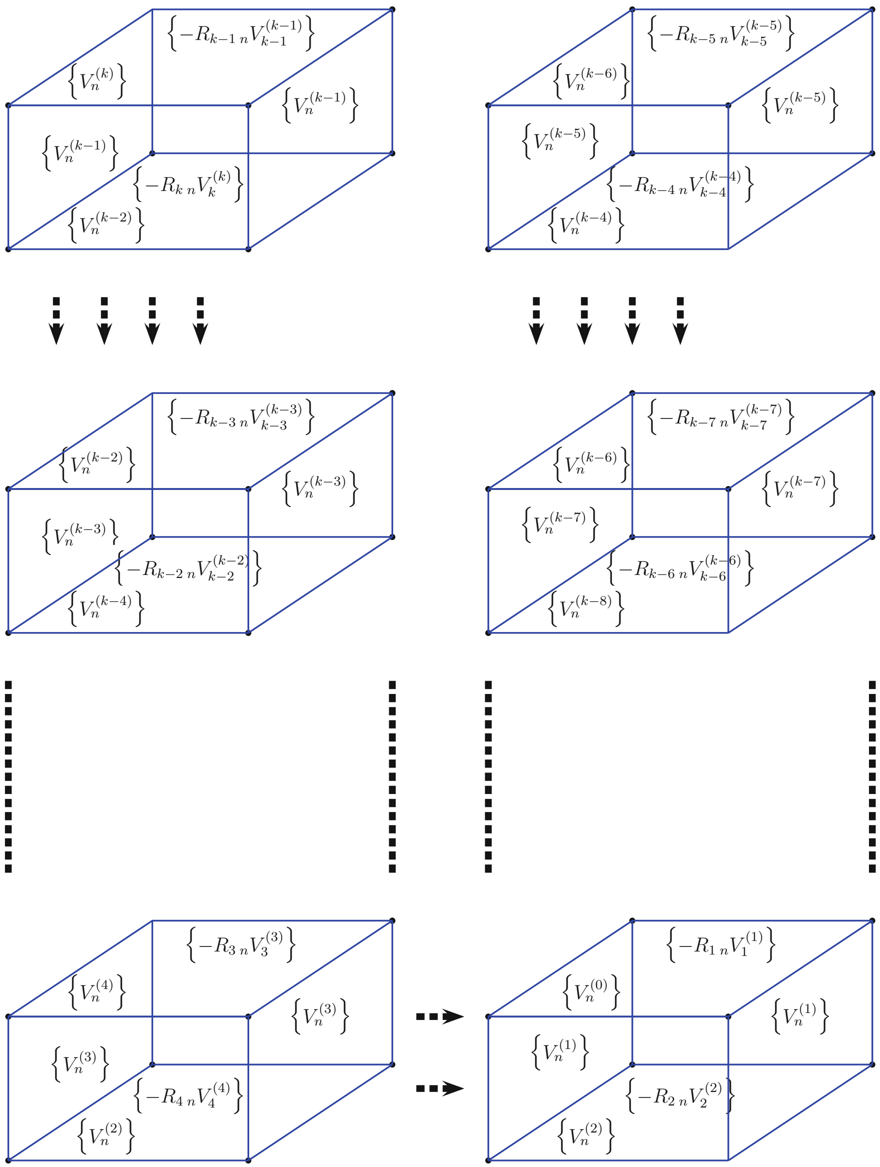

$\left\{-R_{k-1\,n}V_{k-1}^{(k-1)}\right\}$
$\left\{V_n^{(k)}\right\}$
$\left\{V_n^{(k-1)}\right\}$
$\left\{V_n^{(k-1)}\right\}$
$\left\{-R_{k\,n}V_k^{(k)}\right\}$
$\left\{V_n^{(k-2)}\right\}$
$\left\{-R_{k-5\,n}V_{k-5}^{(k-5)}\right\}$
$\left\{V_n^{(k-6)}\right\}$
$\left\{V_n^{(k-5)}\right\}$
$\left\{V_n^{(k-5)}\right\}$
$\left\{-R_{k-4\,n}V_{k-4}^{(k-4)}\right\}$
$\left\{V_n^{(k-4)}\right\}$
$\left\{-R_{k-3\,n}V_{k-3}^{(k-3)}\right\}$
$\left\{V_n^{(k-2)}\right\}$
$\left\{V_n^{(k-3)}\right\}$
$\left\{V_n^{(k-3)}\right\}$
$\left\{-R_{k-2\,n}V_{k-2}^{(k-2)}\right\}$
$\left\{V_n^{(k-4)}\right\}$
$\left\{-R_{k-7\,n}V_{k-7}^{(k-7)}\right\}$
$\left\{V_n^{(k-6)}\right\}$
$\left\{V_n^{(k-7)}\right\}$
$\left\{V_n^{(k-7)}\right\}$
$\left\{-R_{k-6\,n}V_{k-6}^{(k-6)}\right\}$
$\left\{V_n^{(k-8)}\right\}$
$\left\{-R_{3\,n}V_3^{(3)}\right\}$
$\left\{V_n^{(4)}\right\}$
$\left\{V_n^{(3)}\right\}$
$\left\{V_n^{(3)}\right\}$
$\left\{-R_{4\,n}V_4^{(4)}\right\}$
$\left\{V_n^{(2)}\right\}$
$\left\{-R_{1\,n}V_1^{(1)}\right\}$
$\left\{V_n^{(0)}\right\}$
$\left\{V_n^{(1)}\right\}$
$\left\{V_n^{(1)}\right\}$
$\left\{-R_{2\,n}V_2^{(2)}\right\}$
$\left\{V_n^{(2)}\right\}$

Fig. 4 Hardware construction of the inverse matrix computations

$$V = \begin{bmatrix} 1 & -\frac{a_{12}}{||V_2^{(1)}||} & & \cdots & \cdots & \\ 0 & \frac{1}{||V_2^{(1)}||} & & & & \vdots \\ & & \frac{1}{||V_3^{(1)}||\cdot||V_3^{(2)}||} & & & \vdots \\ \vdots & 0 & \vdots & \ddots & & \\ & \vdots & 0 & & \ddots & \vdots \\ 0 & 0 & \cdots & \cdots & 0 & \frac{1}{||V_n^{(1)}||\cdot||V_n^{(2)}||\cdots||V_n^{(n-1)}||} \end{bmatrix}$$

$$R = \begin{bmatrix} 1 & a_{12} & & \cdots & \cdots & a_{1n} \\ 0 & ||V_2^{(1)}|| & & & & V_2^{(2)} \cdot A_n \\ 0 & & ||V_3^{(1)}|| \cdot ||V_3^{(2)}|| & \vdots & \cdots & \vdots \\ \vdots & 0 & \vdots & \ddots & \cdots & \vdots \\ 0 & \vdots & 0 & & \ddots & V_{n-1}^{(n-1)} \cdot A_n \\ 0 & 0 & \cdots & \cdots & 0 & ||V_n^{(1)}|| \cdot \cdots \cdot ||V_n^{(n-1)}|| \end{bmatrix}$$

These two matrices provide the same results as in the Figs. 2 and 3.

6 Hardware Construction Concept

The construction of the hardware for this algorithm will be recursive. The process is summarized in the following steps:

1. $V_1^{(0)}$ and construct $R_{1,2}$
2. $V_1^{(2)}$ and construct $R_{1,3}$ $R_{2,3}$
3. $V_3^{(3)}$ and construct $R_{1,4}$ $R_{2,4}$ $R_{2,4}$ $\cdots$
4. $V_{n-1}^{(n-1)}$ and construct $R_{1,n}$ $R_{2,n}$ $\cdots$ $R_{n-1,n.}$

The proposed hardware will not be trivial for FPGA realisation. Figure 4 represents the computation process of our hardware, see [14] for details on the constructed hardware.

7 Conclusion

This research article points at partial reconfiguration and matrix inverse computation. The mathematical analysis of this paper and the proposed hardware will be a hard task for FPGA realisation. This algorithm is new and is based on the recursive dynamic process. Although the concept of partial reconfiguration will address hardware near systems. The general approach in this research article in this paper is numerical [15–23], in particular the computations presented in section five. The inverse matrix computations from the linear recursive process, its theoretical construction in hardware will have the following impact:

1. Creations of real values hardware instead of integer valued hardware.
2. Solve the generalised inverses matrix problem.
3. Coding matrices.

From the previous resulting research, the inverse computations particular result will serve computer algorithms, particularly numerical matrix based computations. The hardware construction of this algorithm is still open and will have a great research impact on real valued hardware creation. The analyses developed in this paper will be significant for computer scientists and computer engineers.

References

1. B. Osterloh, H. Michalik, S.A. Habinc, B. Fiethe, in *Conference Publication.* Dynamic partial reconfiguration in space applications, pp. 336–343 (2009)
2. A. Donato, F. Ferrandi, M.D. Santambrogio, D. Sciuto, in *IEEE International Soc Conference.* Operating system support for dynamically reconfigurable soc architectures, pp. 233–238 (2005)
3. Upegui, http://lslwww.epfl.ch/upegui/docs/DPR.pdf
4. Xilinx Inc, http://www.xilinx.com/supportdocumentation/white-papers/wp374-partial-reconfig-xilinx-FPGAs.pdf (2012)
5. G.H. Golub, C.F. Van Loan, *Matrix Computations*, 2nd edn. (Johns Hopkins University Press, Baltimore, 1989)
6. Forsythe, E. George, M.A. Malcolm, C.B. Moler, *Computer Methods for Mathematical Computations* (Prentice Hall, Englewood Cliffs, 1977)
7. M.T. Heath, *Scientific Computing: An Introductory Survey* (Mcgraw-hill, Boston, 1997)
8. R.L. Burden, J.D. Faires, *Numerical Analysis* (Brooks/Cole Publishing Company, London, 1997–2001)
9. N.J. Higham, *Accuracy and Stability of Numerical Algorithms* (Siam, Philadelphia, 1996)
10. E.A. Mbock, *Algorithms in Matlab, the Reality of Abstraction and the Power of Pseudocodes* (Optimus Verlag, 2012)
11. J.W. Demmel, *Applied Numerical Linear Algebra* (Siam, Philadelphia, 1997)
12. Y. Meng, in *Proceedings of the 2006 ieee International Conference on Robotics and Automation.* An agent-based mobile robot system using configurable soc technique (Orlando, Florida, 2006) pp. 3368–3373

13. M. Long, A. Gage, R. Murphy, K. Valavanis, in *Proceedings of the IEEE International Conference on Robotics and Automation* Application of the distributed field robot architecture to a simulated demining task (Barcelona, Spain, 2005) pp. 3193–3200
14. E.A. Mbock, *Partial Reconfiguration and the Novel Algorithm for Matrix Inverses Computations* (The World Congress on Engineering and Computer Science, San Francisco, USA, 2013), pp. 118–122
15. G.H. Golub, ChF Van Loan, *Matrix computations*, 2nd edn. (Johns Hopkins University Press, Baltimore, 1989)
16. M.T. Heath, *Scientific Computing: an Introductory Survey* (Mcgraw-hill, Boston, 1997)
17. N.J. Higham, *Accuracy and Stability of Numerical Algorithms* (Siam, Philadelphia, 1996)
18. R.D. Skeel, J.B. Keiper, *Elementary numerical Computing with Mathematica* (Mcgraw-hill, New York, 1993)
19. B.D. Hahn, *Essential Matlab for Scientists and Engineers* (Wiley, New York, 1997)
20. D.R. Hill, D.E. Zitarelli, *Linear Algebra Labs with Matlab*, 2nd edn. (Prentice Hall, Upper Saddle River, 1996)
21. R.E. Larson, B.H. Edwards, Elementary Linear Algebra, *3rd edn.* (D.C. Heath and Company, Lexington, 1996)
22. S.J. Leon, *Linear Algebra with Applications*, 5th edn. (*Prentice Hall, Upper Saddle River,* 1998)
23. G. Strang, *Linear Algebra and its Applications*, 2nd edn. (Academic Press, Orlando, 1980)

Chapter 16
Algorithmic Analysis of the Pseudoanalytic Cryptographic Algorithm

Ariana Guadalupe Bucio Ramirez, Cesar Marco Antonio Robles Gonzalez, Marco Pedro Ramirez Tachiquin and Rogelio Adrian Hernandez Becerril

Abstract In order to protect the information, the study and development of a new cryptographic method is a hard duty, due to the high amount of techniques to decrypt and obtain the information; this works is fully dedicated to analyse the run time of the cipher method employing the Pseudoanalytic Function Theory. The main purpose of this work, is to analyse the cipher method exposed in [3] and calculate the time complexity in order to study the behaviour and look at the possibility to develop an optimized algorithm, preserving the property of confidentiality information, but decreasing the time lapse.

Keywords Algorithm · Cipher · Complexity · Cryptography · Electrical impedance tomography · Mathematics · Pseudoanalytic · Security · Vekua

Every author equally contibuted to the research work.

A. G. Bucio Ramirez (✉)
UPIITA-IPN, Valle del Drave # 31 Colonia Valle de Aragon 3 Seccion,
Ecatepec Estado de Mexico 55280, Mexico
e-mail: ari.bucio@gmail.com

C. M. A. Robles Gonzalez · R. A. Hernandez Becerril
ESIME-IPN, Ciudad de México, Mexico
e-mail: croblesg1101@alumno.ipn.mx

M. P. Ramirez Tachiquin
Postgraduate Section of Mechanical Engineering School,
Instituto Politecnico Nacional Mexico City, Ciudad de México, Mexico
e-mail: mpram@prodigy.net.mx

H. K. Kim et al. (eds.), *Transactions on Engineering Technologies*,
DOI: 10.1007/978-94-017-9115-1_16,

1 Introduction

Nowadays, we are living in a digital era, therefore it is really important to study new techniques for secure communication, thence the importance of cryptography which guarantee the availability of the message only for those who posses the key is important. In order to propose an application based on the Pseudoanalytic Function Theory [2], which it proves be a significant tool in Applied Mathematics (e.g. [4, 6, 8]) and Theoretical Physics.

The Pseudoanalytic Functions have been used to study The Electrical Impedance Tomography Problem [10], this is a mathematical problem that remains open, posed by A. P. Calderon in 1980 [5], who showed that the solution of the problem exists, is unique and steady. By the employ of this characteristics, it was possible to propose a new cryptographic method presented in [3], an important features does the algorithm presents is the *Confidentiality* and the *level of security*.

The focus of this work, will be the performance of the proposal cryptographic method showed in [3], by means of the analysis with the influence of the parameters through the time complexity, even though this operations are limited by the hardware characteristics.

2 Preliminaries

As it has been shown in [3], the idea of presenting a cryptographic algorithm emerges from the study of the Electrical Impedance Tomography problem, when applying mathematical tools based on the Pseudoanalytic Function Theory [6], even though it is important to remark that these numerical techniques does not provide yet a suitable solution for this problem (see e.g. [4] and [7]).

Perhaps, the main advantage of this method is that any attempt to break the ciphering method, would be equivalent to fully solve an arbitrary case for the Electrical Impedance Tomography problem, which still remains open, for more details about this method see [3]. As it has been shown in several works (see e.g. [1] and [6]), the two-dimensional case of the Electrical Impedance Equation is fully equivalent to a Vekua equation [9]:

$$\partial_{\bar{z}} W - \frac{\partial_{\bar{z}} p}{p} \bar{W} = 0, \tag{1}$$

where

$$W = \sqrt{\sigma}(\partial_x u - i\partial_y u), \ \partial_{\bar{z}} = \partial_x + i\partial_y, \ \text{and}\ p = \sqrt{\sigma_1(x)^{-1}\sigma_2(y)}.$$

Professor L. Bers [2] proved that the general solution of the equation [1] can be expressed in terms of the Taylor series in formal powers:

$$W = \sum_{n=0}^{\infty} Z^{(n)}(a_n, z_0; z), \tag{2}$$

where a_n and z_0 are complex constants, $z = x + iy$ and $i^2 = -1$. A more detailed description of the analytic construction of $Z^{(n)}(a_n, z_0; z)$ can be found in [2, 6, 8].

For constructing of the so-called Taylor series in formal powers, hereinafter called formal powers, for the integral expressions showed in [3], let us consider a collection of $\mathbf{Q} = Q + 1$ points located in the closed interval [0, 1] and a radius **R**. The elements of the array $Z^{(0)}[q]$, corresponding to the numerical approach of the formal power $Z^{(0)}(1, 0; z)$, as it shows:

$$Z^{(0)}[q] = \sqrt{\sigma(x[q], y[q])}. \tag{3}$$

The succeeding elements q of the arrays $Z^{(n)}[q]$ will be approached according to the following variation of the classical trapezoidal integration method [3]:

$$\begin{gathered} Z^{(n)}[q] = F[q]\cdot \\ \cdot Re \sum_{h=0}^{q} \left(G^*[h]Z^{(n-1)}[h] + G^*[h+1]Z^{(n-1)}[h+1]\right)\cdot \\ \cdot(x[h+1] - x[h] + i(y[h+1] - y[h])) + G[q]\cdot \\ \cdot Re \sum_{h=0}^{q} \left(F^*[h]Z^{(n-1)}[h] + F^*[h+1]Z^{(n-1)}[h+1]\right)\cdot \\ \cdot(x[h+1] - x[h] + i(y[h+1] - y[h])). \end{gathered} \tag{4}$$

This integration process will be performed for every angle θ, in which the close intervals $[0, 2\pi)$ are subdivided, considering an array of **S** angles where, we are going to approach N formal powers, obtaining a system $\mathbf{N} = 2N + 1$ vectors each one with **S**. After the construction of the formal powers, the application of standard Gram-Schmidt orthonormalization process to the set, will take place:

$$\left\{Z^{(n)}[\mathbf{K}; s]\right\}_{n=0,s=0}^{\mathbf{N},\mathbf{S}} \tag{5}$$

from which will derive the matrix $\mathbf{U}_{[\mathbf{N},\mathbf{S}]}$. For more details we want to refer the reader to [3].

3 Cipher and Decipher Method

The cipher method posed in [3], which employs a secret key, common to the sender and the recipient. For the construction of the key, this method employ a pseudo-random matrix $\mathbf{A}_{[\mathbf{Q},\mathbf{S}]}$, **S** maximum number of radii, **Q** maximum number of points and N maximum number of formal powers.

By the employ of the last parameters, the encryption process proceeds by the approximation of N Formal Powers [3] constructed as it shows in (4), then as it mention before, a standard Gram-Schmidt orthonormalization process takes place to obtain the matrix $\mathbf{U}_{[\mathbf{N},\mathbf{S}]}$. So, the data to be ciphers will be in the matrix $\mathbf{B}_{[\mathbf{M},\mathbf{N}]}$, and the cipher process continue until the matrix $\mathbf{C}_{[\mathbf{M},\mathbf{S}]}$ has been completed, as it shows:

$$C^m = \sum_{n=0}^{\mathbf{N}} b_{m,n} U^n. \tag{6}$$

The most important difference between the cipher and decipher process, is illustrated in Fig. 1 and for decrypting the message, we will calculate the inner products between the vectors C^m that belong $\mathbf{C}_{[\mathbf{M},\mathbf{N}]}$, and U^n belonging to $\mathbf{U}_{[\mathbf{N},\mathbf{S}]}$:

$$b_{m,n} = \langle C^m, U^n \rangle. \tag{7}$$

4 About the algorithm analysis

In the Computer Science, the analysis of algorithms is greatly important to determine the necessary resources for program execution; mostly, the running time (efficiency) of an algorithm which is stated as function of the input length for the number of steps (time complexity). A theoretical classification which estimates the behaviour of an algorithm by the increases or decreases in the parameters, also know as "*run-time analysis*", for the proposed cypher method the parameters to change, will be N formal powers and S points per radii, obtaining the running time expressed in seconds. This analysis is useful for the verification of the algorithmic efficiency posed in [3].

Table 1, shows the behaviour when the employed parameters change, showing the best case for $N = 5$ and $S = 101$ and the worst case for $N = 55$ and $S = 101$, this behaviour also be appreciated in Table 2, and full result could be illustrated in Fig. 2. In light of these results, we can conclude that the behaviour of the algorithm will change, when increases the number of N formal powers, such like it does when S number of points grows, thus, this cypher method posses a polynomial complexity, based on the time analysis of the algorithm complexity. With the performing of different tests, the numerical precision in the algorithm is not directly related by the influence in these parameters.

Since, the security of the information is the main purpose of cryptography, we ought to remark, that preserving confidentiality is the most important characteristic of this algorithm, because trying to decipher the data $b_{m,n}$, employing any different method to (7), would be equivalent to solve the Electrical Impedance Tomography problem posed in [5], as we mention before, this problem still remains open. Other

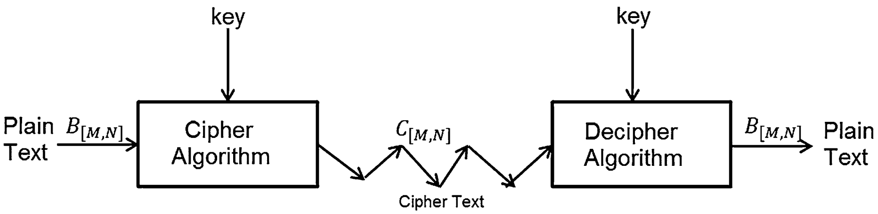


Fig. 1 Cryptographic algorithm employing Pseudoanalytic Function Theory

Table 1 Run time of the cipher method employing 101 points per radii

Formal powers N	Number of points Q	Points per radii S	Time (s)
5	100	101	0.497
15	100	101	1.265
25	100	101	2.040
35	100	101	2.823
45	100	101	3.609
55	100	101	4.390

Table 2 Run time of the cipher method employing 401 points per radii

Formal powers N	Number of points Q	Points per radii S	Time (s)
5	100	401	1.931
15	100	401	4.981
25	100	401	8.075
35	100	401	11.134
45	100	401	14.499
55	100	401	17.372

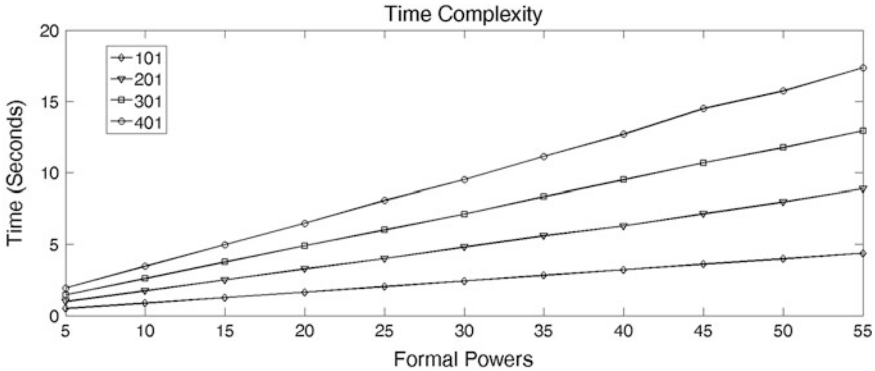


Fig. 2 Running time while employing $Q = 100$ also non fixed N and S

important target of cryptography is the level of security for this algorithm this is warranted because if any parameter of the secret key is changed you will need to construct all the numerical calculations again. As a future work, the optimization of the algorithm will be done, in order to employ it on real time applications. In this optimization process we are thinking about the possibility of applying an authentication encryption process as keyed-hash message authentication code, which simultaneously will provide to the algorithm confidentiality, data integrity and authenticity for security on the message.

References

1. V.V. Kravchenko, *Applied Pseudoanalytic Function Theory, Series*: *Frontiers in Mathematics*, ISBN: 978-3-0346-0003-3 (2009)
2. L. Bers, *Theory of Pseudoanalytic Functions* (IMM New York University, New York, 1953)
3. A. Bucio Ramirez, R. Castillo-Perez, M.P. Ramirez Tachiquin, *On the Numerical Construction of Formal Powers and their Application to the Electrical Impedance Equation*, Proceedings of 8th International Conference on Electrical Engineering, Computing Science and Automatic Control, IEEE Catalog Number: CFP11827-ART, ISBN:978-1-4577-1013-1, pp. 769–774 (2011)
4. M.P. Ramirez Tachiquin, M.C. Robles Gonzalez, R.A. Hernandez-Becerril, A. Bucio Ramirez, *First characterization of a new method for numerically solving the Dirichlet problem of the two-dimensional electrical impedance equation*, J. Appl. Math. **2013**(493483), p. 14 (2013)
5. J.G. Webster, *Electrical Impedance Tomography* (Adam Hilger Series on Biomedical Engineering, New York, 1990)
6. A.P. Calderon, *On an Inverse Boundary Value Problem*, Seminar on Numerical Analysis and its Applications to Continuum Physics, Soc. Brasil. Mat., pp. 65–73 (1980)
7. A. Bucio Ramirez, R. A. Hernandez-Becerril, C.M.A. Robles Gonzalez, M.P. Ramirez Tachiquin, A. Arista-Jalife, *Construction of a New Cryptographic Method, Employing Pseudoanalytic Function Theory,* Proceedings of the World Congress on Engineering and Computer Science 2013, vol 1. WCECS 2013, ISBN: 978-988-19252-3-7, pp. 96-101 (2013)
8. C.M.A. Robles Gonzalez, A. Bucio Ramirez, M.P. Ramirez Tachiquin, *New characterization of an improved numerical method for solving the electrical impedance equation in the plane: an approach from the modern pseudoanalytic function theory*. IAENG Int. J. Appl. Math., **43**(1), IJAM_43_1_03 (2013)
9. K. Astala, L. Päivärinta, *Calderon's inverse conductivity problem in the plane*. Ann. Math. **163**, 265–299 (2006)
10. I.N. Vekua, *Generalized Analytic Functions, International Series of Monographs on Pure and Applied Mathematics* (Pergamon Press, London, 1962)

Chapter 17
Toward Modeling Emotion Elicitation Processes Using Fuzzy Appraisal Approach

Ahmad Soleimani and Ziad Kobti

Abstract This paper investigates using a fuzzy appraisal approach to model the dynamics for the emotion generation process of individuals. The proposed computational model uses guidelines from OCC emotion theory to formulate a system of fuzzy inferential rules that is capable of predicting the elicitation of different emotions as well as transitioning between different emotional states as a result of an occurred event, an action of self or other individuals, or a reaction to an emotion triggering object. In the proposed model, several appraisal variables such as event's desirability and expectedness, action's praise-worthiness and object's degree of emotional appealing were considered and thoroughly analyzed using different techniques. The output of the system is the set of anticipated elicited emotions along with their intensities. Results showed that the proposed computational model is an effective and easy to implement framework that poses an acceptable approximation for the naturally sophisticated dynamics for elicitation and variation of emotional constructs in humans.

Keywords Computational models of emotion · Emotion elicitation · Emotional intelligence · Emotion regulation · Fuzzy automata · OCC theory

1 Introduction

Emotions play an important role in shaping our desires and tendencies. According to contemporary research findings (e.g., [1, 3, 5, 9]), emotions deeply influence the process of decision making and other cognitive tasks of humans. Furthermore, they

A. Soleimani (✉) · Z. Kobti (✉)
Department of Computer Science, University of Windsor, 401 Sunset Ave,
Windsor, ON, Canada
e-mail: soleima@uwindsor.ca

Z. Kobti
e-mail: kobti@uwindsor.ca

H. K. Kim et al. (eds.), *Transactions on Engineering Technologies*,
DOI: 10.1007/978-94-017-9115-1_17,

contribute in the development of coping system that help us to adapt our thoughts and behaviors to the changes that take place in the environment [11].

Considering the deep mutual impact between affect and cognition, studying emotions has attracted a great deal of research works across a large spectrum of disciplines starting from humanistic sciences such as psychology and cognitive science to applied sciences and engineering and arriving at public well being and healthcare. A great deal of affect-enabled applications and commercial products started to emerge in the market as a result of the recent “affect-awareness” research campaign [18].

Within the field of information technology and computer science, *Affective Computing* can be considered the fruitful outcome of studying and modeling emotions by IT researchers. Despite the relatively young age of this field, it has managed to turn into a well-established research area with its own professional meetings and scholarly journals. According to its founder, Picard [13], AC is “computing that relates to, arises from, or deliberately influences emotions”. AC is aimed at filling the gap between highly emotional people and emotional challenged machines [2]. An AC system is involved in building computer artifacts that are more emotionally intelligent, i.e., to recognize (e.g., from person’s facial expressions or physiological signals measured by wearable sensors), represent (e.g., by building computational models) and respond to (e.g., in service robots or avatars) affective states.

2 Computational Models of Emotions

In the process of building computational models of emotion, different approaches such as appraisal (e.g., [8, 12]), dimensional (e.g., [7, 15]), adaptation (e.g., [11]) can be used. Appraisal theory, non-arguably is the most widely used approach in the recent computational models of emotion [20].

2.1 Appraisal Theory

Based on this theory, emotions are outcomes of previously evaluated situations attended by the subject individual and the connection between emotions and cognition is highly emphasized. Accordingly, emotional responses are generated based on an appraisal or assessment process performed continuously by the individual on situations and events that take place in the environment and are perceived relevant by the individual.

According to the appraisal theory which was formally proposed by Smith and Lazarus [17], in order to evaluate the different situations that arise in the relationship between an individual and its environment, a set of appraisal variables or dimensions needs to be considered. Scherer [16] and Frijda [6] argue that these appraisal variables should be able to address the affective-relevant aspects of the situation, such as relevance of the situation, degree of situation expectancy, coping potentials, etc.

2.2 OCC Theory

The emotion process model suggested by Ortony, Clore and Collins known as OCC [12] is a robust and well-established appraisal theory for emotion dynamics that was highly influential in the field of studying emotions within IT. Consequently, a considerable number of computational models of emotions can be seen nowadays where OCC was the basis for them (e.g., [4, 7, 11]).

The popularity of OCC among computer scientists can be attributed to the fact that this theory was founded on a well-defined constraint-satisfaction architecture approach with a finite set of appraisal dimensions used as criteria for classifying different emotions. Such an approach taken in OCC makes it computationally tractable and hence, understandable by computer specialists.

At the heart of the OCC theory, the elicitation dynamics of 22 emotions divided in three categories were studied thoroughly. These three categories are corresponding to the elicitation causes of each class of emotions. Accordingly, the first class includes those emotions that are elicited as a result of the occurrence of relevant external events (see Fig. 1). The second and the third classes are related to reactions to self or others actions as well as being exposed to emotion triggering objects respectively (see Fig. 2).

3 Proposed Computational Model

The proposed model is an OCC based model at which a fuzzy approach was taken to perform the required appraisal processes. Accordingly, by using guidelines from the OCC theory, the elicitation dynamics of all classes of emotions were modeled.

With respect to event-originated emotions, according to Fig. 1, the first appraisal variable that divides the emotions of this class into two sets is the orientation of the event; determining the utility of the event to be either directed toward the agent itself or some other agent(s). This evaluation process yields to a first level of classification of the emotions into *for self* or *for others* categories. Another classification takes place for *self* emotions group based on the prospective aspect of the event which indicates if the event has already taken place (prospect = False) or would possibly take place in the future (prospect = True). A prospective emotion, e.g., *hope* transforms into a post-prospect emotion of *satisfaction* in case of confirmation or *disappointment* in case of disapproval according to some temporal dynamics explained in Sect. 4.

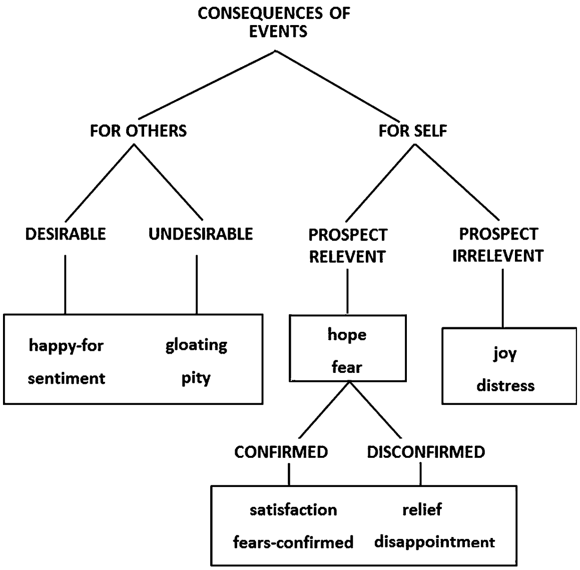


Fig. 1 OCC event-originated emotions. Adopted partially from [12]

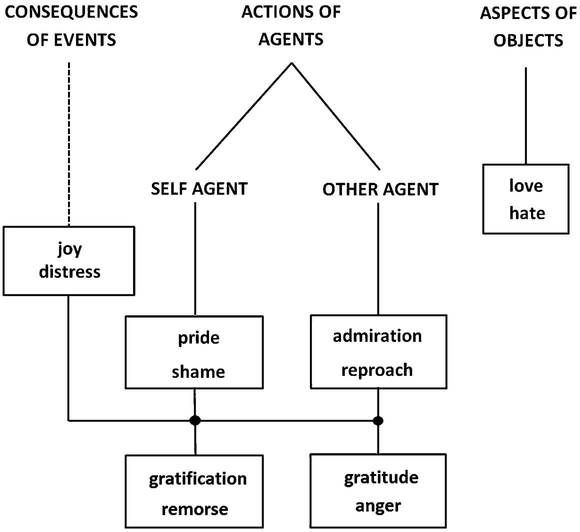


Fig. 2 OCC non-events-originated emotions. Adopted partially from [12]

3.1 Events

Event-originated set of OCC emotions contains emotions whose eliciting conditions are directly linked to an appraisal process performed on external events that take place in the environment of the agent.

3.1.1 Event's Desirability

In OCC theory, desirability is close in meaning to the notion of utility. When an event occurs it can satisfy or interfere with agent's goals, and the desirability variable has therefore two aspects; one is corresponding only to the degree to which the event in question appears to have beneficial (i.e. positively desirable) consequences; and the other is corresponding to the degree to which it is perceived as having harmful (i.e. undesirable) consequences.

The desirability of an occurred or prospective event poses the most influential factor in the specification of the emotion type that will be triggered in the agent along with its intensity. A fuzzy approach was considered to determine the desirability level of events. Accordingly, a fuzzy scale for desirability consisting five fuzzy sets was considered as follows:

$Desirability = \{HighlyUndesired,\ SightlyUndesired,\ Neutral,\ SlightlyDesired,\ Highlydesired\}.$

The above desirability level is linked to an evaluation process that takes into account the impact of the event on the set of goals of the agent. Two other fuzzy variables, *Impact* that indicates the event's degree of influence on the goals of the agent; and *importance* that reflects the importance or preference of each goal were considered. Hence,

$Impact = \{HighlyNegative,\ SlightlyNegative,\ NoImpact,\ SlightlyPositive,\ HighlyPositive\}$
$Importance = \{ExtremlyImportant,\ SlightlyImportant,\ NotImportant\}$

Considering the fact that an event can have an impact on multiple goals whereas each goal has its own importance level, the problem of measuring the desirability of an event would turn into solving a system of fuzzy rules [4]. With regards to the composition of the fuzzy rules in the resulted fuzzy system, a combination of the *sup_min* composition technique proposed by Mamdani [10] and the weighted average method for defuzzification [14] was considered. Using the composition approach explained in [4], we can apply the *sup_min* operator on *Impact*, *Importance* and *Desirability,* and hence, the matching degree between the input and the antecedent of each fuzzy rule can be determined. The final output of this fuzzy system will be the fuzzy value of desirability. We use the following formula based on the weighted average method for defuzzifing the above combined fuzzy result:

$$Desirability_f(e) = y_{final} \ = \frac{\sum \mu_{comb}(\bar{y}).\bar{y}}{\sum \mu_{comb}(\bar{y})}$$

where $\bar{y}$ is the mean of each symmetric membership function.

3.1.2 Events Prospect

As discussed earlier in this article, a group of OCC emotions are prospective emotions, meaning that they are some transient emotional states that reflect a kind of uncertainty with respect to the occurrence possibility of some events. Hence, these emotional states eventually turn into stable emotions once the uncertainty factor was removed. The prospective attribute is directly related to the degree of occurrence possibility set by the agent. In other words it reflects a mechanism for event expectedness by the agent.

In the proposed model, a simple but acceptable estimation for this measure, similar to the one used in [4] was adopted. Based on this approach, a learning module is used to enable the agent to learn patterns for the events and consequently to expect the occurrence of future events using a probabilistic approach. The event's patterns are constructed based on the frequency with which an event, say, e_1 is observed to occur right before previous events of e_2, e_3, etc.

A table data structure is used to count the number of iterations for each event pattern. Conditional probability of $p(e_3|e_1, e_2)$ indicates the probability for event e_3 to happen, assuming that events e_1 and e_2 have just taken place. Hence,

$$Likelihood(e_3|e_1, e_2) = \frac{C[e_1, e_2, e_3]}{\sum_i C[e_1, e_2, i]}$$

where c denotes the count of each event pattern. Here, a length of three for the sequence of the event patterns was considered.

For the complete list of steps and equations used in the last two sections, interested readers are referred to the corresponding conference paper [18].

3.2 *Actions*

Another class of emotions in OCC are those originated by the consequences of purposeful actions. Some events can be attributed to the actions of self or some other agent(s). According to this approach, a measure for the *praiseworthiness* attribute of the action needs to be defined. The valence of this attribute will be positive if the action is in-line with the contextual standards or values, e.g., saving a drowning person which will elicit pride or admiration emotions; whereas it will get a negative value if the action violates those standards or values, e.g., mocking a handicapped person which will trigger an emotion of shame or reproach (in this case it can be called the degree of blameworthiness). It is presumed though that these standards are adopted by the agent itself and are active in the evaluation process of the actions. It is important to be clarified that the proposed model keeps itself independent from these standards and for the sake of providing higher generality for the model, it is assumed that they are simply given to the system.

3.3 Compound Emotions

According to OCC model, some emotions are compound emotional states due to the fact that they are related to the consequences of regular events as well as actions-originated events. A compound emotion such as *anger* is triggered when the evaluating agent appraises both the desirability of the event and the attribution of the action led to the event. Hence, a state of anger is interpreted as a combination of *distress* and *reproach* emotions. Therefore, for this type of emotions, the appraisal parameters would include praiseworthiness of the performed action as well as the desirability of the occurred event.

3.4 Objects

Love and *hate* are under this class of OCC emotions. Appraisal variables that affect the intensity of these two emotions are the degree of appealingness and degree of familiarity with the objects. The appealingness can be considered as a function of dispositional attitude towards a category that the object belongs to. Accordingly, the value *attractive* is set if the object has a positive object valence with a familiarity valence less than a certain threshold. Conversely *not attractive* is set if the object has a negative object valence with a familiarity valence above a certain threshold.

4 Problem Formulation

As discussed earlier, emotions in OCC model are divided into three major groups. We strive to keep the formulation of this problem and the calculative modules inline with the original classification of emotions.

4.1 Emotion Calculations

In this section, a set of computational equations is proposed for each emotion in order to anticipate the elicitation of the under study emotion as well as its intensity level. These modules were designed based on the approach presented in the previous section along with some guidelines from the OCC emotion theory. In these formulas, e is an occurred event, subscript $_p$ stands for *potential* and subscript $_t$ stands for *threshold*, p_i reflects an agent and t is an indicator for time, a is an action performed by self or some other agent, and obj is an encountered object.

It is assumed that an emotional state will not be triggered unless its intensity is above a certain threshold level. This assumption was applied in accordance with the real world rule that not any desirable or undesirable feeling would yield into an explicit emotion [12]. Furthermore, according to the formalization of emotions proposed by Steunebrink et al. [19], it is necessary to differentiate between the actual experiences of emotions and those conditions that merely trigger emotions. Hence, a triggered emotion will not necessarily lead to a genuine experience of it, due to the fact that it was assigned an intensity below the minimum experience level.

For brevity, only one emotion from each group is addressed in this paper.

4.1.1 Event-Originated Emotions

As elaborated before, according to the OCC model, event-originated emotions are classified into two groups of *self-related* and *others-related*. This classification was made by considering the consequences of an occurred event to be directed toward either the evaluating agent itself or some other agent. The diagram of Fig. 1 shows that the first group includes the set of {*joy, distress, hope, fear, satisfaction, disappointment, fearsconfirmed, relief*} emotions whereas the second group includes{*happyfor, resentment, gloating, pity*} emotions.

Self-related

In this section, calculation modules for the self-related set of event-originated emotions are presented. Self-related addresses those emotional states that are being elicited in the evaluating agent itself.

Emotion Joy

An agent experiences joy emotion when it is pleased about a desirable event. Hence,

$$\begin{array}{l} IF\ Desirability(p,e,t) > 0 \\ THEN\ JOY_p(p,e,t) = Desirability(p,e,t) \\ IF\ JOY_p(p,e,t) > JOY_t(p,t) \\ THEN\ Intensity(p,e,t) = JOY_p(p,e,t,) - JOY_t(p,t) \\ ELSE\ Intensity(p,e,t) = 0 \end{array}$$

At this point we consider prospected emotion. *Prospect* in the following equations is a binary logical variable that reflects the occurrence prospect for a future event *e*. In case of *Prospect*(*p*, *e*) = *TRUE*, the function of *Likelihood*(*p*, *e*) will return the probability for the occurrence of event e. As an instance of this group of emotions, we consider the equations set for emotion relief as follows:

Emotion Relief

An agent experiences relief emotion when the occurrence of an expected undesirable event is dis-confirmed. Hence,

```
$IF\ FEAR_p(p,e,t) > 0\ AND\ NOT(Occurred(p,e,t_2))\ AND\ t_2 \geq t$
$THEN\ RELIEF_p(p,e,t_2) = FEAR_P(p,e,t))$
$IF\ RELIEF_p(p,e,t_2) > RELIEF(p,t_2)$
$THEN\ Intensity(p,e,t_2) = RELIEF_p(p,e,t_2) - RELIEF_t(p,t_2)$
$reset\ FEAR_P(p,e,t_2) = Desirability(p,e,t_2) * Likeihood(p,e,t_2)$
$ELSE\ Intensity(p,e,t_2) = 0$
```

In the above rules it is simply assumed that once a prospective negative event was disproved, the relief level of the agent would be directly proportional to the level of fear that was experienced by the agent in an earlier time. In addition, although the agent has experienced some relief emotion at time t_2 as a result of disconfirmed negative event e, but we would need to consider the possibility of its occurrence in a later time. This was the reason for recomputing the value of $Fear_p$ since at least one of its parameters (i.e., Likelihood) was changed.

Others-Related

In this section, calculation modules for the *others-related* set of event-originated emotions are presented. Others-related addresses those emotional states that are being elicited in a different agent from the evaluating one.

Emotion HappyFor

An agent experiences *happyfor* emotion if it is pleased about an event presumed to be desirable for a friend agent. Hence,

```
$IF\ Desirability(p_2,e,t) > 0\ AND\ Friend(p_1,p_2)$
$THEN\ IF\ Desirability(p_1,e,t) > 0$
$THEN\ HAPPYFOR_p(p_1,e,t) =(Desirability(p_2,e,t) + Desirability(p_1,e,t))/2$
$ELSE\ THEN\ HAPPYFOR_p(p_1,e,t) =|Desirability(p_2,e,t) - Desirability(p_1,e,t)|$
$IF\ HAPPYFOR_p(p_1,e,t) > HAPPYFOR_t(p_1,t)$
$THEN\ Intensity(p_1,e,t) = HAPPYFOR_p(p_1,e,t,) - HAPPYFOR_t(p_1,t)$
$ELSE\ Intensity(p_1,e,t) = 0$
```

4.1.2 Action-Originated Emotions

Non-compound Emotions

For this set of emotions, we consider a function called *Praise* that evaluates and sets the degree of praiseworthiness of an action. A negative value for this function indicates the degree of blameworthiness of the action.

Emotion Shame

An agent experiences *shame* emotion if it is disapproving its own blameworthy action. Hence,

$IF\, Praise(p_1, p_2a,t) < 0\, AND\, (p_1 = p_2)$
$THEN\, SHAME_p(p_1, p_2,a,t) = -Praise(p_1, p_2a,t)$
$IF\, SHAME_p(p_1, p_2,a,t) > SHAME_t(p_1, p_2,a,t)$
$THEN\, Intensity(p_1, p_2,a,t) = SHAME_p(p_1, p_2,a,t) - SHAME_t(p_1, p_2,a,t)$
$ELSE\, Intensity(p_1, p_2,a,t) = 0$

Compound Emotions

For this class of emotions, as stated earlier, we deal with two other implicit emotional states that are involved in the calculations and the intensity level would include an average-like operation between these two emotions. Therefore, beside the value of function *Praise* used in the above equations, it will be necessary to calculate the desirability of the resulted events in the same way that was performed for the set of event-originated emotions.

Emotion Anger

An agent experiences *anger* emotion if it is disapproving a blameworthy action of another agent that led to an undesirable event. Hence,

$IF\, Praise(p_1, p_2a,t) < 0\, AND\, NOT\, (p_1 = p_2)\, AND\, Desirability(p,e,t) < 0$
$THEN\, ANGER_p(p_1, p_2,a,t) = (REPROACH + DISTRESS_p)/2$
$IF\, ANGER_p(p_1, p_2,a,t) > ANGER_t(p_1, p_2,a,t)$
$THEN\, Intensity(p_1, p_2,a,t) = ANGER_p(p_1, p_2,a,t) - ANGER_t(p_1, p_2,a,t)$
$ELSE\, Intensity(p_1, p_2,a,t) = 0$

4.1.3 Object-Originated Emotions

Emotions love and hate belong to this class of OCC emotions. As an example, the calculation rules for emotion love is provided below:

Emotion Love

An agent experiences love emotion if he is attracted to an appealing and object (agent). Hence, we have

$IF\,Appealing(p,obj,t) > 0$
$THEN\,LOVE_p(p,obj,t) = Appealing(p,obj,t)$
$LOVE_t = k/Familiar(p,obj,t),\ k = constant$
$IF\,LOVE_p(p,obj,t) > LOVE_t(p,obj,t)$
$THEN\,Intensity(p,obj,t) = LOVE_p(p,obj,t) - LOVE_t(p,obj,t)$
$ELSE\,Intensity(p,obj,t) = 0$

4.1.4 Algorithms

Several algorithms were used in the different appraisal processes of the proposed model. Here, the pseudo code for one of these algorithms is provided.

Event-Track-State to determine triggered emotions along with their intensities as a result of the occurrence of a series of events

```
Input: $q_0 =< m_0, I_0 >, Mood_{global}, E = \{e_1, e_2, ..., e_k\}, E\,is\,list\,of\,occurring\,events$
$Q = \{< m_i, I_i >, m_i \in Event\_Competent\_Emotions, I_i \in Intensity_{fuzzy}\}$
Output: $q_f = \{< m_1, I_1 >, < m_2, I_2 >, ... < m_k, I_k >\} \subset Q$
Begin
  Defuzzify state $q_i = q_0$ using weighted average method
  For each event $e \in E$
    Begin
      Calculate $Desirability_f$ for event $e$
      Based on the variables of $Orientation$, $Prospect$ do:
      Determine possible emotional state $< m_i, I_i >$from emotion derivation rules
      Obtain $\Delta MoodR_{global}$ for $e$ using PAD look-up table
     Update $\Delta MoodR_{global}$
    End For;
  For each $m_i$ where $I_i > 0$
    Begin
      Print $< m_i, I_i >$
    End For;
End.
```

4.2 Emotion Regulation

In this section, a possible utilization of emotion regulation [8] as a mean for affect interventions and control is considered. Here, the possibility of deliberately influencing hyper emotional states through applying one of the three previously discussed stimuli was investigated. The ultimate goal for such an approach is to control negative emotions and probably transition them into some contrary positive emotions. We use a simple fuzzy automata for this purpose at which states reflect emotions and the edges are the appraisal vectors for each stimulus. Accordingly, a fuzzy transition function between two given states is a transformation of the fuzzy membership function of the first state into the fuzzy membership function of the second state; hence, it can be represented using relational matrices. A detailed description for this approach is provided in one of the experiments in the next section.

5 Simulation Experiments and Discussion

In order to test the performance of the model and verify its functionality under different circumstances, a series of simulation experiments were conducted. For brevity, only two of these experiments are considered here. The goal of the first experiment is study the emotional behavior of the agent as a result of the occurrence of some independent events. In the second experiment, the potential impact of applying a set of stimuli for regulating the hyper emotional state of an agent is studied. In these experiment, $G = \{G_1, G_2, G_3\}$ are the goals of the agent and $E = \{e_1, e_2, e_3, e_4, e_5\}$ is the set of possible events. The fuzzy values of *Importance* and *Impact* for these goals and events are described in Table 1. Table 2 shows the temporal dynamics of both real and prospect events that take place in the system during the simulation time. It is assumed that the time duration for a prospect event is 20 time-steps; meaning that the agent will experience the competent prospect emotion for 20 time-steps before it turns into a deterministic emotion. In addition, it is assumed that the life-time for each deterministic emotion is 20 time-steps as well; meaning that an emotional response starts to deteriorate through a linear function due to normal decay and vanishes completely after that period.

As the first step, the desirability level for all events of E were calculated and the results are reflected in Fig. 3.

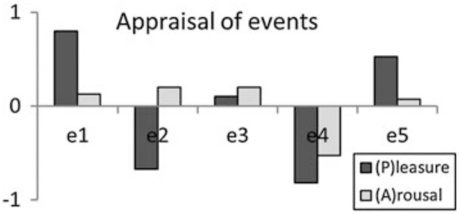
Table 1 List of agent's goals and events along with their impact on each goal for both agents

Goal	G1	G2	G3
Importance	HighlyImportant	SlightlyImportant	HighlyImportant
Event	Impact(G1)	Impact(G2)	Impact(G3)
e_1	HighlyPositive	NoImpact	HighlyPositive
e_2	HighlyNegative	SlightlyPositive	SlightlyNegative
e_3	NoImpact	SlightlyPositive	NoImpact
e_4	HighlyNegative	HighlyPositive	HighlyNegative
e_5	HighlyPositive	HighlyPositive	NoImpact

Table 2 Temporal dynamics of the occurring events

Time	0	10	20	30	40	50	60	70	80	90
Occurrence			e_1			e_4	e_2		e_5	e_1
Prospect		e_2	e_5		e_4		e_5			

Fig. 3 Calculated event's desirability for both agents



5.1 Scenario 1 Unattributed Events

According to Table 2, at time-step = 10, since there is a possibility for the occurrence of e_2 as a negative event, the agent experiences fear emotion. The (actual) occurrence of positive event e_1 at step = 20, caused emotion joy to be triggered in the agent. In addition, at the same step, a certain level of emotion hope was elicited in the agent for the prospect positive event of e_5. At step = 30, due to dis-confirmed e_2, the fear emotion will disappear and gives its room to emotion relief. At step = 40, prospect event e_4 will cause the agent to experience a relatively high level of fear emotion which converts into *fearsconfirmed* at step = 50. At step = 60, negative event e_2 took place and caused the agent to experience a high level of distress emotion. Unlike the earlier prospect occurrence of this event, it was not proceeded by a fear emotion since it was not predicted by the agent. At the same step, the prospect event of e_5 resulted in some degree of hope emotion. This emotion was converted into satisfaction at step = 80 when the occurrence of e_5 was confirmed. Finally, at step = 90, positive event $e_1$took place and caused the agent to experience a high level of joy. Figure 4 shows the big picture of all emotions that were experienced by the agent during the simulation time along with the intensity of each. For instance, it can be seen that the agent

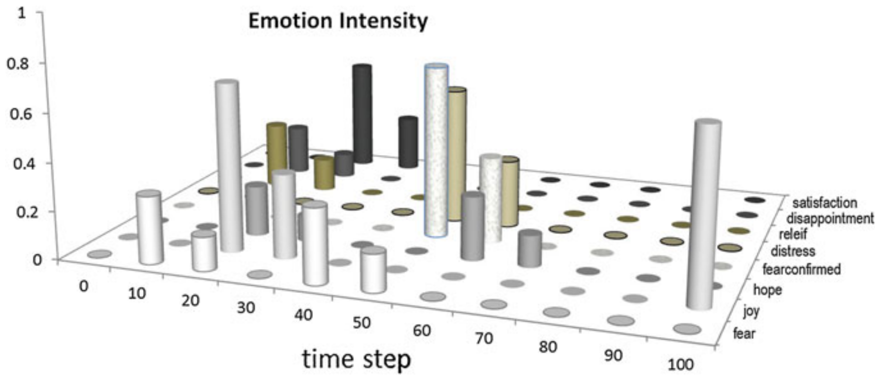


Fig. 4 Intensity of all events-originated emotions for the agent during the simulation

experienced emotion joy for the first time at step = 20 with a high intensity of 0.7 as a result of the occurrence of event e_1. The joy emotion started to deteriorate due to the normal decay and it completely disappeared by step = 40. The agent ended the simulation with another wave of joy emotion as a result of the re-occurrence of e_1.

5.2 Scenario 2: Regulating Hyper Emotional States

This experiment shows a scenario at which a psychotherapist is applying an external stimulus in order to regulate the hyper anger emotion of his patient trying to transition him to calm state. The stimuli to the regulation state machine is a set of 5 video clips of $V = \{v_1, v_2, v_3, v_4, v_5\}$. It is assumed that these video clips have a one-to-one correspondence with the set of events of $E = \{e_1, e_2, e_3, e_4, e_5\}$ explained in Table 1. The ultimate goal for the psychotherapist is to find the best stimulus v_i that would take the patient to the target emotional state.

In this scenario, it would be necessary to obtain the appraisal vector of each stimulus and study its impact on the anger emotion of the agent. Figure 5 represents the full automaton for all possible transitions that can transition the agent from angry state to calm state. The values of the fuzzy membership functions were obtained from applying the $sup - min$ combination operator on the fuzzy rules resulted from the appraisal processes on the two dimension of desirability and arousal. For instance, the fuzzy membership functions for event e_1 were obtained as follows:

$$\begin{bmatrix} 0.89 & 0.14 \\ 0.94 & 0.04 \end{bmatrix}$$

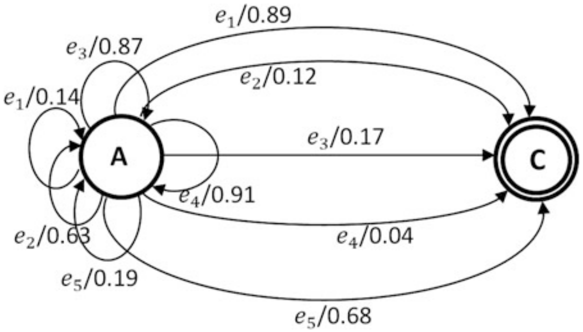


Fig. 5 Fuzzy automaton for transitioning between Anger to calm emotional states

Hence, in order to determine the influence of applying e_1 to angry state we would have,

$$[1 \quad 0] \circ \begin{bmatrix} 0.89 & 0.14 \\ 0.94 & 0.04 \end{bmatrix} = [0.14 \quad 0.89]$$

It is clear that the most reliable stimulus that would transition the angry agent to a calm state would be e_1.

6 Conclusion

In this article a fuzzy appraisal approach for anticipating the set of emotions that will be experienced by an agent based on OCC emotion theory was proposed. These emotions are elicited as a result of either the occurrence of goal-relevant events; actions of self or other agents; or a dispositional reaction to some emotion-eliciting objects. Emotion generation modules were formulated for all 22 emotions of the OCC model according to this ternary classification. The problem formulation was performed based on a set of different appraisal processes used to determine the value for a set of appraisal dimensions such as the desirability of events, degree of event's expectedness, level of involvements, etc. At the core of each assessment process there exist a fuzzy evaluation system that analyzes the competent appraisal variables and generates the value for the output parameters.

The proposed model was able to determine the set of triggered emotions along with their intensities at any point of time during the simulation period. The authors of this article believe that this work is still at the preliminary level and there is much room for further development and research that can use the obtained methods and results to bridge to the relevant disciplines, especially psychology and healthcare.

References

1. A.L. Baylor, A virtual change agent: motivating pre-service teachers to integrate technology in their future classrooms. Educ. Technol. Soc. **1**, 309–321 (2008)
2. R. Calvo, S. Mello, Affect detection: an interdisciplinary review of models, methods and their applications. IEEE Trans. Affect. Comput. **1**, 18–37 (2010)
3. P. Ekman, An argument for basic emotions. Cogn. Emot. **6**, 169–200 (1992)
4. M.S. El Nasr, J. Yen, Flame: fuzzy logic adaptive model of emotions. Auton. Agent. Multi-Agent Syst. **3**(3), 219–257 (2000)
5. C. Elliott, The affective reasoner: a process model of emotions in a multi-agent system, in *Proceedings of the First International Conference on Autonomous Agents* (1997)
6. N.H. Frijda, *The Emotions* (Cambridge University, New York, 1986)
7. P. Gebhard, Alma: a layered model of affect, in *Fourth International Joint Conference on Autonomous Agents and Multiagent Systems* (2005)
8. J. Gross, R. Thompson, Emotion regulation: conceptual foundations, in *Handbook of Emotion Regulation.* (Guilford Press, New York, 2006)
9. J. LeDoux, *The Emotional Brain* (Simon and Schuster, New York, 1996)
10. E.H. Mamdani, S. Assilian, an experiment in linguistic synthesis with a fuzzy logic controller. Int. J. Machine Stud. **7**(1), 1–13 (1975)
11. S. Marsella, J. Gratch, Ema: a model of emotional dynamics. Cogn. Syst. Res. **10**(1), 70–90 (2009)
12. A. Ortony, G. Clore, A. Collins, *The Cognitive Structure of Emotions* (Cambridge University Press, Cambridge, 1988)
13. R.W. Picard, *Affective Computing* (The MIT Press, Cambridge, 1997)
14. T.J. Ross, *Fuzzy logic with Engineering applications* (Wiley, New York, 2004)
15. J.A. Russell, Core affect and the psychological construction of emotion. Psychol. Rev. **110**(1), 140–172 (2003)
16. K. Scherer, *Appraisal Considered as a Process for Multi-level Sequential Checking*, ed. by K.R. Scherer, A. Schorr, T. Johnstone appraisal processes in emotion: theory, methods, research pp. 92–120 (2001)
17. C. Smith, R. Lazarus, *Handbook of Personality: Theory and Research, Chap. Emotion and Adaptation* (Guilford Press, New York, 1990), pp. 609–637
18. A. Soleimani, Z. Kobti, Event-driven fuzzy paradigm for emotion generation dynamics, in *Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering and Computer Science 2013*, WCECS 2013, San Francisco, pp. 168–173, 23–25 Oct 2013
19. B.R. Steunebrink, D.M. Meyer, The occ model revisited, in *4th Workshop on Emotion and Computing* (2009)
20. J. Tao, T. Tan, *Affective information processing* (Springer, London, 2009)

Chapter 18
A Graph-Based Model for Quadratic Separable Programs and Its Decentralization

Jaime Cerda, Alberto Avalos and Mario Graff

Abstract This document proposes a Newton step graph-based model for Quadratic separable Problems (QSP). The Newton step is well suited for this kind of problems, but when the problem size grows the matrix-based QSP model will grow in a non linear manner. Furthermore, handling the constraints becomes the main problem as we have to select the right constraints in the different solution steps. When this happens, the sparse matrix representation is the path to follow, but very little has been made in order to fully explode the sparsity structure. Indeed, the Hessian matrix for the QSP model has a very particular structure which can be exploited by using the graph underlying the problem, this is the approach taken in this document. To this end a graph is built derived from the components involved in the Newton step which describes the solution for the QSP problem. Based on this graph, the gradient can be evaluated directly based on the graph topology, as it will be shown, the information needed for such evaluation is embedded within the graph. These links eventually will guide the solution process in this approach. A deeper analysis of these links is done which leads to its complete understanding. It will be seen that the main effect of the link weakening operation is to allow the computation of the exact gradient. However, the solution will be reinforced by taking into account the second order information provided by the linking structure. Furthermore, the link weakening is used to separate coupled problems which in turn leads to a decentralized scheme. Finally, several decentralization approaches for the Newton step graph-based model for QSP are proposed.

Keywords DC-OPF · Decentralization · Graphs · NLP · Optimization · QSP

J. Cerda (✉) · A. Avalos · M. Graff
Universidad Michoacana, Morelia, Mexico
e-mail: jcerda@umich.mx

A. Avalos
e-mail: javalos@umich.mx

M. Graff
e-mail: mgraffg@dep.fie.umich.mx

H. K. Kim et al. (eds.), *Transactions on Engineering Technologies*,
DOI: 10.1007/978-94-017-9115-1_18,

List of Symbols

N	Number of decision variables
L	Number of equality constraints
M	Number of inequality constraints
z_i	Decision variable i
$\lceil z_i \rceil$	Upper limit value of variable z_i
$\lfloor z_i \rfloor$	Lower limit value of variable z_i
$\overline{z_i}$	Slack variable for z_i upper bound
$\underline{z_i}$	Slack variable for z_i lower bound
Δz_i	Variable x_i increment
$f(z)$	Objective function
$g_l(z)$	Equality constraint l
$h_m(z)$	Inequality constraint m
a_{li}	ith coefficient in equality constraint l
b_{mi}	ith coefficient in inequality constraint m
r_l	RHS of equality constraint l
s_m	RHS of inequality constraint m
λ_l	Dual variable for equality constraint l
μ_m	Dual variable for inequality constraint m
$\overline{\rho}_i$	Dual variable for z_i upper bound
$\underline{\rho}_i$	Dual variable for z_i lower bound
$\wp(S)$	Power set of set S
Γ_i	Set of variables connected to z_i

1 Introduction

Quadratic separable problems (QSP) are non linear problems whose objective function can be decomposed as a sum of functions, in this case quadratic functions, which involve only one variable. Previous works [1–4] have presented decentralized approaches for QSPs applied to Electrical Power Systems, which rely on the auxiliary principle problem [5]. In this work a graph based method to achieve this decentralization is proposed, as presented in [6]. To this end, we have chosen to weak the links, which are part of the graph, as an alternative. These links eventually will guide the solution process in this approach, which implies to weak the links which are coupling the problem in order to achieve some goal, perhaps decentralization. A deeper analysis of these links is done which leads to its complete understanding. It will be seen that the main effect of the link weakening operation is to allow the computation of the exact gradient. However, the solution will be reinforced by taking into account the second order information provided by the linking structure. This document is structured as follows. First, the graph based model is presented. Then, the self contained characteristic of the graph is presented. After this a QSP topological model is presented, which gives a standard graph-based representation for QSP, a goes further by proposing a simpler representation for those graphs. Following, a graph analysis is undertaken based on the linear equation each node and its links represents. Finally, different decentralization schemes are presented based on the previous analysis.

2 A QSP Topological Model Proposal

In this section a graph topology for the Newton step method is proposed as described in [7]. For this purpose, let us base the discussion with the QSP described by the objective function given by expression 1 and the constraints described by expressions from 2 to 4. This QSP consists of N variables, L equality constraints, and M inequality constraints.

$$\min_{z_i} \quad \sum_{i=1}^{N} \alpha_i + \beta_i z_i + \gamma_i z_i^2 \tag{1}$$

s.t.

$$\sum_{i=1}^{N} a_{li} z_i = b_l \quad 1 \leq l \leq L \tag{2}$$

$$\sum_{i=1}^{N} c_{mi} z_i \leq d_m \quad 1 \leq m \leq M \tag{3}$$

$$\lfloor z_1 \rfloor \leq z_n \leq \lceil z_N \rceil \quad 1 \leq n \leq N \tag{4}$$

We will show, by means of an example, that this model can be represented with the graph show in Fig. 1.

Each constraint is represented by a dual variable and a set of links which represent the linear terms within the constraint. The terms in the constraints are represented by links which join the primal variables with the dual variables. The only difference between equality constraints and inequality constraints is the kind of links used to build the linking structure. In the case of equality constraints, the linking structure will be active along the whole solution process. On the other hand, the linking structure for the inequality constraints will be active only when the constraint is binding.

Let us take as a example the Economical Dispatch model [8], whose model is described by equations from 5 to 8. The economical cost model for each generator is given as a convex quadratic function, generally $C_g(p_g) = \alpha_g + \beta_g p_g + \gamma_g p_g^2$.

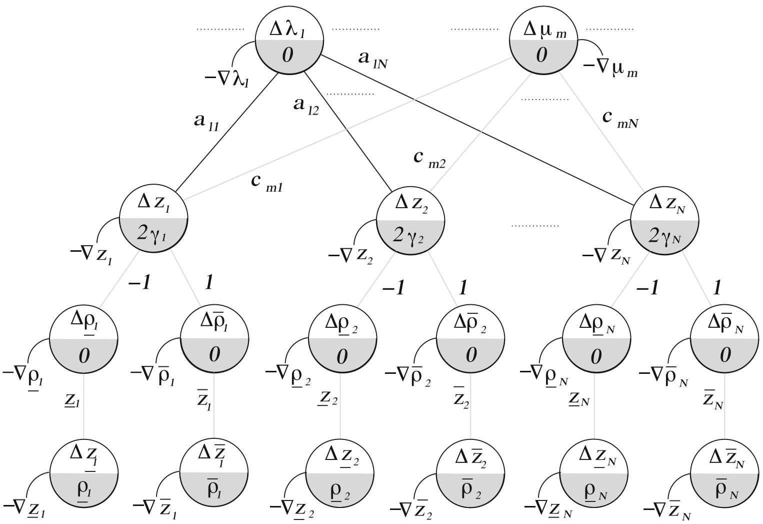


Fig. 1 Proposed topology for the Newton step method

$$\max_{p_g,\, q_l} \quad \sum_{l\in\mathbb{L}} B_l(q_l) - \sum_{g\in\mathbb{G}} C_g(p_g) \tag{5}$$

s.t.

$$\sum_{g\in\mathbb{G}_i} p_g - \sum_{l\in\mathbb{L}_i} q_l - Q = 0 \tag{6}$$

$$\lfloor P_g \rfloor \le p_g \le \lceil P_g \rceil \forall g \in \mathbb{G} \tag{7}$$

$$\lfloor Q_l \rfloor \le q_l \le \lceil Q_l \rceil \forall l \in \mathbb{L} \tag{8}$$

In the same way the economical benefit model for each variable load is represented by the concave quadratic function $B_l(q_l) = \beta_l q_l - \gamma_l q_l^2$. The economical dispatch problem maximizes the social welfare which is defined as the difference between the total benefit $\sum_{l\in\mathbb{L}} B_l(q_l)$ of using the energy and the cost $\sum_{g\in\mathbb{G}} C_g(p_g)$ invested to produce that power. In this model, a variable number of elements can be attached to it such as generators (p_g), variable loads (q_l) as well as a constant load (Q). If there are more than one constant load attached, these can be summed up and treated as a compound fixed load. The economical dispatch study goal is to find an optimal allocation for the energy each generator will produce such that meets the load demands. This solution has to be within the feasible space which the generators, the variable loads and the total fixed load impose.

Table 1 Table describing all the components which involve the Newton step

$$\mathscr{L}(z)=\lambda\tilde{g}(p_g,q_l)+\sum_{g\in\mathbb{G}}(C_g(p_g)+\underline{\rho}_g\tilde{h}_{\underline{p_g}}(p_g,\underline{p_g})+\overline{\rho}_g\tilde{h}_{\overline{p_g}}(p_g,\overline{p_g}))$$
$$-\sum_{l\in\mathbb{L}}(B_l(q_l)-\underline{\mu}_l\tilde{h}_{\underline{q_l}}(q_l,\underline{q_l})-\overline{\mu}_l\tilde{h}_{\overline{q_l}}(q_l,\overline{q_l}))$$

z	$\nabla(\mathscr{L}(z))$	$H(\mathscr{L}(z))$										
		p_g	q_l	$\underline{p_g}$	$\overline{p_g}$	$\underline{q_l}$	$\overline{q_l}$	λ	$\underline{\rho}_g$	$\overline{\rho}_g$	$\underline{\mu}_l$	$\overline{\mu}_l$
p_g	$\beta_g+2\gamma_g p_g-\lambda-\underline{\rho}_g+\overline{\rho}_g$	$2\gamma_g$						-1	-1	1		
q_l	$-\beta_l+2\gamma_l q_l+\lambda-\underline{\mu}_l+\overline{\mu}_l$		$2\gamma_l$					1			-1	1
$\underline{p_g}$	$\underline{\rho}_g\underline{p_g}$			$\underline{\rho}_g$					$\underline{p_g}$			
$\overline{p_g}$	$\overline{\rho}_g\overline{p_g}$				$\overline{\rho}_g$					$\overline{p_g}$		
$\underline{q_l}$	$\underline{\mu}_l\underline{q_l}$					$\underline{\mu}_l$					$\underline{q_l}$	
$\overline{q_l}$	$\overline{\mu}_l\overline{q_l}$						$\overline{\mu}_l$					$\overline{q_l}$
λ	$\sum_{g\in\mathbb{G}}p_g-\sum_{l\in\mathbb{L}}q_l-Q$	-1	1									
$\underline{\rho}_g$	$\lfloor p_g\rfloor-p_g+\underline{p_g}^2/2$	-1		$\underline{p_g}$								
$\overline{\rho}_g$	$p_g-\lceil p_g\rceil+\overline{p_g}^2/2$	1			$\overline{p_g}$							
$\underline{\mu}_l$	$\lfloor q_l\rfloor-q_l+\underline{q_l}^2/2$		-1			$\underline{q_l}$						
$\overline{\mu}_l$	$q_l-\lceil q_l\rceil+\overline{q_l}^2/2$		1				$\overline{q_l}$					

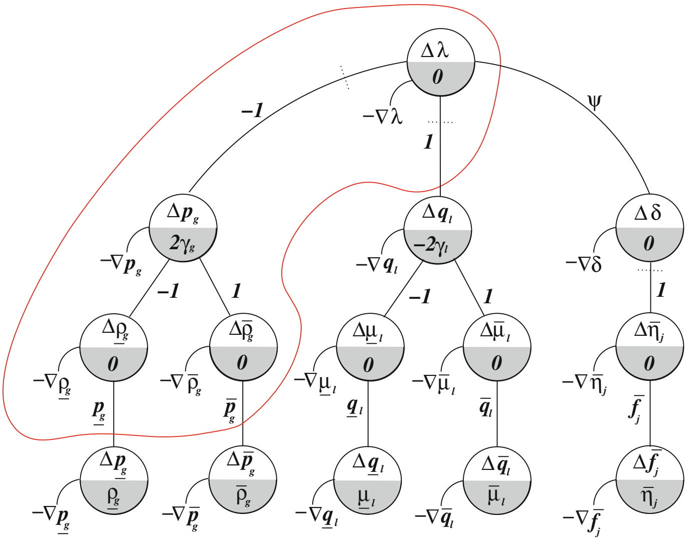


Fig. 2 Subgraph for a primal variable (i.e. p_g)

Table 1, shows all the elements involved in the Newton Step computation, and its corresponding graph is presented in Fig. 2.

3 Self Contained Graphs

An interesting fact when this kind of graphs is used is that the information to rebuild the gradient is contained in the graph topology. First, the case where the gradient for a primal variable, p_g, is analyzed. The discussion will be focused on the subgraph delineated in Fig. 2. From Table 1, it is known that $\nabla p_g = \beta_g + 2\gamma_g p_g - \lambda - \underline{\rho}_g + \overline{\rho}_g$.

Figure 3 shows the gradient evaluation process for a primal variable. Let us suppose that every node has the value related to the variable which itself represents. Therefore, the node gradient evaluation starts by taking into account the gradient information within the node which in this case would be $\beta_g + 2\gamma_g p_g$, as shown in Fig. 3a.

Then it starts to evaluate the portion of the gradient which is a function of the variables contained by the neighbours of p_g, as shown in Fig. 3b, c and d.

Now, let us turn the attention to the case where the gradient for a dual variable is to be found, λ in this case. The discussion will be focused on the subgraph delineated in Fig. 4.

Figure 3 shows the gradient evaluation process for a dual variable. From Table 1 it is known that $\nabla_\lambda = RHS - p_g + q_l + \psi\delta$. As before, the gradient evaluation starts by taking into account the information contained within the node itself, in this case *RHS*, as shown in Fig. 5a. Then the evaluation of the links attached to this node and the variables at the other extreme of the link is performed as shown in Figs. 5b, c and d.

From the previous discussion, as there exist only dual variables and primal variables, the gradient for every variable can be derived straightforwardly from the graph topology. Therefore, the graph can be said to be self-contained as no external information is needed.

4 An Equivalent Graph Representation

In this section, once the characteristics of this graph have been analyzed, a simpler graph model representation is derived in order to make its handling easier. This simplification is based on two main observations: first, the bounding structures are fixed, and second the different kind of variables can be represented in such a way that the content of that node will be inferred by its representation.

4.1 An Equivalent Graph Bounding Structure Representation

The subgraphs which represent the bound on the variables are well defined, furthermore, in [9] an approach to ignore bounding slack variables is presented.

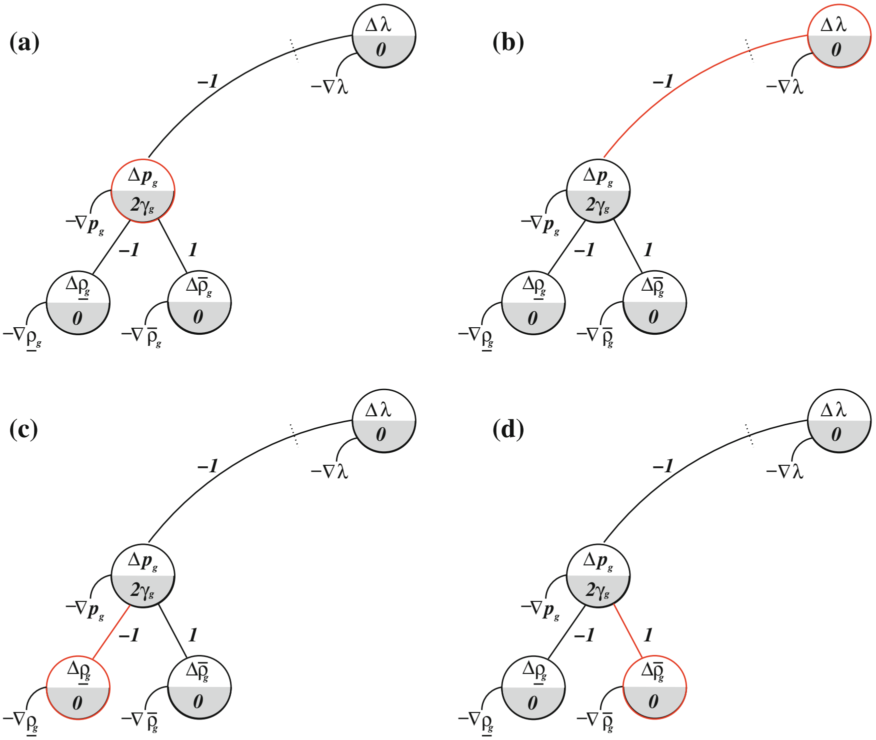


Fig. 3 Graph-based gradient evaluation for a primal variable (i.e. p_g). **a** $\nabla_{p_g} = \beta_g + 2\gamma_g p_g$, **b** $\nabla_{p_g} = \beta_g + 2\gamma_g p_g - \lambda$, **c** $\nabla_{p_g} = \beta_g + 2\gamma_g p_g - \lambda - \underline{\rho}_g$, **d** $\nabla_{p_g} = \beta_g + 2\gamma_g p_g - \lambda - \underline{\rho}_g + \overline{\rho}_g$

Therefore a special graph notation will be derived in order to handle them, as shown in Fig. 6. Here all the links and nodes contained in such subgraph are embedded within the triangle. The link value will be as follows:

$$link.value = \begin{cases} 1 & \text{if the constraint is lower binding,} \\ -1 & \text{if the constraint is upper binding,} \\ -- & \text{if is not binding.} \end{cases}$$

This leads to the representation shown in Fig. 7.

4.2 An Equivalent Node Type Representation

From the previous section, it has been learnt that the gradient is embedded within the graph topology and the information attached to each node. Therefore, a simplification for the graph representation will be derived. The last section has presented a representation for the bounding structures which control the limits on the

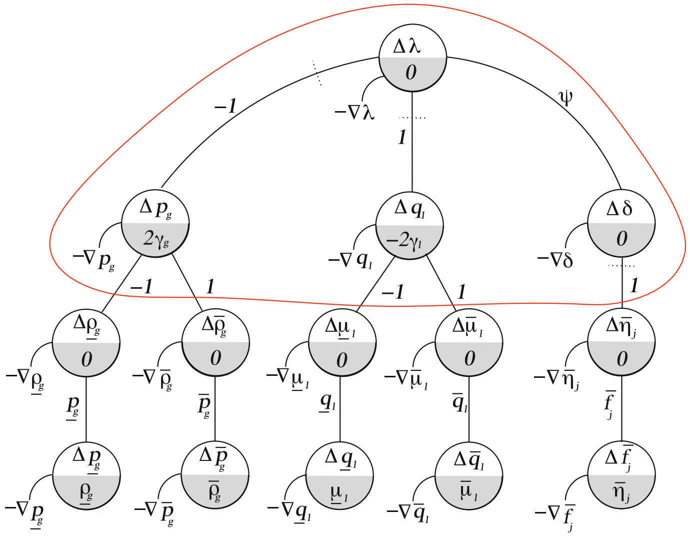


Fig. 4 Subgraph for a dual variable (i.e. λ)

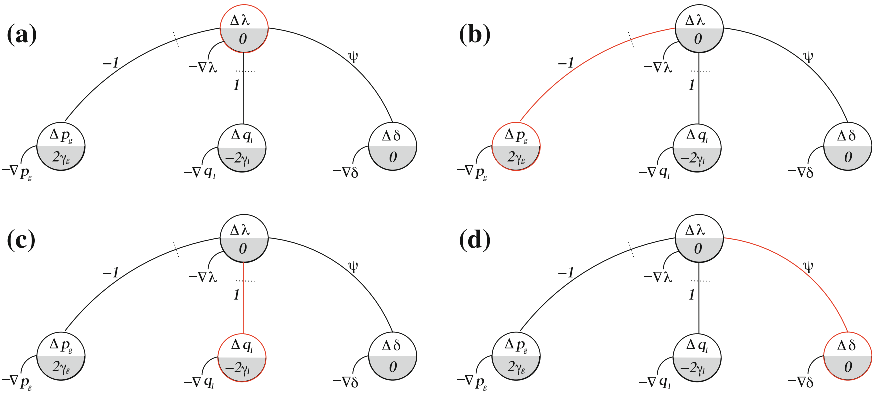


Fig. 5 Graph-based gradient evaluation for a dual variable (i.e. λ). **a** $\nabla_\lambda = RHS$, **b** $\nabla_\lambda = RHS - p_g$, **c** $\nabla_\lambda = RHS - p_g .. + q_l$, **d** $\nabla_\lambda = RHS - p_g .. + q_l .. + \psi\delta$

primal variables. Therefore, now the only nodes in the remaining graph are those representing the primal variables and the dual variables. As mentioned above, the primal variables set can be further divided into two sets. The first one represents those variables which are part of the objective function. The second one contains those primal variables which appear only within the constraints. An instance of these would be the variable representing the electrical angle in the electric power market example (i.e. δ). These two subsets will be called objective and non-objective variables respectively. Therefore, there are three kinds of variables which have to be represented within the graph. These representations are shown in

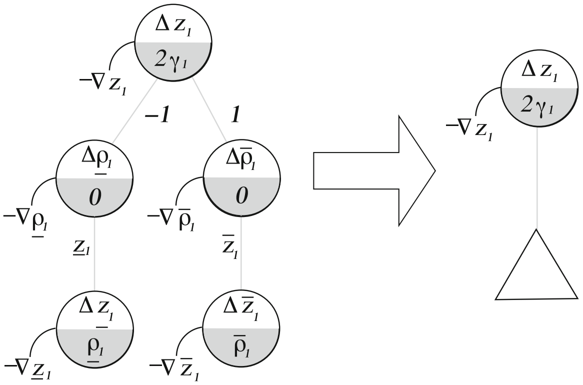


Fig. 6 Bound subgraph representation

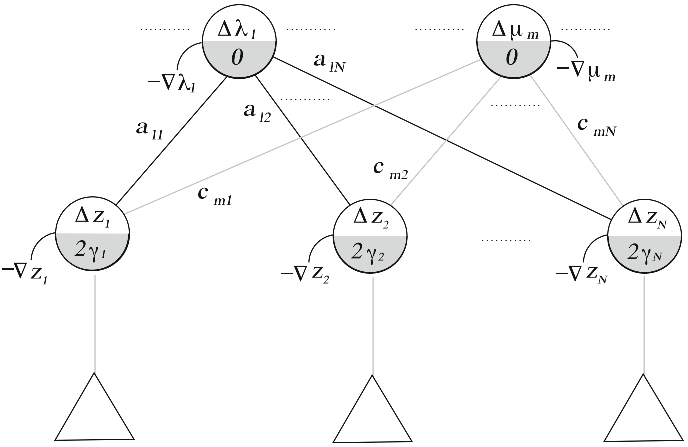


Fig. 7 Hessian topology—modified representation

Fig. 8. Based on the type of variable this node is representing, the information attached to it will be known. This information is as follows

- Objective variables: Attached to this node will be the linear coefficient as well as the quadratic coefficient in order to be able to compute its gradient,
- Non objective variables: To this node there will be no additional information since its coefficients in the constraints are given by the values of the links which are attached to it,
- Dual variables: The information attached to this kind of node will be the right hand side of the constraints which will allow its gradient computation.

This leads to the representation shown in Fig. 9, where z_2 is assumed as a nonobjective variable. Based on this graph the appropriate classes for each type of

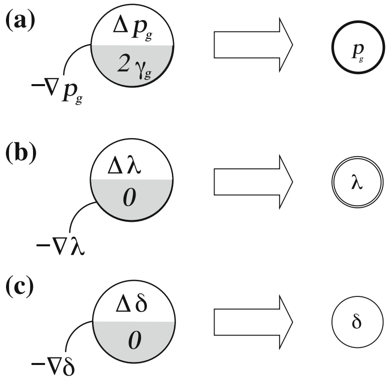


Fig. 8 Variable representation. **a** Primal variable, **b** dual variable, **c** non objective variable

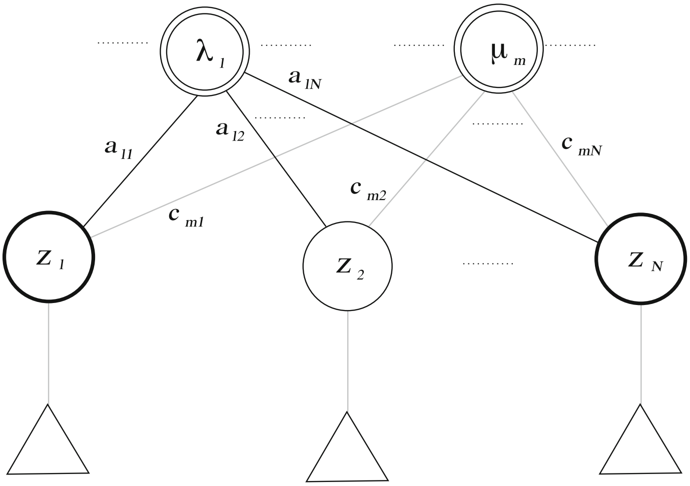


Fig. 9 Hessian topology—final equivalent representation

variable can be defined. Once these definitions have been implemented, the operations to solve the graph can be implemented straightforward.

5 Graph Analysis

A node and the links which are attached to it represent an equation. In this section the analysis for a node and the equation it represents is done. To this end let us extract the equation corresponding to z_i from the system of linear equations which describes the Newton step. This is given by Eq. 9.

$$\frac{\partial^2 \mathcal{L}(z)}{\partial^2 z_i}\Delta z_i + \sum_{\forall j \in \Gamma_i} \frac{\partial^2 \mathcal{L}(z)}{\partial z_i \partial z_j}\Delta z_j = -\nabla_{z_i}\mathcal{L}(z) \tag{9}$$

solving for Δz_i leads to

$$\Delta z_i = \frac{-\nabla_{z_i}\mathcal{L}(z) - \sum\limits_{\forall j \in \Gamma_i} \frac{\partial^2 \mathcal{L}(z)}{\partial z_i \partial z_j}\Delta z_j}{\frac{\partial^2 \mathcal{L}(z)}{\partial^2 z_i}} \tag{10}$$

This can be rewritten as

$$\Delta z_i = \frac{-\nabla_{z_i}\mathcal{L}(z)}{\frac{\partial^2 \mathcal{L}(z)}{\partial^2 z_i}} - \sum_{\forall j \in \Gamma_i} \frac{\frac{\partial^2 \mathcal{L}(z)}{\partial z_i \partial z_j}}{\frac{\partial^2 \mathcal{L}(z)}{\partial^2 z_i}}\Delta z_j \tag{11}$$

This expression can be thought as the improvement in the solution for the component in the orthogonal axis z_i. It can be split into two parts. The first part, described by expression Eq.12, is a component which always appear in any decentralization approach for the graph.

$$\frac{-\nabla_{z_i}\mathcal{L}(z)}{\frac{\partial^2 \mathcal{L}(z)}{\partial^2 z_i}} \tag{12}$$

This is the contribution based on $-\nabla_{z_i}\mathcal{L}(z)$ just like in the steepest descent methods. However the length of the step will be reinforced with the second order information provided by $1/\frac{\partial^2 \mathcal{L}(z)}{\partial^2 z_i}$. This will be the case for the primal variables, however for the dual variables there will not be second order information, and therefore the gradient step size will have to be controlled by other means.

The second part, described by expression Eq. 13, is composed by all the second order contributions which will be collected by z_i from its neighbours (i.e. Γ_i).

$$-\sum_{\forall j \in \Gamma_i} \frac{\frac{\partial^2 \mathcal{L}(z)}{\partial z_i \partial z_j}}{\frac{\partial^2 \mathcal{L}(z)}{\partial^2 z_i}}\Delta z_j \tag{13}$$

This part will have a variable number of components which will depend on the applied decentralization approach. These can go from taking into account the second order information from all the neighbors to the other extreme where no second order information from them will be collected at all. The first approach would be the full centralized Newton step and the second one would result in the steepest descent reinforced with the second order information for the same orthogonal axis. Nevertheless, between these two approaches there is a plethora of options about which of the components of second order information can be taken into account. In fact there are $|\wp(\Gamma_i)|$ choices and the choice at any point will be based on the particular decentralization approach. Let us define $\mathcal{L}_h$ as the set of links which are taken into account for this process. With this in mind expression 13 can be rewritten as expression 14.

$$-\sum_{\forall j \in \Gamma_i (i,j) \in \mathcal{L}_h} \frac{\frac{\partial^2 \mathcal{L}(z)}{\partial z_i \partial z_j}}{\frac{\partial^2 \mathcal{L}(z)}{\partial^2 z_i}} \Delta z_j \tag{14}$$

6 The Graph and Its Decentralization

In this section the graph and its decentralization is addressed. To this end an operation over the links of the graph, called *link weakening*, is defined as in [10]. Then, three different approaches to decentralize the graph are proposed. The first one is a complete decentralized, the second approach is based on the notion of the kind of variables within the graph (i.e. primal and dual variables); and the third approach will be based on agency definitions. To this end let us modify the system which we have been using along the document to the one shown in Fig. 10a. The graph representation corresponding to this example is shown in Fig. 10b, where the minus sign represents a -1 value for the link.

6.1 Link Weakening

Before going into the decentralization approaches, let us define the link weakening operation which allows the decentralization process. This operation labels the links with one of the following two labels.

- HARD—This labeling will be granted to those links which are not part of the decentralization process. The graph reduction process will take into account these links,
- SOFT—These links provide the means to decentralize the graph. If the link possesses this property, then the reduction process will not pass through them. Nevertheless, by using its connectivity, they will provide a means to retrieve the actual value of the variable at the other end of the link which will allow the gradient to be computed, as described in Sect. 3.

6.2 A Gradient-Oriented Approach

The first approach is to decentralize the graph in an extreme way by weakening all the links as shown in Fig. 11. This method leads to a model where the gradient method has to be applied at each node in the graph.

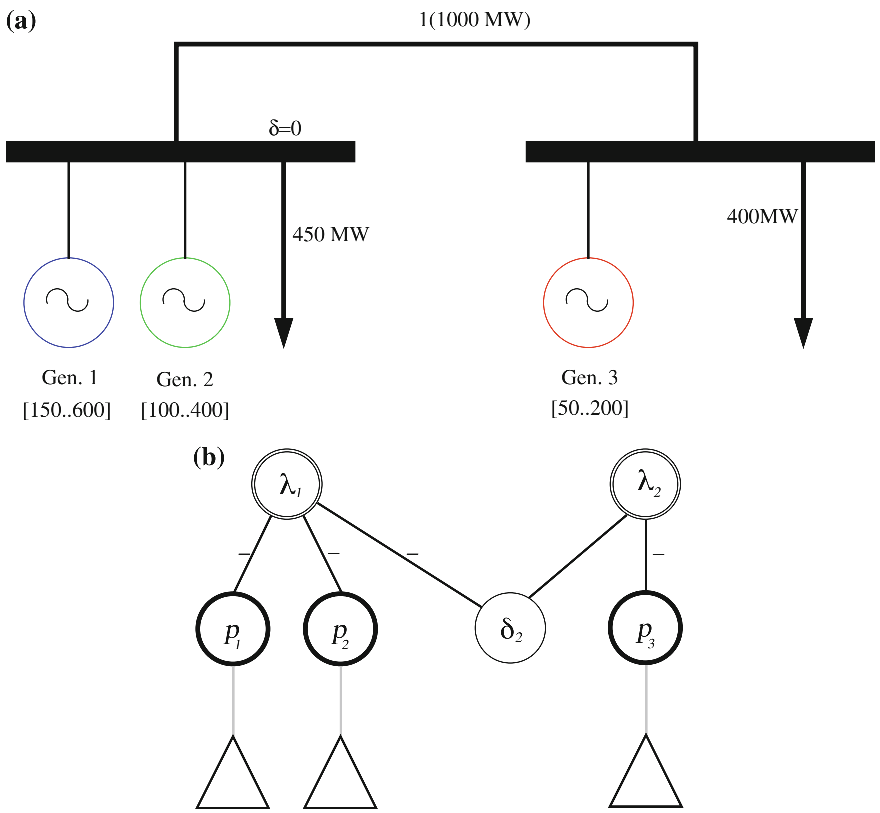


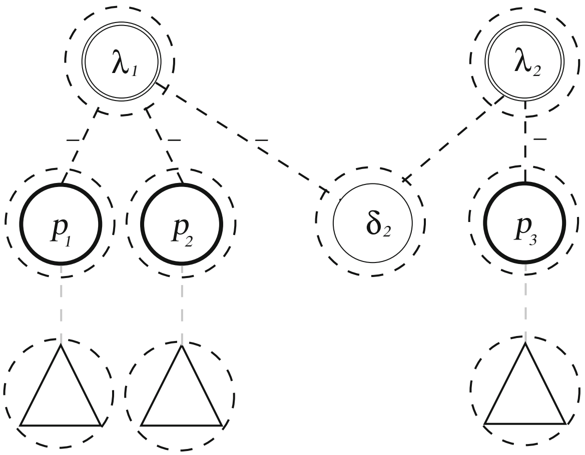
Fig. 10 Two nodes system and its graph representation. **a** Two nodes system. **b** Its graph representation

From this figure and based on Eq. 11 we can assert it will become Eq. 15, where the second order information of all of its neighbors is disregarded. In this case $\mathcal{L}_h = \emptyset$. Therefore Eq. 11 becomes Eq. 15.

$$\Delta z_i = \frac{-\nabla_{z_i}\mathcal{L}(z)}{\frac{\partial^2 \mathcal{L}(z)}{\partial^2 z_i}} \tag{15}$$

Nevertheless, those nodes which have proper second order information will be able to use it in order to speed up the convergence process. In particular, all the nodes related to primal variables have this information. Dual variables do not have second order information at all and therefore they will have to use Eq. 16

Fig. 11 An gradient-based decentralization approach



$$\varDelta z_i = -\kappa \nabla_{z_i} \mathcal{L}(z) \tag{16}$$

The main drawback of gradient methods known also as steepest descent methods is the hardness to estimate κ. Therefore in the primal nodes Eq. 17 holds.

$$\kappa = \frac{1}{\frac{\partial^2 \mathcal{L}(z)}{\partial^2 z_i}} \tag{17}$$

6.3 A Dual-Oriented Approach

The second natural approach to decentralize the graph is the dual-oriented approach. From the model proposed in Sect. 2 it is known the dual variables are in only one layer so if a line across both layers is drawn dissecting the graph, the links which connected the dual variables with the primal variables will be weakened as shown in Fig. 12.

The only dual variables considered in this case are those related with constraints involving two or more primal variables (i.e. bound dual variables are not split from the primal variables set). Let us denote $\mathcal{D}$ as the set of those dual variables. Therefore Eq. 11 becomes Eq. 18, where all the second order information about the dual variables are disregarded by the primal variables. On the other hand, as the dual variables only have links with primal variables, they are now isolated just as in the gradient approach.

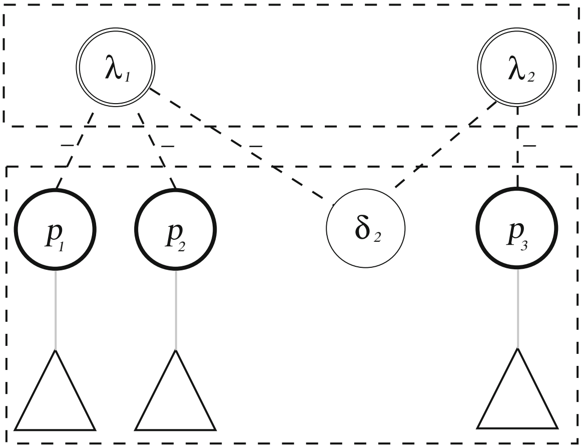


Fig. 12 A dual-oriented decentralisation approach

$$\Delta z_i = \frac{-\nabla_{z_i}\mathcal{L}(z)}{\frac{\partial^2\mathcal{L}(z)}{\partial^2 z_i}} - \sum_{\forall j\in\Gamma_i z_j\notin\mathcal{D}} \frac{\frac{\partial^2\mathcal{L}(z)}{\partial z_i\partial z_j}}{\frac{\partial^2\mathcal{L}(z)}{\partial^2 z_i}} \Delta z_j \tag{18}$$

6.4 An Agent-Oriented Approach

In this approach the decentralization process is made based on concepts drawn from the multi-agent community. A classical definition for an agent is

"An agent is a computer system *situated* in an *environment*, and capable of *flexible autonomous action* in this *environment* in order to meet its *design objectives*" (adapted from [11]).

The interpretation in this work for an agent is an entity which possesses some states or variables and presents an independent and proactive behavior, represented by their objective function. Furthermore, it is situated in an environment which he can sense and act accordingly. However, he is also constrained by its own limitations as well as the constrains presented by the environment. Going back to the problem we are addressing in this document, the DC OPF. There the agents would be acting on behalf of each node. Therefore, their own limitations would be the generation levels for the generators and the power balance at each node. On the other hand the constraints represented by the environment would be represented by the transmission constraints. To cope with this paradigm, the graph is split into subsets of primal variables, links, and dual variables. Based on the membership of these components, the graph is decentralized as shown in Fig. 13. This approach was the one taken for [12, 13]

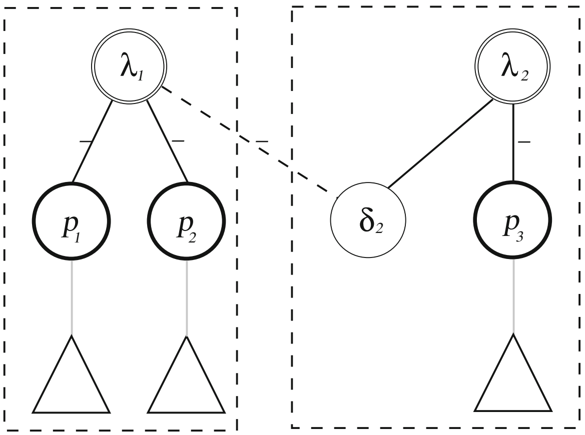


Fig. 13 An agent-based decentralization approach

7 Concluding Remarks

This document has presented a graph model proposal for quadratic separable problems with linear constraints. The ingredients involved in the Newton step, which is based on the Lagrangian of the system, have been discussed. Furthermore, in this representation it has been notorious the sparsity patterns of both types long term and temporary. The first are patterns which are present at the beginning of every Newton step iteration, whereas the second ones are patters which can differ between each Newton step iteration. Then, it has been shown how to transit from the matrix model to the graph model, which reflects the sparsity already noticed in the matrix representation. Following, by analyzing the graph interconnections, it has been uncovered the self-containing characteristic whose basic mean is that no information beyond that involved in the graph is needed to compute the gradient. This allow us to compute the gradient directly from the graph, provided the correct information is attached to each node. Then a standard topological model proposal for the Newton step, when applied to QSP, has been presented. This topological model lends itself to a simpler representation which allows its direct implementation. Then the underlying decentralization principles have been presented based on the analysis of the equation represented by the node and its links. This allow us to compute the gradient directly from the graph, provided the correct information is attached to each node. Finally, three decentralization approaches have been described. The first one is a totally decentralized approach which will lead us to a gradient oriented model reinforced with its proper second order information. The second one is based on the type of variables (i.e. primal or dual), and leads us to a horizontal graph split. The last approach is based on concepts drawn from multi-agents community.

References

1. A.G. Bakirtzis, P.N. Biskas, A decentralized solution of the dc-opf of interconnected power systems. IEEE Trans. Power Syst. **18**(3), 1007–1013 (2003)
2. A. Bakirtzis, P.N. Biskas, N. Macheras, N. Pasialis, A decentralized implementation of dc optimal power flow on a network of computers. IEEE Trans. Power Syst. **20**(1), 25–33 (2000)
3. P. Biskas, A. Bakirtzis, Decentralised opf of large multiarea power systems. IEE Proc. Gener. Transm. Distrib. **153**(1), 99–105 (2006)
4. A.J. Conejo, J.A. Aguado, Multi-area coordinated decentralized dc optimal power flow. IEEE Trans. Power Syst. **13**(4), 1272–1278 (1998)
5. G. Cohen, Optimization by decomposition and coordination: a unified approach. IEEE Trans. Autom. Control **2**(2), 222–232 (1978)
6. J. Cerda, D. De Roure, A graph-based decomposition method for quadratic separable programs, in *SIAM Conference on Optimization*, (Society for Industria and Applied mathematics, SIAM, Ed. Boston, Massachusetts, USA) 10–13 May 2008
7. J. Cerda, A. Avalos, A graph model proposal for convex non linear separable problems with linear constraints, in *Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering and Computer Science*, WCECS 2013, San Francisco, CA, USA, pp. 129–133, 23–25 Oct 2012
8. J. Cerda, D. De Roure, A graph-based economical dispatch model, in *Proceedings of WORLDCOMP 2010: Foundation of Computer Sciences*, ed. by H.R. Arabnia, G.A. Gravvanis, A.M.G. Solo (CSREA Press, Las Vegas, Nevada, USA) pp. 132–138, 6–9 April 2010
9. P. Cira, J. Cerda, J.J. Flores, A graph-based economical dispatch model, in *Procedia Technology: Proceedings of The 2012 Iberoamerican Conference on Electronics Engineering and Computer Science*, vol. 3 (Elsevier, Guadalajara, Mexico) pp. 304–315, 23–25 Oct 2012
10. J. Cerda, M. Graff, A graph-based decentralization proposal for convex non linear separable problems with linear constraints, in *Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering and Computer Science*, WCECS 2013, San Francisco, CA, USA, pp. 41–45, 23–25 Oct 2012
11. N. Jennings, M. Wooldridge, Intelligent agents: theory and practice. Knowl. Eng. Rev. **10**(2), 115–152 (1995)
12. J. Cerda, D. De Roure, E. Gerding, An agent-based electrical power market simulator, in *AAMAS '08: Proceedings of the 7th International Joint Conference on Autonomous Agents and Multi-agent Systems. Richland*, SC: International Foundation for Autonomous Agents and Multiagent Systems, 2008, pp. 1655–1656
13. J. Cerda, D. De Roure, An agent-based decentralisation approach for the electricity power market, in *Proceedings of the 2010 IEEE Electronics, Robotics and Automotive Mechanics Conference*, ed. by P. Kellenberger (IEEE, Cuernavaca Morelos, México), pp. 666–671, Sept 2010

Chapter 19
Predicting Congestions in a Ship Fire Evacuation: A Dynamic Bayesian Networks Simulation

Parvaneh Sarshar, Jaziar Radianti and Jose J. Gonzalez

Abstract In this paper, some new simulation results achieved from our proposed simulation model for analyzing congestions in ship evacuation are presented. To guarantee a safe evacuation, this model considers the most important real-life factors including, but not limited to, the passengers' panic, the age and sex of the passengers, the structure of the ship, and so on. The qualitative factors have been quantized in order to compute the probability of congestion during the entire evacuation. We then utilize the dynamic Bayesian network (DBN) to predict congestion and to handle the non-stationarity of the scenario with respect to the time. Considering the most important scenarios and running the simulation, we demonstrate the distinct effects of these factors on congestion. The role of decision supports (DS), i.e. smartphone evacuation applications and rescue team presence on congestion is also studied. In addition, the impact of congested escape routes on the evacuation time is also investigated. The presented model and results of this paper are possible decision support tools for maritime organizations, emergency management sectors, and rescuers onboard the ships, which try to alleviate the human or property losses.

Keywords Congestion · Decision support systems · Dynamic bayesian networks · Evacuation time · Simulation · Ship fire

P. Sarshar (✉)
Universitetet i Agder (UiA), Jon Lilletuns vei 9, Room A3 074C, 4879 Grimstad, Norway
e-mail: parvaneh.sarshar@uia.no

J. Radianti
Universitetet i Agder (UiA), Jon Lilletuns vei 9, Room A3 050, 4879 Grimstad, Norway
e-mail: jaziar.radianti@uia.no

J. J. Gonzalez
Universitetet i Agder (UiA), Jon Lilletuns vei 9, Room A3 081, 4879 Grimstad, Norway
e-mail: jose.j.gonzalez@uia.no

H. K. Kim et al. (eds.), *Transactions on Engineering Technologies*,
DOI: 10.1007/978-94-017-9115-1_19,

1 Introduction

Unfortunately, we witness different types of disasters all around the world each year. If we look for very recent disasters in 2013, we can find many natural or man-made disasters including Typhoon Soulik in China, Asiana Boeing 777 crashed and burned in San Francisco, or Boston marathon bombings. One of these hazardous situations is fire onboard a ship which requires an emergency evacuation either from some parts of the ship or from the entire ship. Ship fire is not a rare accident. In 2013, several cruise fires have been reported from all over the world. For instance, fire on royal Caribbean cruise in Bahamas or the Carnival Triumph ship in Alabama. All these crisis situations dealt with a large number of people and needed a proper and safe evacuation in time.

Ship emergency evacuations especially during fire have always been very critical and vital. Many studies have been done on how to have the best evacuation during crisis situations and how different factors influence the evacuation and evacuation time. These factors can be passengers' characteristics (age, health, gender, interrelationship, and the degree of panic), onboard crew (skill and training level), and the ship structure (the number of decks, size and location of emergency exits, seats arrangement, etc.). The aforementioned factors can cause to a bigger problem during evacuation, i.e., congestion in the escape routes.

The international maritime organization (IMO) has defined congestion as [1]: (1) initial density equal to, or greater than, 3.5 persons/m^2; or (2) accumulation of more than 1.5 persons per second between ingress and egress from a point. During an emergency evacuation, congestion can become critical and lead the evacuation time to exceed from the designated standards. This can be a surplus threat along with the actual hazard (fire) for passengers' lives during evacuation. Based on [2], a passenger may waste up to 89 % of their personal evacuation time held up in congestion.

Most of the investigations on the evacuation procedure have also looked at the concept of congestion, clogging or jamming (see, e.g., [2–8]). This is an evidence to show the critical importance of these concepts. To prevent or solve congestion during emergencies, a variety of models, simulations, algorithms, applications, software developments, and sensors technologies have been examined and/or corresponding regulations have been proposed in the literature. For instance, in [3, 4], a social force model has been employed to scrutinize congestion caused by crowds. Several cellular automaton models have been proposed in e.g. [4, 5] that focus on the occurrence of a congestion. Furthermore, Ferscha and Zia [6] developed a wearable device, called LifeBelt to handle the evacuation procedure and the inevitable congestion. Fujihara and Miwa [8] used opportunistic communications for handling congested routes. They showed that a correct guidance can effectively reduce the congestion and evacuation time. IMO has also organized some regulations [1] that should be applied by marine organizations to avoid congestions during ship evacuation.

Going deep in the literature reveals the fact that the literature on congestion still lacks appropriate models to dynamically describe complex and uncertain situations in time. In paper [9], we proposed a new congestion model based on DBN, which is a powerful tool to model highly uncertain conditions and to predict complex phenomena in time. In this paper, we expand the idea of congestion prediction by having a deeper insight into our proposed model and analyzing the factors affecting congestion. This helps us to find out more crucial factors affecting congestion and to predict how the probability of congestion changes during an evacuation process in different time steps in more realistic scenarios. We also study how the existence of decision supports (DS) and the reliability of extinguishing systems (ES) can influence the probability of congestion. The model and the results support the decision making process for the rescue team members and passengers onboard.

This work is a part of SmartRescue project, so the term decision supports (DS) in this paper refers to the existence of rescue team and an application developed in SmartRescue project for helping rescuers and passengers in emergency situation to automatically obtain the overview of the immediate threats by using sensors in the smartphones. The novelty of the paper arises from the utilization of the DBN for congestion modeling and analysis in ship evacuation, which has not been addressed in the literature before. The proposed model considers the most important factors causing congestion on escape routes. The corresponding input of the model has been derived from the literature which makes the model more realistic and trustworthy.

The remainder of this paper is structured as follows. Section 2 describes the idea of DBN. In Sect. 3, we illustrate our proposed DBN model. The results gained in this research are presented in Sect. 4 and eventually, Sect. 5 concludes the paper.

2 Dynamic Bayesian Network

The DBN is an extension of the Bayesian network (BN) [10] to model dynamic systems, which vary over the time. In this regard, a DBN allows a probabilistic graphical model to describe the level of uncertainty with variety of applications targeting the computational complexity reduction and reasoning under uncertain situations. Analysis of highly complex phenomena and decision making in complex situations where, e.g., variables are highly interlinked and/or data is ambiguous, are also parallel achievements of utilizing DBNs.

In DBNs, the system state at time t is shown by a set of random variables $\mathrm{X}^{\mathrm{t}} = \mathrm{X}^{\mathrm{t}}_{1} \ldots \mathrm{X}^{\mathrm{t}}_{\mathrm{n}}$. If the system state depends only on the immediate preceding state (i.e., k = 1), the system is called first-order Markov (Markov Chains) with the transition distribution $\mathrm{P}\left(\mathrm{X}^{\mathrm{t}} \middle| \mathrm{X}^{\mathrm{t}-1}\right)$ given by [11]:

$$P(X^t | X^{t-1}) = \prod_{i-1}^{n} P\left(X_i^t \middle| P_a\left(X_i^t\right)\right). \tag{1}$$

More details on the concept of DBN can be found in [11, 12]. Given GeNIe [12] as a domain to implement a DBN, the procedure of the design is quite similar to the one reported in [9]. Accordingly, we start from a static BN containing variables (nodes) and dependencies (arcs). Defining the conditional probability table (CPT) of the nodes, we then quantify the considered qualitative factors, which have the most important impact on occurring congestion. In each column of CPTs, the probability values are normalized to fall into the range of [0, 1]. As a result, the sum of the probabilities in each column must meet the unity.

To obtain the aforementioned probability distributions, there exist several methods in the literature. For instance, if the exact knowledge of data is not at hand, they can be determined based on suggestions from associated experts. Another classical method is to use maritime incident databases provided. Figures 2 and 3 display examples of the DBN developed in GeNIe. For a detailed description on developing a DBN in GeNIe refer to [9, 13].

In this paper, the proposed DBN models the factors affecting congestion occurrence, assisting rescue teams to obtain a sound understanding of the evacuation procedure. This enables them to make the best decisions in case of emergency.

3 Proposed DBN Model for Congestion

In this section, we first overview the main factors causing congestions during evacuation of a ship burning in fire. These factors are employed in the DNB model designed in GeNIe. The impact of congestion on the evacuation time is also studied.

To reduce the mathematical complexity of the proposed model, we consider some realistic assumptions, targeting a simple model. The simplification results in a very high-speed simulation model and very well suits for the optimization of the evacuation process, more especially in complex situations. This simplified model can be easily expanded into more complex one, if needed.

In this regard, we designed a simplified layout of a passenger ship, which is shown in Fig. 1a. According to this structure, the first deck of the ship consists of 3 compartments, A, B, and C, as well as a corridor that is divided into 3 sections, D1, D2, and D3. Stairs (or ramps for emergency exits) S1 and S2 link the three former sections to the compartment E which stands for the embarkation area (assembly station) in the second deck. Notice that compartment D2 has no direct access to the compartment E. It has the access via the compartments D1 and D3 and then the stairs.

The associated bidirectional graph illustrating the connections between the aforementioned compartments is shown in Fig. 1b. People are able to move from A to D1 or from B to D2 and vice versa, whereas for arriving at E, passing from S1 or S2 is indispensable. The loops connected to each compartment, present the fact

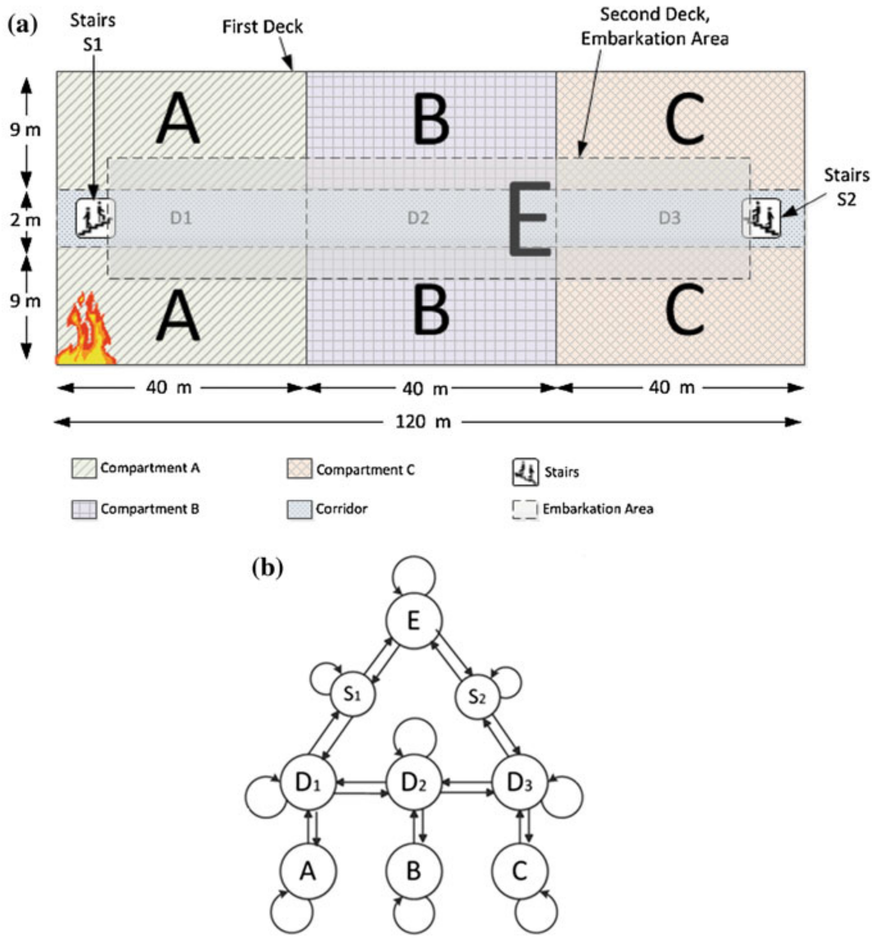


Fig. 1 The simplified ship layout (**a**) and the corresponding bidirectional graph (**b**)

that passengers might stay in their current locations as well (due to panic, congestion, or etc.).

To make the model more realistic, we made some assumptions:

- The ship engine is located in compartment A (back of the ship), and the fire starts in compartment A.
- There is no fire in compartment E. This compartment is the final destination of the passengers and the rescue team.
- All the passengers are equally spread all over the ship.
- Please note that, we have divided the entire evacuation time into 15 time steps. The standard evacuation should be finished in 15 time steps. Time step 0 is when the fire is detected.

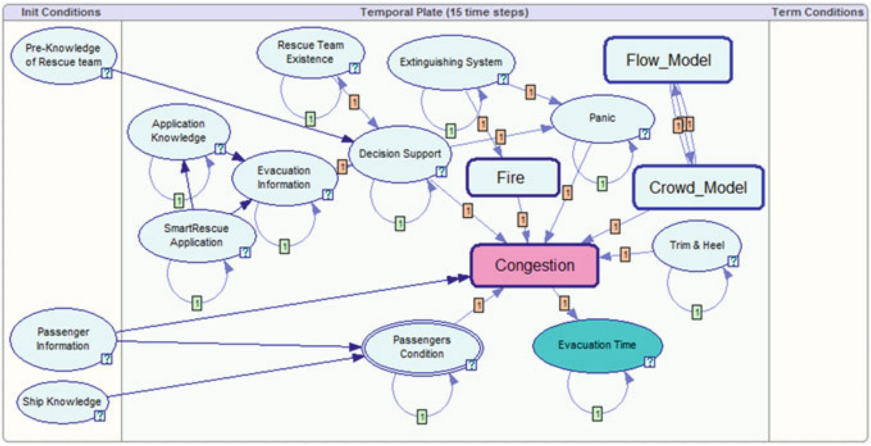


Fig. 2 Our proposed DBN model for congestion

- By evacuation, we mean getting to compartment E.
- We assume there are 1,000 passengers and 35 crews onboard.

Figure 2 represents the proposed DBN model for congestion. Herein, the oval shapes show the nodes and the arcs determine causal relationships between the nodes. The double-ovals called deterministic nodes, represent either constant values or values that are algebraically determined from the states of their parents. The states of the nodes are provided in Table 1. The rectangular shapes are called sub-models that have a group of nodes inside. Figure 3 illustrates the sub-model "Fire" with its nodes and states. Table 1 lists detailed information on cause and effect relationship of the nodes presented in Fig. 2.

The initial probabilities and the CPTs are defined for the entire network. The CPTs are obtained according to the data in the literature, related databases, authors' reasoning. In future the input data to this DBN model can also be collected from the sensors installed on the ship and on the passengers' smartphones, which is our main goal of this research. The simulation results are presented and discussed in the next section.

4 Results

We are able to run the simulation for variety of scenarios in our model, but we have carefully selected the most realistic and critical scenarios. In order to have a smooth simulation, AIS sampling algorithm has been used in GeNIe. The aim of these simulations which have been targeted in 3 different parts is a three-fold:

Table 1 The description of the DBN nodes and their states

Sub-models	Number of nodes	Nodes	Number of states	States of each node	Description
Fire	8	A/B/C D1/D2/D3 S1/S2	4	No Starting Developed Burnout	This sub-model consists of 8 nodes, and each node has 4 states. This sub-model can show how the probability of fire is dynamically changing through time steps
Congestion	8	A/B/C D1/D2/D3 S1/S2	2	Free Congested	This sub-model shows the probability of congestions in each of the compartment
Crowd	8	A/B/C D1/D2/D3 S1/S2	3	Empty Some Many	This sub-model shows the crowdedness of passengers in each compartment
Flow	24	In_A/Out_A/ Decrease_A/ In_B/Out_B/ Decrease_B/ In_C/Out_C/ Decrease_C/ In_D1/Out_D1 Decrease_D1/ In_D2/Out_D2 Decrease_D2/ In_D3/Out_D3 Decrease_D3/ In_S1/Out_S1/ Decrease_S1/ In_S2/Out_S2/ Decrease_S2	2	True False	This sub-model computes how the flow of passengers moves among different compartments

(continued)

Table 1 (continued)

Sub-models	Number of nodes	Nodes	Number of states	States of each node	Description
		Extinguishing system	2	Reliable Unreliable	This node shows the status of the extinguishing system of the vessel, if they are working properly or not
		Trim and heel	2	Yes No	This node shows if the ship is vertically or horizontally imbalanced
		Rescue team existence	2	Yes No	This node represents if the passengers have the guidance of the rescue team members rather than being alone on their own during a disaster
		Pre-knowledge of rescue team	2	Yes No	This node depicts if the rescue team members have been trained properly
		Application knowledge	2	Yes No	"Application" is an emergency mobile app that is developed in the SmartRescue project. The node represents if the passenger is familiar with the app or not
		SmartRescue application	2	Working Not-working	This node indicates if the application is working on mobile phones or not
		Evacuation information	2	Yes No	This node explains if the passengers are receiving any information about the evacuation procedure using their smartphones which are equipped with SmartRescue application
		Decision support	2	Yes No	In this paper, decision support (DS) means if the rescue team members are doing their assigned tasks during emergencies properly, if they have been trained efficiently, if the SmartRescue app is working and the passengers have the sufficient knowledge about it
		Panic	2	Yes No	This node represents the passengers' psychological condition if they are panicking or not
		Passenger condition	2	Safe Injured	In this model, even though passengers can also have the state "dead", it is not considered due to the fact that a dead passenger cannot trap in congestion, so does not influence the evacuation time

(continued)

Table 1 (continued)

Sub-models	Number of nodes	Nodes	Number of states	States of each node	Description
		Passenger information	10	Female_under_ 30 Female_between_30&50 Female_above_ 50 F_Disabled_ under_50 F_Disabled_ above_50 Male_under_30 Male_between_ 30&50 Male_above_50 M_Disabled_ under_50 M_Disabled_ above_50	Based on IMO [1], the passengers in ships are divided into 10 different groups based on the gender, age and disability
		Ship knowledge	2	Yes No	This node indicates if the passengers have any pre-knowledge about the structure of the ship
		Evacuation time	2	Exceeded Standard	This node denotes the duration of evacuation process, if it is on time or exceeding the expected duration

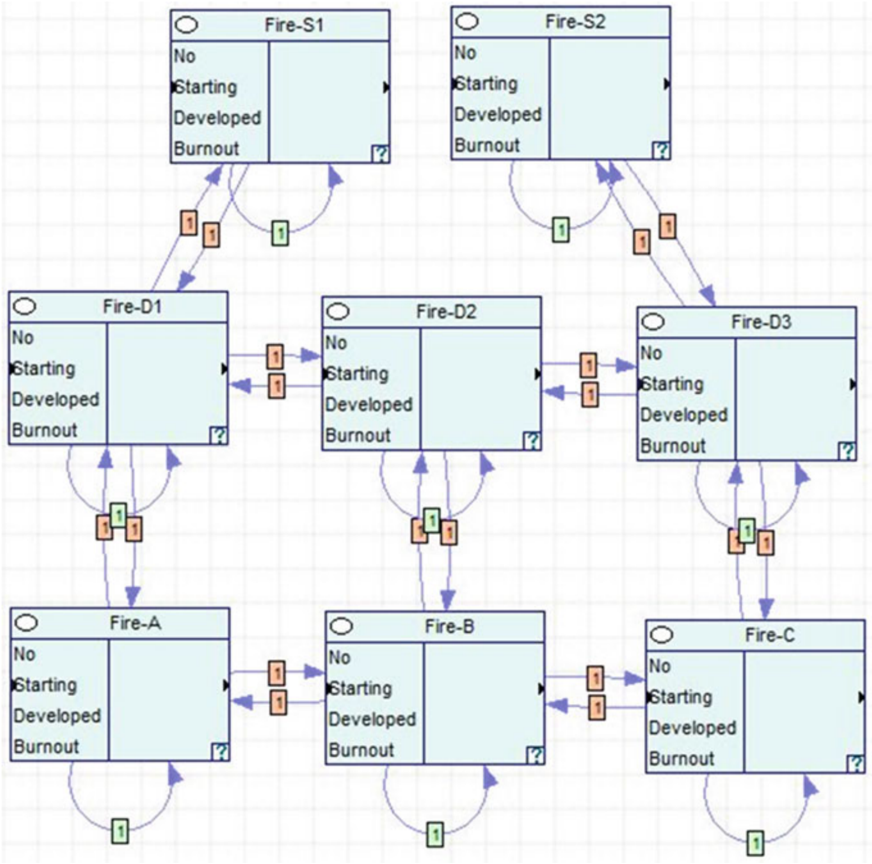


Fig. 3 A view of inside sub-model "Fire", with 8 nodes and 4 states for each node

- To observe the impact DS and ES in different time steps of the evacuation procedure on congestion, and to study the dynamicity of congestion in different time steps of the evacuation. (Part 1)
- To learn how the congestion in different compartments can impact the total evacuation time. (Part 2)
- To study how the probability of congestion can affect the probability of factors/ nodes DS and ES. (Part 3)

To gain these objectives, the assumptions made in Sect. 3 are applied. In addition, we have also studied worst-case scenarios, where all the evidences achieved from sensors, rescue team members, or smartphones present the worst situations. For instance, the status of "Panic" is "Yes", which means passengers are panicking, the ship is not stable, so the evidence entered for the node "Trim and Heel" is "Yes", and there are "Many" passengers in compartment A (node "A" in the sub-model "Crowd" has the status "Many"). In this section, the results are presented in three different parts and in each part several scenarios are simulated.

In part 1, we study the influence of factors DS and ES on the probability of congestion occurrence for compartment A and show how this probability changes dynamically through 15 time steps.

We ran the simulation for three different scenarios based on the assumptions above and some more assumptions. (a) We assumed that there are evidences available for congestion status in all compartments except compartment A that is our under-study compartment for congestion. (b) There is congestion in compartments B and C from time steps 1–3. (c) There is congestion in compartments D1, D2, and D3 from time steps 2–5. (d) There is congestion in compartments S1 and S2 from time steps 6–10. (e) As the focus of this study is on above mentioned factors, there will be no evidence available for the rest of the factors/nodes in our developed network, and their conditional probability distribution over time is calculated accordingly. The scenarios for this part are as follows.

1 Worst-case scenario: When there are no decision supports (DS) and no extinguishing system (ES) available.
2 Partial-case scenario: When there are no DS and ES at the beginning but after some time step 8 and 4, there will be DS and ES available respectively. That means, we assumed:

- From time step 0–8 there is no DS and then from time step 9–14 we have DS.
- From time step 0–4, there is no ES and later on we have ES.

3 Best-case scenario: When there are both DS and ES.

Figure 4 illustrates how the probability of congestion occurrence changes in 15 time steps in compartment A. As can be seen, in the worst-case scenario (indicated as "Worst" in this figure), the probability of congestion occurrence is 0 % at the beginning, then within 3 time steps this probability grows rapidly to 100 % and it remains at the pick till the end of 15 time steps. This represents the fact that evacuation from compartment A cannot be done successfully in the span of the standard evacuation time for this case.

In the second scenario, called partial-case (indicated as "Partial" in this figure), the trend for probability of congestion is exactly similar to the worst-case one from time step 0–7, but by having the ES active and DS in hand later in the evacuation procedure, the probability of congestion occurrence falls intensely to about 27 % in time step 12. For the next two time steps, the trend is still slightly downward, that means the probability of having a smooth evacuation from compartment A is around 80 %.

In the best-case scenario (indicated as "best" in this figure), the probability of occurring congestion in compartment A increases sharply from 0 % in time step 0–100 % in time step 3 and it remains constant for 3 time steps. This is due to the fact that the flow of passengers has got to the bottlenecks, so facing congestion for

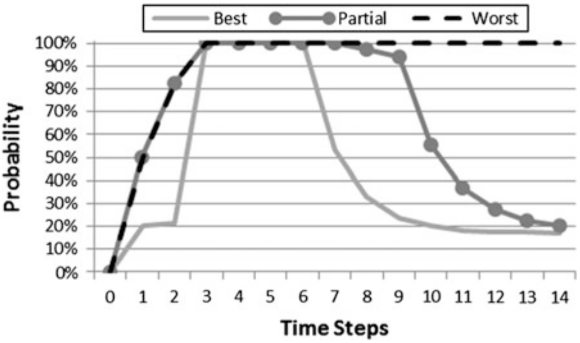


Fig. 4 The probability of congestion occurrence in compartment A

passengers in compartment A is inevitable even with the full support of crew members and smartphone application. After time step 6, the probability of congestion decreases significantly to 20 % after 4 time steps and this trend remains almost persistent to the end. The result of this scenario proves how vital DS are for the congestion and evacuation handling and shows DS decreases the probability of congestion significantly. As the best-case scenario is the optimum condition, it is expected to have 0 % chance of congestion at least in the last time steps, but the role of other factors on congestion presented in the network including size of the fire, trim and heel of the ship, passengers' panic, and etc. shouldn't be ignored. As based on our assumptions, the evidences for the aforementioned factors showed the worst status, thus they prevent the probability of congestion to turn to 0 %, which is completely realistic.

In part 2, we also compare three different scenarios to investigate the impact of congestion in different compartments on the total evacuation time in each time step. In the first scenario, we consider that stairs S1 and S2 are fully congested (indicated as S1 and S2 in the figure) and the rest of the compartments are free. In the second scenario, we assume compartments A, B, and C are congested (indicated as A and B and C in the figure) in all 15 time steps and the rest of the compartments are free. In the last scenario, we assume two ends of the corridor (indicated as D1 and D3 in the figure) are congested in the entire 15 time steps and the rest of the locations are free.

Figure 5 explains how influential congestion of different compartments is on the evacuation time. When S1 and S2 are congested, the probability that evacuation time will be extended reaches from 0 to 100 % in 8 time steps and it remains constant afterwards. On the other hand, for the second and third scenarios, the probability of the evacuation time to exceed standards climbs to 70 and 60 % respectively. These results prove that stairs are the most critical compartments, they are considered as possible bottlenecks. Interestingly, based on our model, compartments A, B, and C are more critical than the corridor when it comes to the evacuation time.

In part 3, another interesting experiment is done to compute the expected probability of existence of DS and reliability of ES, when the status of congestion in compartment A is "Free" in the entire evacuation time (15 time steps). In other

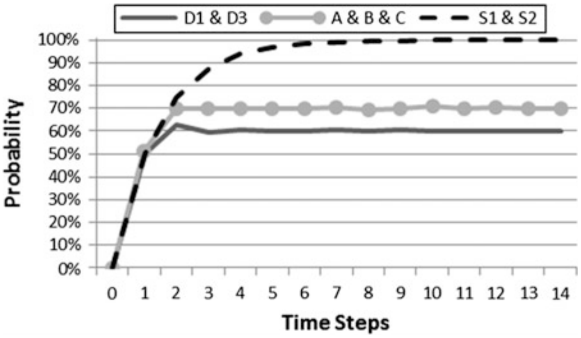


Fig. 5 The probability of the extension of the evacuation time for different scenarios

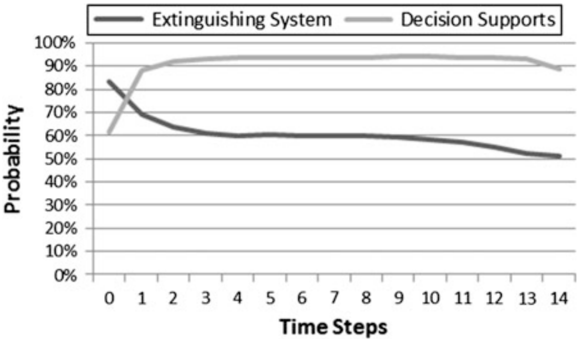


Fig. 6 The probability of existence of DS and reliability of ES when compartment A is "Free"

words, we want to know what should be the minimum probability of DS and ES in order to avoid congestion in compartment A in each time step. So we assume, compartment A is free in all 15 time steps and then GeNIe calculates the probability of DS and ES using Bayes's rule. The results are shown in Fig. 6.

5 Conclusion and Future Work

This paper has presented a DBN model for analyzing congestion during evacuating a ship burning in fire. The proposed model has considered the most significant factors affecting congestion occurrence. We have studied the real-time variations of the congestion and its impact on the evacuation time. It has been shown that the existence of DS can reduce the probability of occurring congestions, considerably. This reduction is up to about 80 % for a fully available DS. The results of this paper assist rescuers to make more accurate decisions. Further investigations towards studying the usage of smartphone sensors and their impact on congestion in bottlenecks can be carried out in future works.

Acknowledgments The research reported here has been partially funded by the research grant awarded to the SmartRescue project by Aust-Agder Utvikling- og Kompetansefond.

References

1. IMO, Interim guidelines for evacuation analyses for new and existing passenger ships, MSC/Circ. 1033 (2007)
2. S. Gwynne, E. Galea, C. Lyster, I. Glen, Analysing the evacuation procedures employed on a Thames passenger boat using the maritime EXODUS evacuation model. Fire Technol. **39**, 225–246 (2003)
3. D. Helbing, I. Farkas, T. Vicsek, Simulating dynamical features of escape panic. Nature **407**, 487–490 (2000)
4. H. Klüpfel, T. Meyer-König, J. Wahle, M, Schreckenberg, Microscopic simulation of evacuation processes on passenger ships. in *ACRI*, pp. 63–71 (2000)
5. A. Kirchner, A. Schadschneider, Simulation of evacuation processes using a bionics-inspired cellular automaton model for pedestrian dynamics. Phys. A **312**, 260–276 (2002)
6. A. Ferscha, K. Zia, On the efficiency of lifebelt based crowd evacuation, in *Proceedings of the 2009 13th IEEE/ACM International Symposium on Distributed Simulation and Real Time Applications*, pp. 13–20 (2009)
7. V. Ha, G. Lykotrafitis, Agent-based modeling of a multi-room multi-floor building emergency evacuation. Phys. A **391**, 2740–2751 (2012)
8. A. Fujihara, H. Miwa, Effect of traffic volume in real-time disaster evacuation guidance using opportunistic communications, in *2012 4th International Conference on Intelligent Networking and Collaborative Systems (INCoS)*, pp. 457–462 (2012)
9. P. Sarshar, J. Radianti, O.-C. Granmo, J.J. Gonzalez, A dynamic Bayesian network model for predicting congestion during a ship fire evacuation, in *Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering and Computer Science*, WCECS 2013, San Francisco, USA, pp. 29–34, 23–25 Oct 2013
10. J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference* (Morgan Kaufmann, San Mateo, 1988)
11. J.-L. Molina, D. Pulido-Velázquez, J.L. García-Aróstegui, M. Pulido-Velázquez. Dynamic Bayesian networks as a decision support tool for assessing climate change impacts on highly stressed groundwater systems. J. Hydrol. **479**, 113–129 (2013)
12. D.S. Laboratory (1998) GeNIe & SMILE (Online), http://genie.sis.pitt.edu/about.html#genie
13. P. Sarshar, J. Radianti, J.J. Gonzalez, Modeling panic in ship fire evacuation using dynamic Bayesian network, in *Innovative Computing Technology (INTECH), 2013 Third International Conference*, 2013, pp. 301–307

Chapter 20
An Intelligent Composition Algorithm for Automatic Thematic Music Generation from Extant Pieces

Abhijit Suprem and Manjit Ruprem

Abstract Recently, research on computer based music generation utilizing composition algorithm has drawn attention. The goal of this research is to produce new music, never heard before, using the algorithm developed and presented in this paper. The developed algorithm uses learning technique and probability and statistical analysis. The algorithm uses note sequences and other musical parameters such as note length, pitch, accidentals, modifications (intensity, speed), and note sequence repetition density for the preparation of a probability table that will generate new music. We used thematic music pieces (from same theme) as input music for analysis using learning followed by statistical analysis. We used MATLAB for analysis and MC Music editor for display. This research study is the first of its kind to create thematic music pieces effectively in a computer-based environment. The outcome of this research has a wide range of usage: waiting-music during automated phone-calls, background music in airports, airplanes, and restaurants, and so on. The work was extended to include variations of frequency and the shape of the note sequences for analysis.

Keywords Machine learning · MATLAB · Music generation · Music theory · Pattern recognition · Statistical analysis

1 Introduction

This chapter presents a new method for creating music in the composition algorithm domain. Due to the advent of electronics, software, and advanced computing methods, music creation is possible without using real instruments. In general such

A. Suprem (✉)
Department of Electrical and Computer Engineering, Lyles College of Engineering,
California State University, Fresno, CA, USA
e-mail: asuprem@mail.fresnostate.edu

M. Ruprem
Buchanan High School, Fresno, CA, USA

H. K. Kim et al. (eds.), *Transactions on Engineering Technologies*,
DOI: 10.1007/978-94-017-9115-1_20,

this approach to music generation is called algorithm-based music creation and utilizes such commercial software as Finale Notation Software and others. Music using these software systems still has to be manually written. We have developed a learning followed by probability and statistics based algorithm to create music that mimics existing styles of the composers of some specified input by reiterating the thematic elements and generating new music based on input music pieces. The algorithm incorporates some basic rules of music theory and through the learning method, learns other rules in order to create its own music.

2 Literature Review

The primary composition algorithm approaches are fractals, stochastic and L-systems [1]. These approaches incorporate various techniques. The developed techniques can be categorizes as knowledge-based, grammar, evolutionary, and learning techniques. There are various mathematical models used for each of the approaches for music generation, ranging from the complex methods for fractals, stochastic, and L-systems, and Markov models to more simplistic methods such as probability tables and statistical models [2]. Stochastic learning technique is implemented in the area of algorithmic composition in order to create new music from a set of random, existing music pieces. This research utilizes the principle of learning technique to analyze the existing musing to eventually generate thematic new music pieces. The algorithm has been tested and validated. In the sequel we have briefly review fractal, Markov and L-system.

2.1 Fractals

Music composition with fractals follows the notion that music is repetitive at various perspectives. Taking the example of Mozart's Symphony No. 40 from Leach and Fitch's paper on fractal-based music composition (Fig. 1), the repetitions can be represented and analyzed using fractal mathematics (Fig. 2) [3]. This can aid in not only discerning patterns but also in creating a mathematical model for music that corresponds to real-world music theory and its constraints.

2.2 Markov Model

Markov chain modeling can also be utilized for music generation. Markov modeling is a stochastic method for analyzing and learning from environmental data [4]. The environment in such a case, i.e. music generation, is a database of music pieces in a common format. This method analyzes existing data and generates an

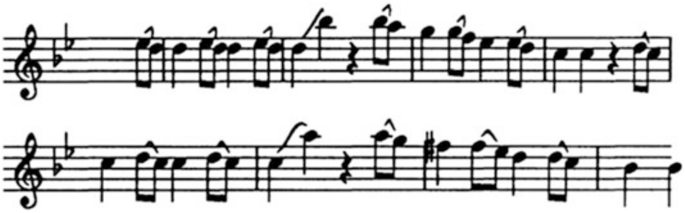


Fig. 1 Example of music input [3]

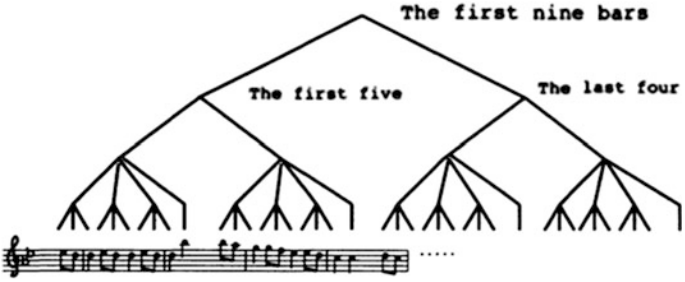


Fig. 2 Fractal based music [3]

optimal matrix of actions for each event in order to maximize the highest reward, which is a fixed number assigned to the model upon generation of structured music [5]. Given a larger database, the model can achieve convergence towards the optimal matrix faster. Under Markov chain modeling exist two subcategories: controlled and autonomous learning. Controlled learning involves user supervision in order to guide the system's learning. Autonomous learning is also known as unsupervised learning as the agent's only information comes from the input, which is sheet music in this work.

2.3 L-Systems

An L-system is a useful method of qualifying music and generating new music based on given rules. An L-system is a fractal generator that obeys grammatical rules [6]. Music is inherently structured as a language with syntax rules that must be followed. However, as music generation is largely a creative process requiring some exploration (i.e. deviation from extant pieces), using L-systems can be a setback [7].

3 Music Generation via Pattern Recognition and Statistical Analysis

Creation of new music (music that has not been heard before) using computer algorithm is a new research area. Computer creates new music either from scratch or using old music pieces from the same theme (classical, rock, instrumental, slow-speed, high-speed, hip-hop, baroque, etc.). The latter approach draws more attention because a listener likes to enjoy to listen to music of the same theme that s/he has been acquainted with. The objective of this experiment is to create a program that can create unique music pieces using pattern recognition and statistical analysis.

Patterns are unique sequences, which can be classified and clustered. To create a pattern recognition software, knowledge is needed about the item being identified. In this case, the algorithm needs to identify the notes, their length, and shape. Pattern recognition itself is the study of how machines can observe environment and extract all the repeating processes, learn to distinguish unique patterns, and make decisions based on the sequence of patterns [8].

Statistical analysis is a means to analyze probability of patterns. A statistical approach involves creating a data table for use later. Such a probability table is created with the following procedure:

- Identification of current index in dataset
- Defining and developing relationship between current and previous indices
- Recording of the relationships

With this process, any dataset can be defined with relationships between contents of the dataset. These relationships are recorded and used during music generation.

4 Methodology

To develop such an algorithm, appropriate integrated development environment (IDE) is necessary. As much of the analysis is around sequences of notes an IDE that can deal with arrays and matrices will work well. As such MATLAB was used in this research because MATLAB has been designed with matrix manipulation in mind. MATLAB follows BASIC language syntax, and has command toolboxes for specific fields. For this study, only the basic commands and the matrix toolbox were required. MATLAB script language was learned from "Numerical Methods with MATLAB" by Recktenwald [9]. A software tool was necessary to convert sheet music to a code format. The standardized musical notation format ABC was used. The ABC specification is listed at www.norbeck.nu/abc/abcbnf.txt. An ABC notation decoding software, MC Music Editor was used to convert sheet music to notation format and to play back output new music pieces.

5 Machine Learning and Analysis Technique

The various techniques that we used in this research are systematically outlined below.

5.1 *Supervised Machine Learning*

Supervised machine learning is a form of artificial intelligence that deals with pattern recognition based on a training input [10]. The agent (software system) is given data that contains patterns. Identification of correct patterns in the data (the correct patterns are known to the trainer) leads to a numerical reward to the agent. Incorrect pattern identification leads to a negative reward. The agent is programmed to follow positive rewards and the methods the program uses to correctly identify patterns in data are given more weight. With repeated training, the agent learns to use the correct method to identify patterns [11].

5.2 *Reinforcement Learning*

A reinforcement learning algorithm learns the optimal policy in an environment by choosing actions with highest future rewards [12]. Of particular interest is the Q-learning algorithm. The algorithm has three components: (i) and agent, which learns the environment, (ii) a dynamic or static environment made of states, where various actions can be completed in order to achieve a predetermined objective, with rewards for achieving the objective, and (iii) a goal state for the agent to reach. The environment is modeled for the agent with a matrix known as the R-Matrix, with dimensions M, X, and N, where M is number of states and N is number of actions per state. Each element of the matrix is defined as a state-action pair, and each state action pair has a reward associated with it. Generally, all possible state-action pairs are given a zero reward, all impossible state-action pairs are given negative reward, and the goal state-action pair is given the highest reward. The agent uses the R-Matrix to build the Q-Matrix, which is a model of the shortest path from any state to the goal state. The Q-Matrix has the same dimensions as the R-Matrix. Each state action-pair in the Q-Matrix has a reward value used for choosing optimal learning mechanism. The model for Q-Learning is as follows:

$$Q(state, action) = R(state, action) + (\gamma * \mathrm{Max}(Q(next\ state, all\ actions)))$$

where,

- Q(state, action) is state-action pair for the particular action the agent has chosen
- R(state, action) is the reward currently assigned to the state-action pair in the R-matrix

- γ, a value from 0 to 1, is the agent's consideration for future rewards and the reduction factor for rewards
- Max[Q(next state, all actions)] is the maximum rewards possible in the next state

The learning rate defines which actions are chosen. A higher learning rate leads to less exploration and more exploitation, i.e. the agent will choose actions that lead to higher rewards, and vice versa [13].

5.3 Statistical Modeling

Statistical modeling can be utilized in music generation as a probability-based prediction method. Such modeling has two phases: information retrieval and prediction.

5.3.1 Information Retrieval

Music can be stored in various standardized formats to ease the retrieval process. Currently, there are four standards for music storage: (i) Humdrum, (ii) ABC notation, (iii) MusicXML, (iv) Humdrum, and (v) Portable Document Format (PDF).

- Humdrum: Humdrum is an older format for music storage. Each line contains a note and its duration. Multiple staffs are represented by tab delimiters on each line.
- ABC notation: The ABC notation format is a highly simplistic representation of music that can depict various musical elements and events such as ties, triplets, tempo, and volume. However, the notation is less verbose than required and is not as standardized as other formats. Therefore, it can be difficult to set up input data. ABC notation was used in a preliminary test of the algorithm. The absence of more advanced elements and events makes the notation format unfit for the algorithm.
- Portable Document Format: The Portable Document Format (PDF) is one of the most widespread music notation formats. Majority of sheet music can be found in PDF format. The PDF format, however, stores music graphical and current Optical Music Recognition (OMR) techniques are not advanced to properly characterize and translate graphical music to a text form.
- MusicXML: MusicXML (Music Extensible Markup Language) is a more widespread format for music representation. Currently, there exist many music pieces in the MusicXML format. Further, there are various APIs in different languages for efficient data retrieval from XML documents by DOM (Document Object Model) traversal. The MusicXML format was used for this research. As MusicXML 3.0 can represent various notation symbols and musical structures, it is a flexible choice for this research.

5.3.2 Prediction

The prediction phase involves using retrieved data to build a knowledgebase for generating future music. As noted, there are machine learning methods, Markov chains, and fractal-based methods that can be used in this phase. For the purposes of this research, machine learning methods were used. The pseudocode for the prediction phase is as follows:

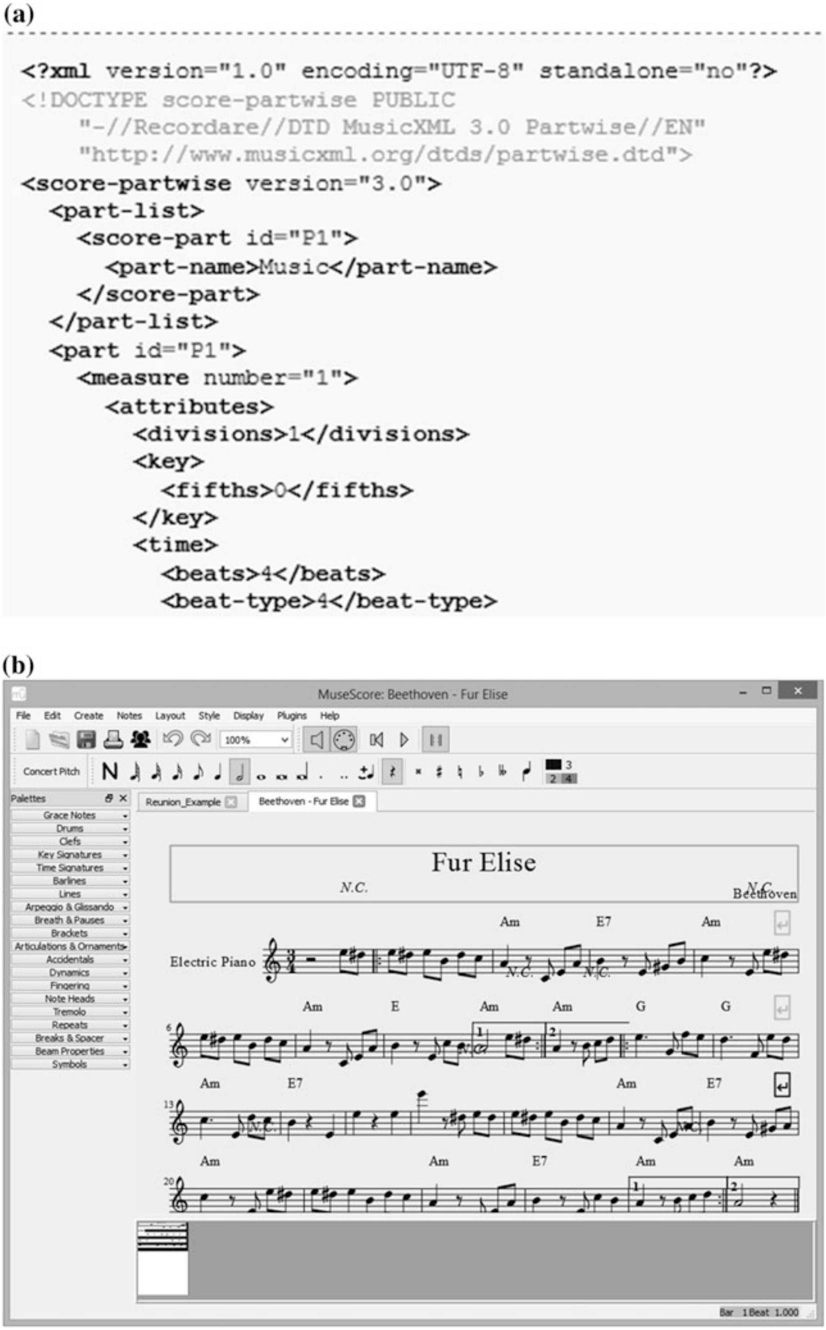
```
a. Collect music and characterize into measures
b. Define each measure based on pitch
c. Determine patterns in music
    a. Rising, falling, peak, trough patterns are characterized
    b. Recurring undefined patterns are stored as reference—these are music
       theory components the system has learned
d. Determine probability of each pattern type
e. Create probability table of pattern types and successive probabilities.
```

6 Implementation

Learning algorithms have pervaded many commercial systems from speech recognition, image processing, and mobile robot navigation to conversation dynamics and data mining [1]. Such systems have become commonplace in today's technological environment, and every day, new techniques are being developed to use learning algorithms in various applications. This paper presents a novel method for music generation based on machine learning method. Music generation has been traditionally been restricted to only probability and statistical models [2]. There have been some research on using artificial intelligence in music generation [3]. The methods used in this research are new as they analyze input data statistically rather than on a note-by-note basis.

6.1 Music Database

Before implementing the analysis algorithm in MATLAB script language, several music pieces were imported using MusixXML. Some music pieces did not have corresponding MusicXML files; however, they had ABC notation formats. These files were used in conjunction with other MusicXML files. Figure 3 shows the XML schema for Fur Elise, as well as the accompanying digital sheet music. Figure 4 shows ABC notation for A Song for Adra (an example) and the accompanying sheet music as a PDF file.



```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<!DOCTYPE score-partwise PUBLIC
    "-//Recordare//DTD MusicXML 3.0 Partwise//EN"
    "http://www.musicxml.org/dtds/partwise.dtd">
<score-partwise version="3.0">
  <part-list>
    <score-part id="P1">
      <part-name>Music</part-name>
    </score-part>
  </part-list>
  <part id="P1">
    <measure number="1">
      <attributes>
        <divisions>1</divisions>
        <key>
          <fifths>0</fifths>
        </key>
        <time>
          <beats>4</beats>
          <beat-type>4</beat-type>
```

Fig. 3 XML Music Format Input for MATLAB. **a** Code mode. **b** Graphical mode

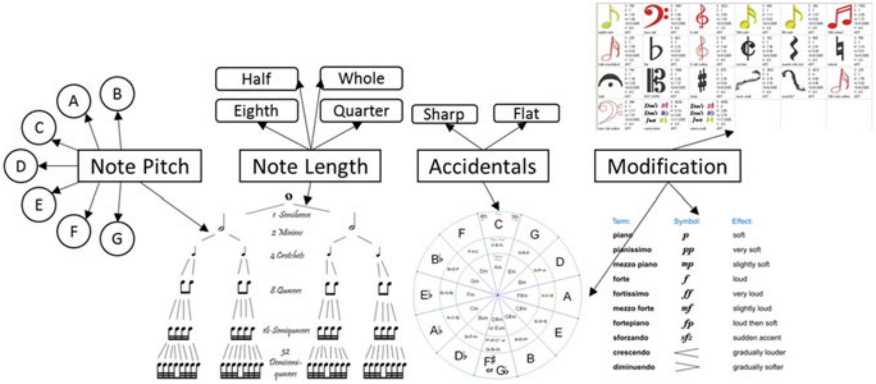


Fig. 4 Database format for music pieces for each note; Note Pitch is defined as standard music notes; Note Length is standard durations for notes; Accidentals refer to half-step increments or decrements in frequency; Modifications refer to musical embellishments

6.2 MATLAB Script Algorithm

The MATLAB script accomplishes three goals: (i) import music pieces and create music database, (ii) analyze patterns in music database using machine learning and statistical analysis and (iii) generate new music based on analysis.

6.2.1 Music Database Creation

The MATLAB language includes various APIs for XML schema parsing [14]. These were used to read MusicXML files and convert the content to readable database format (Fig. 4). A parser was written for files in ABC format. Both parsers (XML and ABC) were modified to create the same output format for music database. The database stores the following for each note: (i) note identifier of the current note (ii) the note identifier of the previous note, (iii) current sequence identifier.

6.2.2 Learning and Analysis Phase

The database is analyzed using statistical methods. As each thematic piece (i.e. rock, classical, instrumental, etc.) has varying signal (sound) morphology, statistical analysis yields a generalized rule-based matrix of note and sequence identifiers for creating new pieces (Fig. 6).

Each sequence has a high probability of some sequences appearing after it, and a lower probability of other sequences. The rule-based matrix contains these probabilities. During music generation, this matrix is used as the source or environment

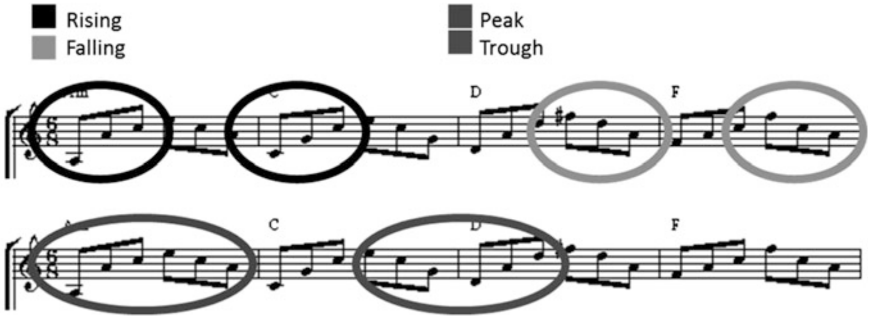


Fig. 5 Defined sequence types for statistical analysis

matrix. The rewards are predetermined to reduce operational time for achieving convergence.

- Note identifier: The note identifier is an identification code that can be used to determine note attributes such as pitch, length, accidentals, or any modifications such as arpeggios, allegros, pianissimo, etc., The note identifier is used to characterize and categorize.
- Sequence identifier: The sequence identifier is an identification code for a sequence of notes. The sequence may be a rising, falling, peaked, or trough style (Fig. 5). The sequence identifier is used to detect patterns within a section of a music piece. The sequence identifier is also used to characterize both input and output music.

6.2.3 Music Generation Phase

The matrix created during statistical analysis is used to generate new music. An initial note identifier and sequence identifier is chosen at random and successive note identifiers are added to the sequence based on the matrix. New sequences (chosen from the matrix) are appended upon completion of each current sequence. The sound morphology of the generated music is compared with source pieces to assign rewards. Higher rewards are assigned for similarities to source pieces. Rewards are recursively incorporated into the matrix by increasing probabilities of sequences applied to generated music. The process is repeated until generated music morphology closely matches source music morphology within predefined error bounds (Fig. 6).

7 Discussion

This research is first of its kind to characterize the music in a computer-based system in the sense that learning is integrated statistical analysis to produce thematic music pieces. This entails interdisciplinary knowledge base in the areas of

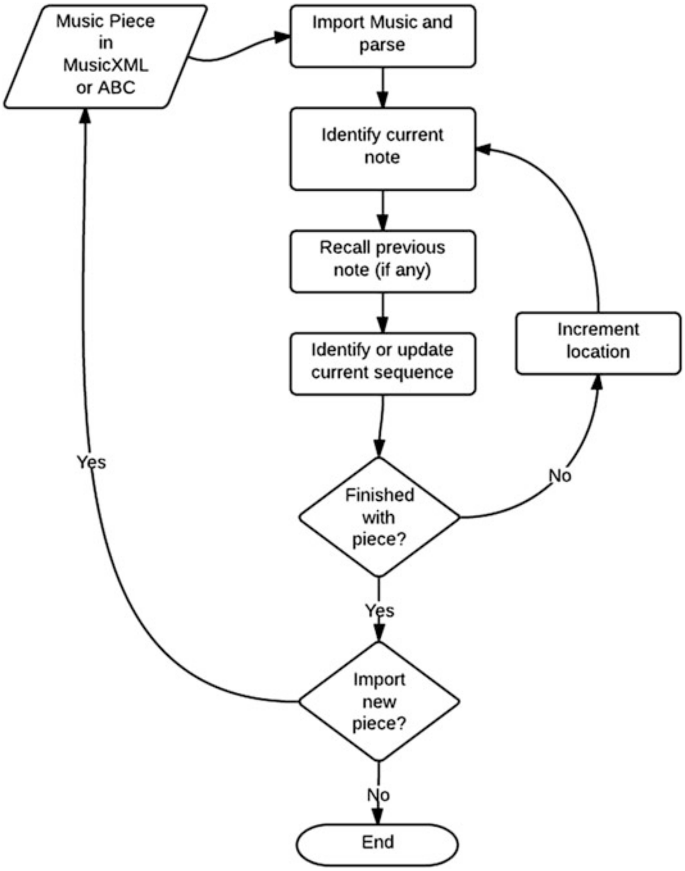
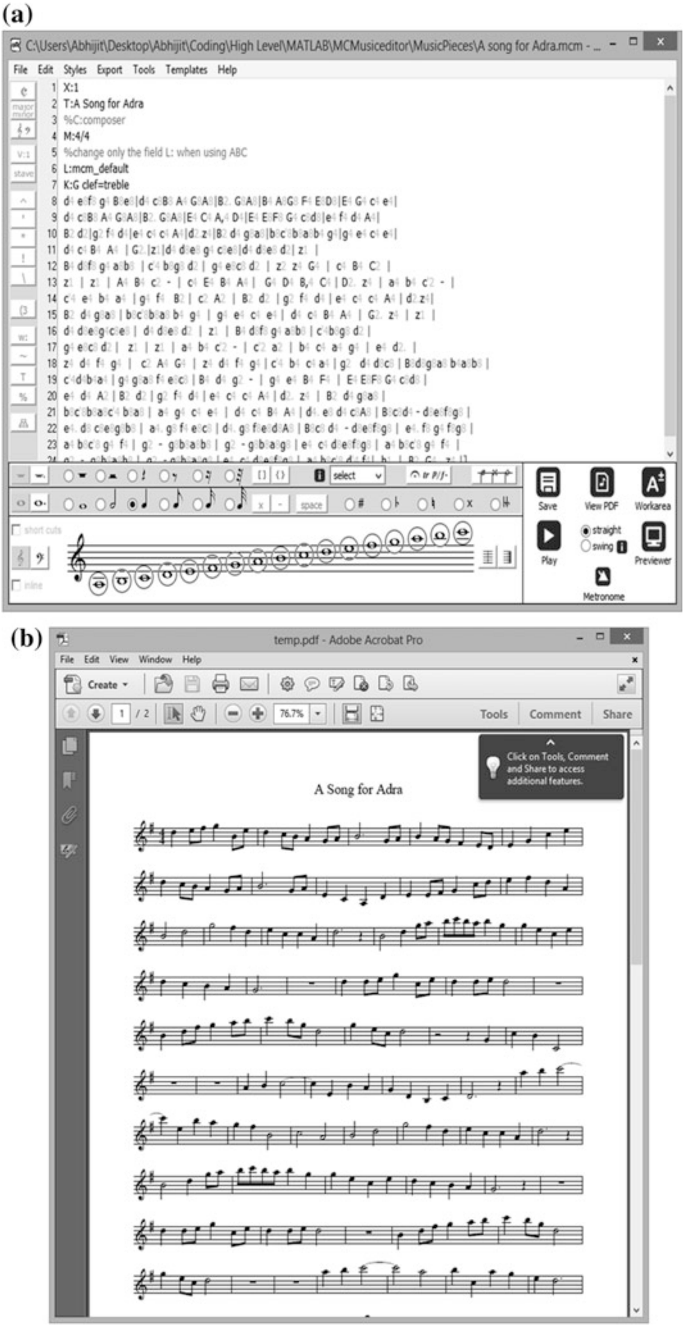


Fig. 6 Statistical analysis algorithm

music, music interpretation (technical viewpoint), programming, and data analysis and prediction. The results are encouraging to motivate the researchers to develop a complete computerized music infrastructure to generate new music from extant pieces.

The developed algorithm [15] can be used to generate any thematic music, provided the input music conforms to the theme. Music from multiple themes can distort the algorithm by contributing dissonant musical structures and harmonics. Consequently, the algorithm can work under single themes. The algorithm can itself be used as a filter for thematic music under an autonomous, unsupervised training period if it has already been implemented under a supervisory period and the probability tables have been stored. Thus, the existing pattern data can aid in identifying new input music under themes.


(a)
C:\Users\Abhijit\Desktop\Abhijit\Coding\High Level\MATLAB\MCMusiceditor\MusicPieces\A song for Adra.mcm - ...
File Edit Styles Export Tools Templates Help
major minor
V:1
stave
select
short cuts
inline
Save
View PDF
Workarea
Play
● straight
○ swing
Previewer
Metronome
(b)
temp.pdf - Adobe Acrobat Pro
File Edit View Window Help
Create
1 / 2
76.7%
Tools
Comment
Share
Click on Tools, Comment and Share to access additional features.
A Song for Adra

```
X:1
T:A Song for Adra
%C:composer
M:4/4
%change only the field L: when using ABC
L:mcm_default
K:G clef=treble
d4 e8f8 g4 B8e8|d4 c8B8 A4 G8A8|B2. G8A8|B4 A8G8 F4 E8D8|E4 G4 c4 e4|
d4 c8B8 A4 G8A8|B2. G8A8|E4 C4 A,4 D4|E4 E8F8 G4 c8d8|e4 f4 d4 A4|
B2 d2|g2 f4 d4|e4 c4 c4 A4|d2 z4|B2 d4 g8a8|b8c'8b8a8b4 g4|g4 e4 c4 e4|
d4 c4 B4 A4 | G2.|z1|d4 d8e8 g4 c8e8|d4 d8e8 d2| z1 |
B4 d8f8 g4 a8b8 | c'4 b8g8 d2 | g4 e8c8 d2 | z2 z4 G4 | c4 B4 C2 |
z1 | z1 | A4 B4 c2 - | c4 E4 B4 A4 | G4 D4 B,4 C4 | D2. z4 | a4 b4 c'2 - |
c'4 e4 b4 a4 | g4 f4 B2 | c2 A2 | B2 d2 | g2 f4 d4 | e4 c4 c4 A4 | d2 z4|
B2 d4 g8a8 | b8c'8b8a8 b4 g4 | g4 e4 c4 e4 | d4 c4 B4 A4 | G2. z4 | z1 |
d4 d8e8g4c8e8 | d4 d8e8 d2 | z1 | B4 d8f8 g4 a8b8 | c'4b8g8 d2 |
g4 e8c8 d2 | z1 | z1 | a4 b4 c'2 - | c'2 a2 | b4 c'4 a4 g4 | e4 d2. |
z4 d4 f4 g4 | c2 A4 G4 | z4 d4 f4 g4 | c'4 b4 c'4 a4 | g2 d4 d8c8 | B8d8g8a8 b4a8b8 |
c'4d4b4a4 | g4 g8a8 f4 e8c8 | B4 d4 g2 - | g4 e4 B4 F4 | E4 E8F8 G4 c8d8 |
e4 d4 A2 | B2 d2 | g2 f4 d4 | e4 c4 c4 A4 | d2. z4 | B2 d4 g8a8 |
b8c'8b8a8c'4 b8a8 | a4 g4 c4 e4 | d4 c4 B4 A4 | d4. e8 d4 c8A8 | B8c8d4 - d8e8f8g8 |
e4. d8 c8e8g8b8 | a4. g8 f4 e8c8 | d4. g8 f8e8d8A8 | B8c8 d4 - d8e8f8g8 | e4. f8 g4 f8g8 |
a4 b8c'8 g4 f4 | g2 - g8b8a8b8 | g2 - g8b8a8g8 | e4 c4 d8e8f8g8 | a4 b8c'8 g4 f4 |
```

Fig. 7 **a** Output music. **b** Output music (Sheet form)

8 Conclusion

This study is a step to understand the effect of existing parameters on the created music in a computerized music system. In this study, extant music pieces were used to create new pieces through machine learning and statistical analysis methods. Significant numbers of music parameters were used in the learning and analysis phase to generate output music (Fig. 7). The increased complexity of the algorithm versus traditional approaches where the generation method utilizes only note pitch and duration leads to longer training periods as the algorithm must, in the probability tables, consider learning weights for additional parameters such as embellishments and accidentals. This leads to a closer conformation of output music to input music as most musical elements are incorporated to the output.

Acknowledgment I would like to acknowledge Dr. Honora Chapman, Director, Smittcamp Family Honors College, for her advice and encouragement. I would also like to acknowledge Dr. agy Bengiamin, Chair of Electrical and Computer Engineering and Dr. Ram Nunna, Dean of Lyles College of Engineering for providing the computing facilities for this research. I also acknowledge my parents for allowing us to accomplish part of this research in a home computing system.

References

1. D. Plans, D. Morelli, Experience-driven procedural music generation for games. IEEE Trans. Comput. Intell. AI Games **4**(3), 192–198 (2012)
2. W. Schulze, B. van der Merwe, Music generation with Markov models. IEEE Multimedia **18**(3), 78–85 (2011)
3. J. Leach, J. Fitch, Nature, music, and algorithmic composition. Comput. Music J. **19**(2), 23–33 (1995)
4. J.A. Whittaker, M. Thomason, A Markov chain model for statistical software testing. IEEE Trans. Softw. Eng. **20**(10), 812–824 (1994)
5. Q. Yuting, J. Paisley, L. Carin, Music analysis using hidden Markov mixture models. IEEE Trans. Signal Process **55**(11), 5209–5224
6. J. Mishra, Classification of linear fractals through L-system. First Int. Conf. Emerg. Trends Eng. Tech. **1**(5), 16–18 (2008)
7. P. Meyer, The fractal dimension of music. Senior Thesis, Columbia University (1993)
8. A. Jain, R. Duin, M. Jianchang, Statistical pattern recognition: a review. IEEE Trans. Pattern Anal. Mach. Intell. **22**(1), 4–37 (2000)
9. G. Recktenwald, *Numerical Methods with MATLAB: Implementations and Applications* (Prentice Hall, Upper Saddle River, 2000)
10. V. Shen, C. Yue-Shan, T. Juang, Supervised and unsupervised learning by using Petri nets. IEEE Trans. Syst. Man Cybern. Part A Syst. Hum. **40**(2), 363–375 (2010)
11. K. Dixon, C. Lippitt, J. Forsythe, Supervised machine learning for modeling human recognition of vehicle-driving situations. Int. Conf. Intell. Robots Syst. **2**(6), 604–609 (2005)
12. P. Kulkarni, *Introduction to Reinforcement and Systemic Machine Learning. Reinforcement and Systemic Machine Learning for Decision Making*. (IEEE, Piscataway, 2012), pp. 1–21
13. G. Maozu, L. Yang, J. Malec, A new Q-learning algorithm based on the metropolis criterion. IEEE Trans. Syst. Man Cybern. Part B: Cybern. **34**(5), 2140–2143 (2004)

14. XML Documents, MATLAB documentation center—data import and export, http://www.mathworks.com/help/matlab/ref/xmlread.html
15. A. Suprem, M. Ruprem, A new composition algorithm for automatic generation of thematic music from existing music pieces., Lecture Notes in Engineering and Computer Science, in *Proceedings of The World Congress on Engineering and Computer Science 2013, WCECS 2013*, USA, San Francisco, 23–25 Oct 2013, pp. 808–812

Chapter 21
On Combining Boosting with Rule-Induction for Automated Fruit Grading

Teo Susnjak, Andre Barczak and Napoleon Reyes

Abstract The automation of post-harvest fruit grading in the industry is a problem that is receiving considerable attention in the realm of computer vision and machine learning. Classification accuracy with automated systems in this domain is a challenge given the inherent variability in the visual appearance of fruit and its quality-determining features. While the accuracy of automated systems is of paramount importance, the usability and the interpretability of machine learning solutions to the operators are also crucial since many sophisticated algorithms involve numerous tunable parameters and are often "black-boxes". This research presents a generalizable machine learning solution that balances the need for high accuracy and usability by decomposing the problem into sub-tasks. A powerful boosting algorithm (AdaBoost.ECC) with low interpretability is employed for learning fruit-surface characteristics. The classification outputs of boosting then become inputs for rule-induction algorithms (RIPPER and FURIA), generating human-interpretable rule sets that are amenable for review and revisions by operators. Using seven datasets of different fruit varieties, the performance of the proposed method was compared against a manually calibrated commercial fruit-grading system. The results showed that the proposed system is able to match the accuracy of machines calibrated by domain experts having many years of experience, while providing simpler rule sets possessing high interpretability and usability while yielding knowledge discovery.

T. Susnjak (✉)
School of Engineering and Advanced Technology, Massey University,
Albany, New Zealand
e-mail: T.Susnjak@massey.ac.nz

A. Barczak · N. Reyes
Institute of Natural and Mathematical Sciences, Albany, New Zealand
e-mail: A.L.Barczak@massey.ac.nz

N. Reyes
e-mail: N.H.Reyes@massey.ac.nz

H. K. Kim et al. (eds.), *Transactions on Engineering Technologies*,
DOI: 10.1007/978-94-017-9115-1_21,

Keywords AdaBoost.ECC · Boosting · Classification · Decomposition strategies · Fruit grading · FURIA · Machine learning · RIPPER · Rule-induction

1 Introduction

The increasing demand for smarter and more effective solutions for sorting and grading of fresh produce in industry has lured considerable research activities combining computer vision (CV) and machine learning (ML) techniques. In particular, as evident from the literature [1–4], the automation of this industrial process is now considered as one of the most active application domains for integrated CV and ML solutions.

Automated inspection is important since it provides a more objective and thus consistent grading of fresh produce over manual inspection [1]. Fruit quality, for example, is commonly determined based on the extracted features representing the size, shape, color and the presence of blemishes and foreign materials [1]. The ability of manual inspection to deliver accurate grading diminishes with the increase in the number of factors that have to be considered [4], which raises the need for consistent and objective grading. Automatic grading is also more efficient since it increases the volume of produce that can be inspected, thus elevating productivity. Though the financial benefits of lowering the labor costs are certainly one driving factor towards automated grading, in some cases it is also crucial since certain fruit varieties are seasonal and grow in isolated regions where it is difficult to secure the required labor force.

Recent fruit grading research involving ML and CV include: [5] citrus fruit sorting based on rottenness caused by fungi by way of hyper-spectral imaging and the use of neural networks and CART ML algorithms, [6] blueberry sorting and detection of foreign materials using near infra-red (NIR) spectral imaging with linear discriminant analysis (LDA), [7] bruise detection on pears from hyper-spectral images using the Mahalanobis, Euclidean distance as well as maximum likelihood classification, [8] apple grading by classifying blemishes using the *k*-means clustering and quadratic discriminant analysis, [9] mango grading based on maturity levels using Gaussian mixture models (GMM), [10] investigations into raisin grading comparing SVMs, ANN, Bayesian Networks, decision trees and application of SVMs to automatic grading of Chinese jujube fruit [11]. Notable solutions using hyper-spectral imaging without ML techniques have been applied to detecting citrus canker lesions on grapefruit [12] with later modifications in [13] to accomplish the same task but in real-time using only two-band spectral imaging and pomegranate aril classification [14].

Despite the advantages associated with the automated sorting and grading of fruit using CV and ML techniques, two main obstacles deter its greater proliferation in the industry for the following reasons: (1) insufficient accuracy for certain fruit varieties [15, 16], and (2) the usability of the ML components in classification.

Naive CV techniques like segmentation and color thresholding are sometimes sufficient to accurately determine the quality of fruit as the recent survey [2] points out. In many cases though, the critical determinant of fruit quality is the presence of multiple types of blemishes, fruit features and foreign materials, which render simplistic approaches ineffective [8]. ML is perfectly suited to providing solutions for this task; however, given that the standards of what constitutes a certain type of a blemish and its degree of severity are non-uniform across different geographic locations and may even undergo re-adjustments within the same location from one crop to the next [15], the *one-size-fits-all* classifiers trained by ML experts off-site are not well suited and contribute to low grading accuracies. In addition, unforeseen environmental conditions produce defects like hail or frost damage for which the response in providing new classifiers must be immediate and therefore requires on-site re-training of the classifiers. The issue then becomes, which types of ML algorithms should be employed by non-ML experts for these types of problems? This question has to be carefully considered since using more complex algorithms are tempting; however, they usually come with more tunable parameters that need to be set appropriately, which makes them harder to use for the non-initiated and posses internals which are often more opaque [17].

In this work, a generalizable ML solution to the challenges of fruit grading is proposed and an appropriate strategy of applying different families of ML techniques to address the real-world industry requirements is demonstrated. The solution decomposes the classification task into multiple phases in order to address the problem of maintaining classification accuracy and adaptability, as well as the usability of ML and the interpretability of their outputs. A classification architecture is devised which employs a sophisticated boosting algorithm (AdaBoost.ECC [18]) for learning blemish and surface-type features at the initial layer. The outputs of this classifier subsequently act as inputs to the next classification layer, which are combined with the global fruit surface features like color, size and shape. The second layer is represented by a state-of-the-art rule-induction learning algorithm. Though rule-based algorithms are often not the most accurate inducers [19], they provide the advantages of high usability and interpretability of outputs [20]. For this, both the RIPPER [21] and FURIA [20] algorithms for generating the final fruit grading classifiers were experimented with.

The unique integration of complementary and as yet unexplored ML techniques for the fruit-grading problem domain comprises the novel contribution in this work.[1] The research here addresses the classification challenges in the industry, and responds to the calls [3] to discover new ways of combining ML techniques for this problem domain. The data sets used in the experiments were collected from a commercial fruit sorting machine. The capabilities of the proposed method are demonstrated by comparing the accuracies achieved by the commercial machine to sort seven fruit varieties using settings derived manually by a domain

[1] Preliminary results related to the current chapter were published at the Proceedings of The World Congress on Engineering and Computer Science 2013 [22].

expert, against the accuracies of the combination of boosting for surface feature classification and RIPPER and FURIA for the overall fruit grading.

To the authors best knowledge, in the literature, there are only two reports of boosting algorithms that were used in relation to the domain of fruit grading. [15] used a boosting algorithm called RealBoost [23] to perform a pixel-wise classification of potato surfaces in order to detect blemished regions, whereas [24] experimented using several ML algorithms, of which AdaBoost [23] is one of them to explore the accuracies of differentiating stems from calyxes on apples. Although both employed boosting algorithms on problem domains concerning fresh produce, neither addressed the overall grading aspect of each individual fruit. To the authors' best efforts, no instances of research or industrial applications of rule-based induction for the purpose of fruit grading has been uncovered.

The remainder of the paper is structured as follows: Sect. 2 describes the rationale behind the strategy of decomposing a given problem into multiple learning sub-tasks. Section 3 provides a brief overview of the proposed machine learning algorithms for usage on this type of a problem domain, while Sects. 4 and 5 present the methodology and the experimental results respectively, before the concluding remarks in the succeeding section.

2 Machine Learning Problem Decomposition

According to [25], *decomposition* generally describes the process of breaking down a given task or a system into smaller units. The idea is not new to machine learning and can be traced back to Samuel [26] in the 1960s with his decomposition approach application to the checkers playing programs. The motivation behind decomposition is to reduce a complex problem into more manageable sub-tasks, that can then be combined in order to solve the initial problem. The definition of such a goal-subgoal hierarchy can serve as a powerful and effective approach to reformulating a classification problem. Although the reduction in processing complexity might seem as a primary driver for employing this strategy, research indicates that decomposing a problem can also improve the classification accuracy of existing approaches [27]. Additional advantages inherent within the decomposition strategy are the increase in the comprehensibility of the original problem, the maintenance of simpler classification models as well as the flexibility that enables the usage of different types of inducers on each of the sub-problems [28].

The complexity of a learning task often refers to it comprising of high dimensionality (features) data. The challenge of performing machine learning in high dimensionality domains is a well understood problem. The principal difficulty arises from the fact that as the dimensionality (or the number of features) of a learning problem increase, a fixed-sized training dataset covers an ever decreasing fraction of the possible sample input space. With the growth in the sample dimensionality, the generalizability on such a domain becomes exponentially more difficult [17]. For example, even when presented with a trivial problem of learning a Boolean function B, where $B = \{0{,}1\}$ and dimensionality $d = 50$, the total number of samples representing the input space becomes as large as 2^{50}. If the problem domain lends itself, then one possible solution is to explicitly decompose the learning task into learning sub-task h_1 and h_2, each comprising of d_1 and d_2 dimensions where $d_1 + d_2 = d$. In this case the size of the total input space for learning a given Boolean function would be considerably reduced to $2^{d}{}_{1} + 2^{d}{}_{2}$.

In a domain where it is costly in terms of time resources to gather large datasets of samples, the importance of lowering the dimensionality of the learning problem becomes even more acute. The domain of fruit sorting is one such area, since each image sample must be carefully inspected and correctly labeled with the appropriate class. The learning problem for fruit classification in this case lends itself well to this form of sub-tasking which can be reformulated into a hierarchical decomposition, where the outputs of one sub-problem become the inputs to another. In this instance, features relevant to blemish classification are extracted and used only for the learning of the blemish classifiers, whose output becomes the new input feature for the induction of the global fruit grading classifier.

3 Proposed Two-Stage Classification Strategy

This research proposes decomposing the fruit grading problem into two classification tasks: (1) the training and classification of fruit surface features and blemishes, (2) the training and grading of the overall fruit based on the combination of the blemish-classification output and the general fruit color/appearance characteristics. The utilization of a more sophisticated ensemble-based inducer (AdaBoost.ECC seen in Algorithm 1) with low interpretability for learning the blemish classifier, while employing less powerful rule-induction algorithms (RIPPER and FURIA) to generate final fruit grading classification rules with high interpretability forms the heart of the proposed strategy.

Algorithm 1: AdaBoost.ECC

```
Given: $(x_1,y_1),...,(x_m,y_m)$ where $x_i \in X, y_i \in Y$ to make uniform over all incorrect labels
Output: Hypothesis $H_{final}(x) = \arg\max_{\ell \in Y} \sum_{t=1}^{T} g_t(x)\mu_t(\ell)$
Initialize $\tilde{D}_1(i,\ell) = [\![\ell \neq y_i]\!]/(m(k-1))$ where $m$ and $k$ are the number of samples and class
labels respectively and $[\![\pi]\!]$ evaluates to 1 if proposition $\pi$ holds, otherwise 0.
for $t = 1$ to $T$ do
    Compute coloring $\mu : Y \to \{-1,1\}$
    Let $U_t = \sum_{i=1}^{m} \sum_{\ell \in Y} \tilde{D}_t(i,\ell)[\![\mu_t(y_i) \neq \mu_t(\ell)]\!]$
    Let $D_i = \frac{1}{U_t} \cdot \sum_{\ell \in Y} \tilde{D}_t(i,\ell)[\![\mu_t(y_i) \neq \mu_t(\ell)]\!]$
    Train weak learner on examples $(x_1,\mu_t(y_1)),...,(x_m,\mu_t(y_m))$ weighted according to $D_t$
    Get weak hypothesis $h_t : X \to \{-1,1\}$
    Compute the weight of positive and negative votes $\alpha_t$ and $\beta_t$ respectively
    Define: $g_t(x) = \begin{cases} \alpha_t & \text{if } h_t(x) = 1 \\ \beta_t & \text{if } h_t(x) = -1 \end{cases}$
    Update $\tilde{D}_{t+1}(i,\ell) = \frac{1}{\tilde{Z}_t} \cdot \tilde{D}_t(i,\ell)\exp\{(g_t(x_i)\mu_t(\ell) - g_t(x_i)\mu_t(y_i)) \cdot \frac{1}{2}\}$
    where $\tilde{Z}_t$ is the normalization factor so that $\tilde{D}_{t+1}$ will sum to 1.
```

The combination of ensemble-based machine learning methods with boosting and weak underlying models, have recently experienced a widespread use due to their effectiveness at addressing many challenging classification problems. Following the success of the binary-class AdaBoost [29] algorithm, [18] proposed AdaBoost.ECC (error-correcting codes) in order to overcome the limitations of its predecessor and to thus extend boosting to multiclass scenarios. AdaBoost.ECC elegantly merges error correcting output coding (ECOC) principles with boosting. The algorithm repeatedly calls a weak learner (decision stump) on samples with variable weights, for a predetermined T rounds. A *coloring* function μ is defined which decomposes the multiclass problem into a binary one by re-labeling sample class-memberships. After each round, the *coloring* function μ then becomes the vehicle for iteratively generating the columns of the coding matrix which is used by ECOC methods for the resolution of predictions. An additional distribution $\tilde{D}$ is maintained to maximize the error correcting ability of each column in the coding matrix. The evaluation of final classifier H, on a sample x is computed as being the class label l, which receives the highest weighted vote from all class labels returned by each weak classifier $h_t(x)$.

Rule-based learning is one of the oldest and well studied paradigms within machine learning [19]. Its distinguishing feature is its high applicability to domains where the comprehensibility of the induced model is of prime importance, and where manual revision and adaptation of the induced models is necessary. RIPPER (*Repeated Incremental Pruning* to *Produce Error Reduction*) is a state-of-the-art algorithm in this genre, and recently FURIA (*Fuzzy Unordered Rule Induction Algorithm*) has been proposed as its extension and an improvement over the original. The idea here is to conduct two sets of experiments whereby the boosting algorithm is combined alternatively with RIPPER and FURIA for generating fruit grading rules.

RIPPER constructs rules in a greedy manner. The rules consist of conjunctions of predicates and a consequent part which designates a class to which the covered instances of that rule are assigned to. RIPPER learns rules one class label at a time, beginning with the smallest class in terms of the number of samples. Samples are removed from the training set incrementally with each subsequent antecedent that covers them. The training set is divided into a growing and a pruning set that signify two phases of the rule induction process. The growing phase specializes the rule by inducing and appending each new antecedent according to the *information gain* criterion. This is then followed by the pruning phase which removes the antecedents it considers to have overfitted the data according to its rule-value metric. Both the growing and the pruning phases are repeated until all the samples of the given class are covered or until the complexity of the rules exceeds the total description length metric. Following this, a sophisticated optimization phase is executed involving the re-running of the growing and pruning steps, and replacing existing antecedents with alternative and newly generated ones.

Whereas RIPPER produces hard and inflexible decision boundaries between different classes, FURIA proposes introducing a softer transition between class boundaries through *fuzzy rules*. It also departs from its predecessor by inducing rules for each class using the one-versus-all method which frees up the classifier from a strict order in which it must be evaluated. Arguably, this increases the comprehensibility as well as the knowledge discovery quality of its rules since they no longer implicitly embody the negated conditions of the previous rules [20]. This however introduces a problem during classification of unseen samples, where a sample may not satisfy any of the generated rules. FURIA addresses this by devising a rule stretching mechanism that generalizes the rules further to ensure a maximum coverage.

A more thorough exposition of the RIPPER and FURIA algorithms can be found in [20, 21] respectively. Meanwhile, examples of recent applications of FURIA can be found in [30, 31].

4 Methodology

The experiments comprised of seven datasets involving six fruit varieties. Table 1 outlines the details of each of the datasets.

The datasets themselves were obtained from packing houses from different locations around the world. The equipment and software used to capture the images and extract the features into datasets originated from a commercial fruit sorting equipment, manufactured by Compac Sorting Limited.[2] The key components of the equipment associated with the capture of the images are: (1) the conveyor belt, upon which reside the individual fruit cup holders which rotate the fruit on a single axis and making its entire surface visible, (2) the computer vision cabinet, which resides on top of the conveyor belt and contains the necessary lighting for multiple cameras (Fig. 1). The cameras are capable of capturing and synchronizing the rotating images from both the visible and infra-red spectra (Fig. 2a, b).

Each dataset was randomly split into halves representing the training and test datasets. Following best practices [32], the splits were stratified in order to ensure an equal proportion of samples from each class in both datasets. For each fruit variety, the images from the training dataset were used for calibrating the computer vision components. This entailed manually selecting pixels representing dominant hues in order to achieve color segmentation of regions signifying the quality of a given fruit, as well as the regions that identify the background (Fig. 2c). The software then classified the remaining pixels into the selected colors based on similarity measures. For each fruit variety, key blemish types that determine the grading quality of each specimen were identified. Thereupon, within the segmented fruit surface areas, further regions of interest (ROI) were manually identified as representing these surface features (Fig. 2d). Generically termed here

[2] Compac Sorting is one of the world industry leaders in automated fruit sorting with its headquarters in Auckland, New Zealand.

Table 1 Dataset characteristics

Fruit variety	Dataset samples			Fruit attributes		Blob attributes	
	Training set	Test set	Infra-red	Classes	Features	Classes	Features
Gala apples	78	79	yes	5	9	3	31
Plums	73	71	yes	3	8	5	31
Oranges	63	60	no	5	5	3	31
Navel-split oranges	61	62	no	3	5	5	31
Nectarines	296	291	yes	4	6	8	31
Peaches	133	107	yes	3	3	5	31
Pears	104	104	yes	5	7	3	31

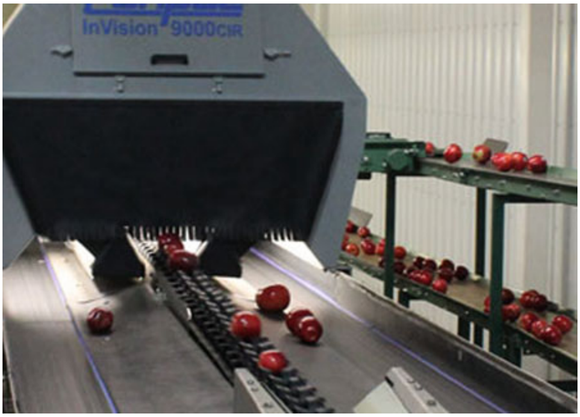


Fig. 1 Example of the commercial machine used to capture images and extract features for the training of classifiers

as *blobs*, and though they signify defects of varying type and severity, they can equally represent natural surface features like stems or calyxes depending on the fruit variety. These ROI were identified by manually selecting pixels that were most representative of both the blob features and normal fruit-surface areas which the software once again used to categorize the remaining pixels based on similarity measures. Finally, using the same segmentation technique from previous steps, the blob ROI were themselves segmented into dominant colors which would then serve as input features for the blob classifiers (Fig. 2e).

The blob feature vectors were then extracted into training datasets and manually labeled with the correct class membership. 10–20 blob samples were selected for each blob class. Using AdaBoost.ECC, the blob classifiers for each fruit variety were trained with the ensemble size set to 150. 5-fold cross-validation was used to generate blob classifiers on the training dataset in order to inspect the generalizability of the problem first. Provided that the selected blob dataset was of good quality, then the entire blob training set was used to generate the final classifier. These classifiers were subsequently applied to both the fruit training and test set images (Fig. 3) in order to extract blob feature vectors that would form the training

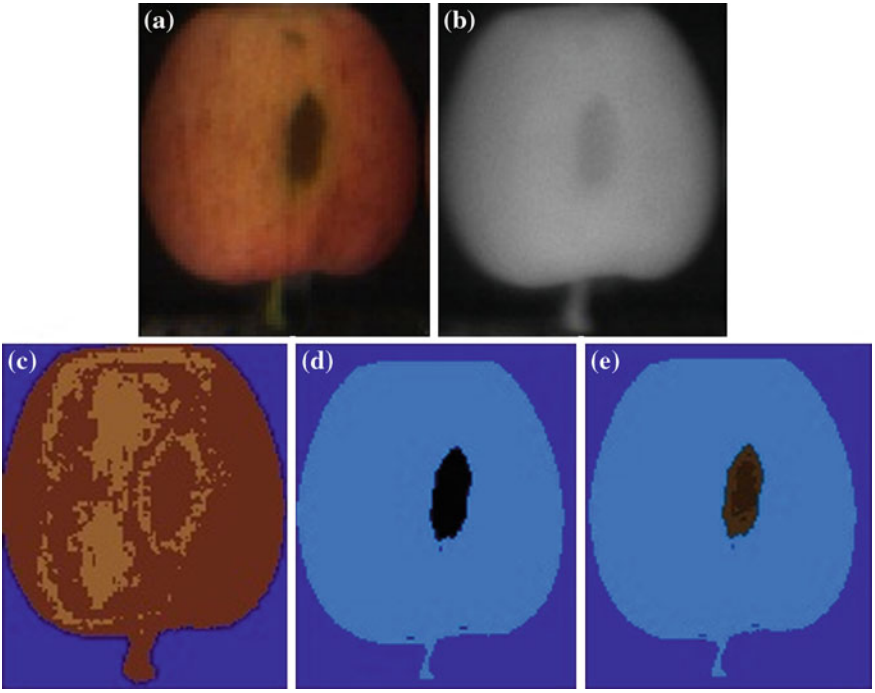


Fig. 2 Example of a gala apple with a *blemish*, imaged under the visible (**a**) and infra-red spectra (**b**). Example of the manual process of selecting dominant fruit colors and that of the conveyor belt in order to achieve segmentation (**c**). Identification of the *blemish* region within the segmented image (**d**) and the manual identification of the *dominant colors* representing the *blemish* region (**e**)

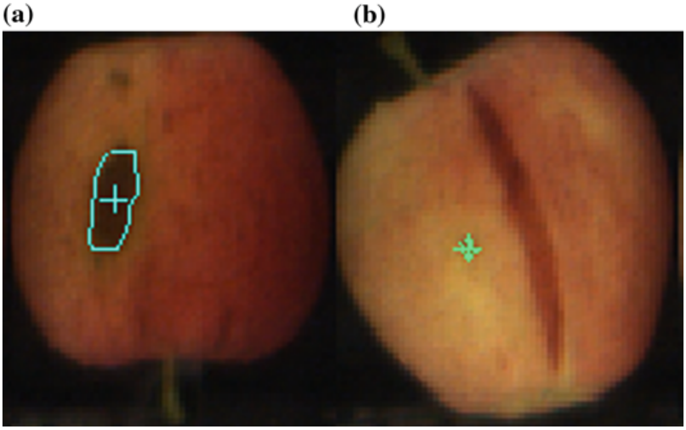


Fig. 3 Example of blob classification using an AdaBoost.ECC classifier on gala apples. Classification of a severe blemish type (**a**), mild blemish type (**b**)

and test sets of the rule-based classifiers. The blob feature vectors comprised (1) the sum of pixels representing the classified blob type (seen in Fig. 2d), (2) the sum of pixels for each of the colors that represent the particular blob class (seen in Fig. 2e). These feature vectors were combined with additional features extracted from the datasets representing global characteristics of each fruit (e.g. surface area covered by each of the selected hues in Fig. 2c).

The second phase involved training the rule-based classifiers on the training sets and testing the generalizability of the outputted classifiers on to the test datasets. The RIPPER and FURIA algorithms were used to train these classifiers. For each algorithm type, five classifiers with different random number seeds were trained. Average classification accuracy and geometric mean measures could then be calculated together with their standard deviations over the five runs. The accuracy of these classifiers was then compared to the accuracy attained by manually designed grading configurations by domain experts. The domain experts followed the same procedure of designing the grading configurations solely from the data available to them from the training datasets, whose generalizability was subsequently evaluated against the test datasets.

For the experiments, AdaBoost.ECC was implemented in C++, while the WEKA [32] machine learning toolkit was used for training RIPPER and FURIA classifiers.

5 Results

All seven datasets contained non-uniform class distributions, therefore both the total accuracy and the geometric mean as measures of generalizability were employed. Presenting only the accuracy has been shown to be inadequate and often misleading on class-imbalanced datasets [33], whereas the geometric mean can be a more meaningful measure of accuracy for biased class-distributions. Sun et al. [34] demonstrated how the geometric mean of recall values of each class *i* of a total of *k* classes can be applied to the multiclass scenario by being calculated as:

$$Geometric\ mean = \left(\prod_{i=1}^{k} Recall_i\right)^{\frac{1}{k}} \tag{1}$$

which yields a single value from 0-1 that presents a balanced performance of a classifier across all classes.

Table 2a shows the total accuracy across all datasets for each method as a percentage with the standard deviations. The proposed methods outperformed the manual strategy on five of the seven datasets. Likewise, RIPPER produced higher accuracy than FURIA on all except the Oranges and Peaches datasets. The accuracy measures in Table 2a are summarized in the form of average mean ranks. These figures indicate that RIPPER was the best performing algorithm on the selected problem sets, while the manual method was least successful.

Table 2 Generalization results for the three methods across all datasets

(a) Accuracy.

Fruit Variety	Manual	RIPPER	FURIA
Gala apples	81	80.5 ±0.7	74.9 ±0.6
Plums	52	54.4 ±4	53.5 ±3
Oranges	53	67 ±3	71 ±2
Navel-split oranges	50	56 ±2	55 ±2
Nectarines	59	59.1 ±1.4	58.4 ±5.6
Peaches	69	70.1 ±3.3	73.6 ±1.3
Pears	44	43.3 ±4.9	42.7 ±2.4
Mean ranks	2.3	1.6	2.1

(b) Geometric mean.

Fruit Variety	Manual	RIPPER	FURIA
Gala apples	0.58	0.58 ±0.004	0.54 ±0.01
Plums	0.35	0.40 ±0.004	0.34 ±0.05
Oranges	-	0.63 ±0.04	0.65 ±0.04
Navel-split oranges	0.64	0.30 ±0.03	0.30 ±0.03
Nectarines	0.39	0.19 ±0.17	0.28 ±0.16
Peaches	0.52	0.63 ±0.03	0.66 ±0.02
Pears	0.33	-	-
Mean ranks	1.8	1.9	2.1

The geometric mean measures are listed in Table 2b. From this perspective, the data indicated that the overall picture changed somewhat. The manual method emerged as the most accurate on four of the seven datasets, while RIPPER and FURIA won out on two datasets each. The average mean ranks reflected this shift by showing that the manual method edged out RIPPER for the best overall performance across all datasets, while FURIA trailed in the last position. The three methods experienced difficulties producing classifiers that produced non-zero hit rates on all classes within every dataset. This is denoted by a missing geometric mean entry on the Oranges and Pears datasets. The low geometric means by both RIPPER and FURIA on the Nectarines dataset, coupled with very high standard deviations may also be an indication that these algorithms experienced difficulties with low hit rates on some classes within this dataset; however, this may also be attributed to the presence of outliers in the limited sample sizes generated from five training runs.

The important outcome from the generalizability results is that they demonstrated the ability of the proposed methods to match and at times even to improve on the accuracy of manual calibration by domain experts that have undergone years of training to attain the required skill level. This must be emphasized since the generalization rates in these experiments for some fruit varieties may seem to be too low and thus ineffectual for commercial applications. However, this is not necessarily the case since it should be noted that fruit grading is subjective and not

Table 3 Example of the confusion matrices for all algorithms on the nectarines dataset from a selected training run

(a) Manual.

A	B	C	D	Classified as
1**37**	6	0	1	A
59	**17**	0	4	B
3	1	**5**	1	D
31	12	1	**13**	C

(b) RIPPER.

A	B	C	D	Classified as
126	12	4	2	A
55	**25**	0	0	B
5	3	**2**	0	D
28	11	2	**16**	C

(c) FURIA.

A	B	C	D	Classified as
130	14	0	0	A
53	**25**	0	2	B
5	3	**1**	1	D
23	12	0	**22**	C

an exact science. This means that some of the datasets in this corpus have different levels of noise in the form of mislabeled samples. Indeed, the consequences of misclassification in this domain are often not critical when an instance is classified one grade up or down from its true classification; albeit, upgrades tend to be more preferable. In that sense, it is mostly misclassifications into grades that are least correlated with the true classification, that pose the biggest concern. Table 3 is an example that demonstrates this in the form of confusion matrices for the Nectarines dataset. The table shows the detailed classification results at the level of each class for the three methods. The confusion matrices indicate that while the generalizability of all methods appeared to be low from Table 2a, b, the actual misclassifications mostly occurred in the proximity of each of the true classes. True positive detection rates were particularly problematic for class D where most of the samples were classified as instances of class A. On this class, both RIPPER and FURIA displayed superiority. Similarly, the manual method outperformed the other strategies on class D, which was also a difficult class to differentiate. While classes A and B contained a large percentage of misclassifications amongst them, the acuteness of these errors was of a lesser significance.

Although the mean ranks shown in Table 2a, b in themselves are informative, for completeness, the Iman-Davenport [35] non-parametric statistical test was carried out on both in order to determine if there existed significant differences amongst them. The null-hypothesis in this case states that there is no difference between the mean ranks and thus the various methods are identical. The critical value for the Iman-Davenport statistic F_F according to the F-distribution with $k-1$ and $(k-1)(N-1)$ degrees of freedom, yielded the value of $F_F(2,18) = 3.63$ at a $\alpha \leq 0.05$ significance level. The mean ranks for the F_F statistic for the Table 2a, b produced the values of 1.0 and 2.0 respectively. This forced the acceptance of the null-hypothesis at the specified level of significance and thus no further *post hoc* statistical tests were justifiable. From this it can be concluded that both the accuracy and geometric mean measures across the seven datasets indicated that the proposed decomposition strategy using the AdaBoost.ECC algorithm for learning surface features, whose output is combined with the FURIA and RIPPER algorithms for learning fruit grading rules is effective at matching the accuracy of manually calibrating grading thresholds by domain experts.

The final analysis considered the complexity and the interpretability of the generated rule sets. The example rules below are taken from the Peaches dataset. The manual method defined one rule for each of the three classes and a total of 13 antecedents:

```
IF (0 ≤ Brown ≤ 3) AND (0 ≤ BrownDefectColour ≤ 40) AND
   (0 ≤ NonDefectBlobColour ≤ 50) AND (0 ≤ ScarDefect ≤ 20)  THEN grade =  A
ELSE IF (0 ≤ Brown ≤ 10) AND (0 ≤ NonDefectBlobColour ≤ 300) THEN grade =  B
OTHERWISE grade = C
```

where *Brown* represent the surface pixel number of this color and *ScarDefect* signifies the size in pixels of a given classified defect. The *NonDefectBlobColour* and *BrownDefectColour* describe features whose primary usage is for the classification of *ScarDefect* instances but is also used as a feature for grading fruit. While this is an example of a succinct rule set for this dataset, it nonetheless presents a problem due to the dependence and high correlation between the features used for identifying surface defects and the defects themselves as they are both used for grading fruit as a whole which would be avoided in a decomposition approach.

Below is the example of rules that were generated by the RIPPER algorithm:

```
IF (Red ≤ 96.61) AND (62.26 ≤ ScarDefect ≤ 335.65)      THEN grade = C
ELSE IF (Red ≥ 98.97)                                  THEN grade = A
ELSE IF (ScarDefect ≤ 4.98) AND (Red ≥ 98.41)          THEN grade = A
OTHERWISE grade = B
```

In this example RIPPER induced four rules and 11 antecedents. While the RIPPER rule set is slightly more complex than the manual rules, in terms of interpretability it is arguably of higher quality since it relies only on two features to derive the rule set which cannot be features used for classifying defects.

The final example is that of the generation of FURIA rules. FURIA induces fuzzified rules and therefore relies on a trapezoidal function that determines the degree of membership for a given value. This function and its intervals are represented here using three values. The range between the positive or negative ∞ and its adjacent real number represents the values that are fully covered. The range between the two floating point numbers is the fuzzy interval that determines the degree of membership. In addition, each rule is associated with a confidence factor (CF). The class with the highest CF is assigned to a candidate sample. The FURIA rule set consists of a total of six rules and 11 antecedents.

```
CLASS WITH MAX SUPPORT:
(Red in [98.95, 98.97, ∞])                                           THEN grade = A (CF = 0.89)
(ScarDefect in [−∞, 0, 1.86]) AND (Red in [96.84, 97.19, ∞])        THEN grade = A (CF = 0.77)
(ScarDefect in [0, 1.86, ∞]) AND  (ScarDefect in [−∞, 61.58, 62.26])  THEN grade = B (CF = 0.87)
(ScarDefect in [0, 70.24, ∞]) AND (Red in [95.7, 95.86, ∞])         THEN grade = B (CF = 0.79)
(Red in [−∞, 98.23, 98.41]) AND (Red in [97.34, 97.55, ∞])          THEN grade = B (CF = 0.83)
(Red in [−∞, 95.22, 95.86]) AND (ScarDefect in [61.58, 62.26, ∞])     THEN grade = C (CF = 0.76)
```

Though the FURIA rule set is more verbose and thus more complex than the previous methods, its interpretability is arguably only minimally affected. Both the manually and the RIPPER generated rules embody within themselves the negated conditions of the previous rules. This means that in order to understand a given rule, it must be combined with the preceding rules. FURIA rule sets on the other hand are explicit and therefore more verbose, but in that they offer more in terms of knowledge discovery.

6 Conclusion

This research considered the problem of automated fruit sorting using machine learning. A novel strategy which decomposes a classification task into two phases for this problem domain was presented. The first phase applies a multiclass boosting algorithm (AdaBoost.ECC) to the sub-problem of classifying surface defects. The outputs of this classifier act as inputs to a rule-induction algorithm that generates interpretable rules. The experiments used the state-of-the-art RIPPER and FURIA algorithms. Most importantly, the results indicated that this approach matched the accuracy of rules devised manually by domain experts with many years of experience. The generated rules tended to be less complex than the manually derived rules, and are also more likely to contribute to knowledge discovery.

Acknowledgment The authors express their gratitude to Compac Sorting Ltd. for providing access to their datasets, grading maps and the necessary software for these experiments.

References

1. T. Brosnan, D.W. Sun, Improving quality inspection of food products by computer vision: a review. J. Food Eng. **61**(1), 3–16 (2004)
2. S. Cubero, N. Aleixos, E. Molto, J. Gómez-Sanchis, J. Blasco, Advances in machine vision applications for automatic inspection and quality evaluation of fruits and vegetables. Food Bioprocess. Tech. **4**(4), 487–504 (2011). doi:10.1007/s11947-010-0411-8
3. C.J. Du, D.W. Sun, Learning techniques used in computer vision for food quality evaluation: a review. J. Food Eng. **72**(1), 39–55 (2006)
4. D. Lorente, N. Aleixos, J. Gómez-Sanchis, S. Cubero, O. García-Navarrete, J. Blasco, Recent advances and applications of hyperspectral imaging for fruit and vegetable quality assessment. Food Bioprocess. Tech. **5**(4), 1121–1142 (2012)
5. J. Gómez-Sanchis, J.D. Martín-Guerrero, E. Soria-Olivas, M. Martínez-Sober, R. Magdalena-Benedito, J. Blasco, Detecting rottenness caused by penicillium genus fungi in citrus fruits using machine learning techniques. Expert Syst. Appl. **39**(1), 780–785 (2012)
6. T. Sugiyama, J. Sugiyama, M. Tsuta, K. Fujita, M. Shibata, M. Kokawa, T. Araki, H. Nabetani, Y. Sagara, Nir spectral imaging with discriminant analysis for detecting foreign materials among blueberries. J. Food Eng. **101**(3), 244–252 (2010)

7. J. Zhao, Q. Ouyang, Q. Chen, J. Wang, Detection of bruise on pear by hyperspectral imaging sensor with different classification algorithms. Sens. Lett. **8**(4), 570–576 (2010)
8. V. Leemans, M.F. Destain, A real-time grading method of apples based on features extracted from defects. J. Food Eng. **61**(1), 83–89 (2004)
9. C.S Nandi, B. Tudu, C. Koley, An automated machine vision based system for fruit sorting and grading, in *Proceedings of the 6th International Conference on Sensing Technology (ICST, IEEE)*, 2012, pp. 195–200
10. K. Mollazade, M. Omid, A. Arefi, Comparing data mining classifiers for grading raisins based on visual features. Comp. Electron. Agr. **84**, 124–131 (2012)
11. H. Zheng, H. Lu, Y. Zheng, H. Lou, C. Chen, Automatic sorting of Chinese jujube (zizyphus jujuba mill. cv. 'hongxing') using chlorophyll fluorescence and support vector machine. J. Food Eng. **101**(4), 402–408 (2010). doi:10.1016/j.jfoodeng.2010.07.028
12. J. Qin, T.F. Burks, M.A. Ritenour, W.G. Bonn, Detection of citrus canker using hyperspectral reflectance imaging with spectral information divergence. J. Food Eng. **93**(2), 183–191 (2009). doi:10.1016/j.jfoodeng.2009.01.014
13. J. Qin, T.F. Burks, X. Zhao, N. Niphadkar, M.A. Ritenour, Development of a two-band spectral imaging system for real-time citrus canker detection. J. Food Eng. **108**(1), 87–93 (2012). doi:10.1016/j.jfoodeng.2011.07.022
14. J. Blasco, S. Cubero, J. Gómez-Sanchís, P. Mira, E. Moltó, Development of a machine for the automatic sorting of pomegranate (*Punica Granatum*) arils based on computer vision. J. Food Eng. **90**(1), 27–34 (2009). doi:10.1016/j.jfoodeng.2008.05.035
15. M. Barnes, T. Duckett, G. Cielniak, G. Stroud, G. Harper, Visual detection of blemishes in potatoes using minimalist boosted classifiers. J. Food Eng. **98**(3), 339–346 (2010)
16. W. Huang, C. Zhang, B. Zhang, in *Identifying Apple Surface Defects Based on Gabor Features and SVM Using Machine Vision*, eds. by D. Li, Y. Chen. Computer and Computing Technologies in Agriculture V, IFIP Advances in Information and Communication Technology, vol. 370 (Springer, Berlin, 2012), pp. 343–350. doi:10.1007/978-3-642-27275-2_39
17. P. Domingos, A few useful things to know about machine learning. Commun. ACM **55**(10), 78–87 (2012). doi:10.1145/2347736.2347755
18. V. Guruswami, A. Sahai, Multiclass learning, boosting, and error-correcting codes, in *Proceedings of the 12th Annual Conference on Computational Learning Theory (COLT'99)* (ACM, New York, 1999), pp. 145–155
19. J. Fürnkranz, D. Gamberger, N. Lavrac, *Foundations of Rule Learning* (Springer, New York, 2012)
20. J. Hühn, E. Hüllermeier, Furia: an algorithm for unordered fuzzy rule induction. Data Min. Knowl. Disc. **19**(3), 293–319 (2009). doi:10.1007/s10618-009-0131-8
21. W. Cohen, Fast effective rule induction, in *Proceedings of the 12th International Conference on Machine Learning*, 1995, pp. 115–123
22. T. Susnjak, A. Barczak, N. Reyes, A decomposition machine-learning strategy for automated fruit grading, in *Proceedings of the World Congress on Engineering and Computer Science (WCECS 2013)*, (San Francisco, 2013), pp. 819–825
23. Y. Freund, R. Schapire, A decision-theoretic generalization of on-line learning and an application to boosting. J. Comput. Syst. Sci. **55**(1), 119–139 (1997)
24. D. Unay, B. Gosselin, Stem and calyx recognition on 'jonagold' apples by pattern recognition. J. Food Eng. **78**(2), 597–605 (2007)
25. O. Maimon, L. Rokach, *Decomposition methodology for knowledge discovery and data mining: Data Mining and Knowledge Discovery Handbook* (Springer, New York, 2005), pp. 981–1003
26. A.L. Samuel, Some studies in machine learning using the game of checkers ii: recent progress. IBM J. Res. Dev. **11**(6), 601–617 (1967)
27. A.J.C. Sharkey, G.O. Chandroth, N.E. Sharkey, A multi-net system for the fault diagnosis of a diesel engine. Neural Comput. Appl. **9**(2), 152–160 (2000)
28. L. Rokach, Decomposition methodology for classification tasks: a meta decomposer framework. Pattern Anal. Appl. **9**(2), 257–271 (2006)

29. Y. Freund, R.E. Schapire, A short introduction to boosting. J. JSAI **14**(5), 771–780 (1999)
30. F. Yuan, X. Li, W. Li-ming, P. Le-ping, S. Ying, in *Knowledge Discovery of Energy Management System Based on Prism, Furia and J48*, vol. 100, ed. by M. Ma. Communication Systems and I.T (Lecture Notes in Electronic Engineering). (Springer, Berlin, Heidelberg, 2011) pp. 593–600
31. V. Kumari, P. Kumar, Fuzzy unordered rule induction for evaluating cardiac arrhythmia. Biomed. Eng. Lett. **3**(2), 74–79 (2013). doi:10.1007/s13534-013-0096-
32. I. Witten, E. Frank, M. Hall, *Data Mining: Practical Machine Learning Tools and Techniques* (Morgan Kaufmann, San Francisco, 2011)
33. F. Provost, T. Fawcett, Analysis and visualization of classifier performance: comparison under imprecise class and cost distributions, in *Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining* (AAAI Press, 1997), pp. 43–48
34. Y. Sun, M. Kamel, Y. Wang, Boosting for learning multiple classes with imbalanced class distribution, in *Proceedings of 6th International Conference on Data Mining ICDM'06*, 2006, pp. 592–602
35. R. Iman, J. Davenport, Approximations of the critical region of the fbietkan statistic. Commun. Stat. A Theor. **9**(6), 571–595 (1980)

Chapter 22
Fuzzy Logic Modeling for Higher Adhesion Strength of Cr/Cr-N Multilayer Thin Film Coating on Aerospace AL7075-T6 Alloy for Higher Fretting Fatigue Life

Erfan Zalnezhad and Ahmed Aly Diaa Mohammed Sarhan

Abstract Adhesion strength of coating is one of the most imperative factors in magnetron sputtering technique. Therefore; exploring the effect of coating parameters on enhancing adhesion strength of coating to substrate is extremely important. In this study, an experimental assessment was carried out to discover the fretting fatigue life of Cr/CrN coated AL7075-T6 alloy with higher adhesion strength to substrate by application of PVD magnetron sputtering technique. A fuzzy logic method was utilized to examine how to achieve higher adhesion of coating regarding to changes in input process parameters including DC power, nitrogen flow rate and temperature for fretting fatigue life Improvement.

Keywords Adhesion · AL7075-T6 · Cr-CrN coating · Fretting fatigue · Fuzzy logic · Magnetron sputters technique

1 Introduction

As light-weight, high strength and high conductivity materials, aluminum alloys are becoming more and more important, particularly in the aircraft and automobile industries for both economic and technical reasons [1]. Dispersion hardening through solution and ageing heat treatments are usually used to induce high static mechanical properties in aluminum alloys. However, these alloys are always subject to different working conditions. Wear and fretting normally originate when the substrate is in contact with other surfaces and they rub each other under normal load, causing share force to act on the surface [2–5]. When two contacting

E. Zalnezhad · A. A. D. M. Sarhan (✉)
Center of Advanced Manufacturing and Material Processing, Department of Mechanical Engineering, Faculty of Engineering, University of Malaya, 50603 Kuala Lumpur, Malaysia
e-mail: ah_sarhan@um.edu.my

E. Zalnezhad
e-mail: erfan@um.edu.my

H. K. Kim et al. (eds.), *Transactions on Engineering Technologies*,
DOI: 10.1007/978-94-017-9115-1_22,

components experience a vibratory motion of small amplitude, it is termed fretting. If these mating components are then subjected to cyclic load, the process is known as fretting fatigue. Fretting fatigue increases the shear and tensile stresses at the contact surface, creating surface flaws which can act as stress concentration sites [6]. For this reason, aluminum alloys are often subjected to surface modifications. Material with proper coating should reveal different properties for working effectively in a given tribological application. Vital coating properties are high adhesion strength to the substrate, low tendency to adhere to the mated material, good wear resistance (high hardness), high fracture toughness and superior chemical and thermal stability. Nowadays, the hard coatings of metal nitrides are used in several tribological applications [7, 8]. The hard coatings are generally deposited via physical vapor deposition techniques such as ion plating, magnetron sputtering, and thermal evaporation, permitting the creation of dense adhesive thin films at low temperatures, which is one of the most important advantages of these sorts of coatings [9]. In the past few decades, transition metal nitride coatings, mostly based on titanium and chromium, have attracted significant attention due to their beneficial potential applications in several fields such as electronics, optical, decorative and magnetic coatings [10, 11]. Cr-CrN thin film coating is particularly appropriate to serve as wear and corrosion protection because of its fine mechanical properties (wear resistance, low friction coefficient and high hardness). Cr-CrN coating using magnetron sputtering technique has the main specific advantages of easily controlled deposition rate and low impurities. Furthermore, this method allows the creation of thin films of numerous crystallographic and morphology structures [12]. Multilayer Cr-CrN coatings are made by changing coating parameters which help the construction of thin films over the crystalline, i.e., lower deposition rate and lower substrate temperature [13]. The traditional method of attaining high strength and hardness at different coating parameters is to use the experimental trial and error approach, which is very time-consuming due to the large number of experiments. Hence, a reliable systematic approach for predicting surface hardness at different parameter conditions is required to cover all the parameter ranges in a low number of experiments [14]. Soft computing techniques are useful when exact mathematical information is not available. In contrast to traditional computing, these techniques suffer from approximation, partial truth, met heuristics, uncertainty, and inaccuracy. One of the soft computing techniques with a significant role in input-output parameter relationship modeling is fuzzy logic system. Artificial intelligence (AI) tools play an important role in manufacturing processes. Compared to other artificial intelligence methods, development of fuzzy logic is moderately easier and it does not need many software and hardware resources. Fuzzy logic was introduced by Zadeh (1965) and is the victorious application of theory of the fuzzy set, as an extension of the set theory by the characteristic function replacement of a set through a membership function whose values range from 0 to 1. A considerable amount of studies have focused on the prediction and measurement of coating surface integrity [15].

In this study, Al7075-T6 substrate was coated with Cr-CrN at different parameter conditions. Each parameter has four levels, namely: substrate temperature, nitrogen percentage and DC power. The fuzzy rule-based method was proposed to investigate surface adhesion of multilayer Cr-CrN coating on AL7075-T6 alloy. The fretting fatigue life of Cr-CrN coated specimens with high adhesion was investigated.

2 Experimental details

The material investigated in this work is Al-7075-T6 aluminum alloy with the following chemical composition (wt%): 4.6Zn; 1.8 Mg; 1.85Cu; 0.06Mn; 0.47Si and 0.28Cr. The ultimate strength and yield stress of Al7075-T6 were attained via a number of tensile tests, and they are: $\sigma_{ut} = 590$ MPa and $\sigma_y = 520$ MPa, respectively.

For sample preparation, the cylindrical shape test specimens, show in Fig. 1a, were machined by lathe turning (CNC LATHE MACHINE, Miyano, BNC-42C5) in accordance with ISO 1143 standard [16]. Fretting fatigue pads were fabricated from AISI 4140 steel plate with hardness of 346HV. Substrate material (179HV) is softer than the pads but Cr-CrN coating (630HV) is harder. The friction pads drawings are depicted in Fig. 1b.

In order to make the films adhere well to the substrates, surfaces must be carefully cleaned before film deposition. Therefore, all substrates were polished with SiC paper of 800–2,000 grit, and were surface mirrored with diamond liquid. The substrates were ultrasonically cleaned in alkali and alcohol baths, respectively, and thoroughly rinsed with distilled water. The samples were then inserted into the chamber for in situ cleaning. The chamber was evacuated to a pressure of 3.7×10^{-5} Torr, and the substrates were heated to 350 °C for one hour. This process mainly removed water molecules, which were absorbed on almost all surfaces. During the last step of cleaning, called ion-etching or Ar^+ sputtering, Ar^+ ions were accelerated by applying substrate bias potential V_s (−200 V) onto the substrates. In the ion-etching process, oxides or chemisorbed nitrogen and/or carbon atoms were removed. A magnetron sputtering machine (SG Control Engineering Pte Ltd) was utilized in order to deposit thin film on the metal. DC generators were selected to facilitate the sputter metals. Sputtering pressure was adjusted to around 5.2×10^{-3} Torr. Table 1 presents the coating parameter conditions used in this experiment, in an investigation of how to increase sputtered CrN thin film adhesion to the substrate. A pure chromium 99.95 % target was selected for exploring the sputtering parameters condition for AL7075-T6 alloy. Pure chromium was initially coated onto the substrate as an interfacial layer for 1 h to improve adhesion between the substrate and second layer of coating (chromium nitride). The deposition time for the second layer was adjusted to 3 h.

The coating procedure was planned using the experimental array shown in Table 1. Adhesion of coating to substrate is the most essential factor to be investigated. The layers were characterized by scanning electron microscopy (FE/SEM-FEG) and focused ion beam technique (Quanta FEG250). Substrate adhesion

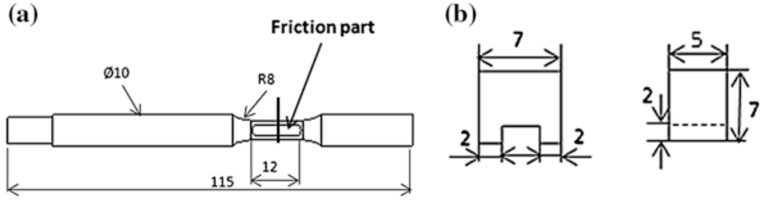


Fig. 1 Drawings of the fretting fatigue specimen and the fretting pad

Table 1 L_{16} (3^4) orthogonal array

Experiment	Parameters combination			Average scratch force (mN)
	A	B	C	845
1	200	150	3	1,033
2	200	200	6	1,030
3	200	250	9	1,372
4	200	300	12	939
5	300	150	6	713
6	300	200	3	1,287
7	300	250	12	1,447
8	300	300	9	1,752
9	400	150	9	1,426
10	400	200	12	1,035
11	400	250	3	794
12	400	300	6	683
13	500	150	12	681
14	500	200	9	823
15	500	250	6	539
16	500	300	3	845

Note
A DC Power (w)
B Temperature (°C)
C Nitrogen low rate (%)

was measured using micro-scratch force equipment (Micro Material Ltd, Wrexham, U.K.). Each experiment was repeated three times and the average values were used for analysis.

A rotating bending fretting fatigue test machine was applied at a frequency of 50 Hz and constant contact pressure of 100 MPa at room temperature to attain the S–N curves. This type of testing was chosen because it produces the greatest amount of stress on specimen surface.

3 Experimental Results

Figure 2a, b show the SEM cross section and surface view of thin film Cr-CrN coating. As it can be seen under the SEM, the coatings consist of two distinct parts: the interfacial layer (pure chromium) and the second layer (CrN) on AL 7075-T6 alloy. The adhesion strength (scratch force) of the coating layers to substrate was measured utilizing the aforementioned equipment. Each measurement was repeated three times, and the measured values are summarized in Table 1.

An initial load (zero) was applied onto a sample by a Rockwell type diamond indenter with a radius of 25 μm at a sliding velocity of 5 μm/s. The load was increased gradually by 9.2 mN/s. Scratch length during the scratch test was 1,000 μm. In the scratch test, critical load Lc could be used to calculate adhesion strength. Critical load magnitude was obtained by applying acoustic signal, friction curve and microscope observation. Acoustic signal produced by film delamination could be used to characterize the critical load (Lc). Scratch adhesion testing was performed on a coated sample to measure Lc. The two images shown in Fig. 3a, b represent the weakest and strongest adhesion strengths.

4 Fuzzy Rule Based Model

Fuzzy logic is a continuous conversion from truth to false conditions, as opposed to the separate true: false transition in binary logic. The possibility theory of fuzzy logic provides a measure of a subset's potential ability to belong to another subset. It can be shown that the probability theory is a special case of the possibility theory [14]. Therefore, fuzzy logic has an extensive scope and range of applications compared to many other statistical methods.

Fuzzy logic in engineering applications utilizes this continuous transition in subset membership to alter a problem from wavy numeric to fuzzy linguistic territories. Fuzzy logic employs conventional language to define variables and uses fuzzy linguistic rules to describe relationships instead of working with numeric values of variables and using mathematical functions. This is especially beneficial in coatings where some variables such as DC power, temperature, and the effect of nitrogen flow rate, have no exact numeric values. It also allows the use of accrued experience and knowledge in the form of rules-of-thumb, which cannot be incorporated into a mathematical formula. The most important power of fuzzy logic is that when correctly choosing fuzzy rules and membership functions, it can simulate very complex and non-linear systems while obviously maintaining the physical inferences and effects of every variable. The fuzzy rule base contains a group of IF–THEN declarations with three inputs, A, B and C, and one output, D. The notion of fuzzy argumentation for three inputs and one output of the fuzzy logic unit are defined as follows:

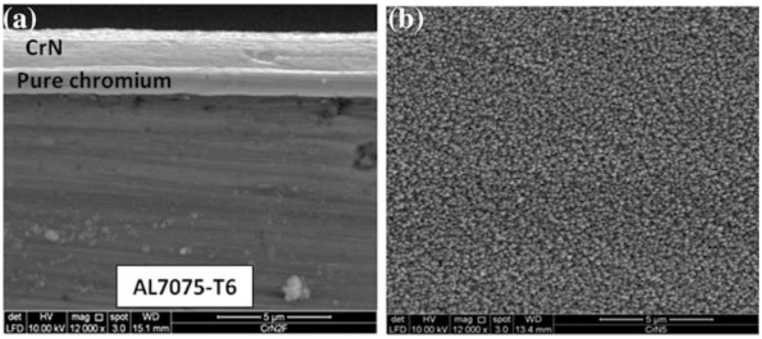


Fig. 2 Typical **a** cross-sectional and **b** surface views of SEM micrograph of Cr-CrN coating on AL7075-T6

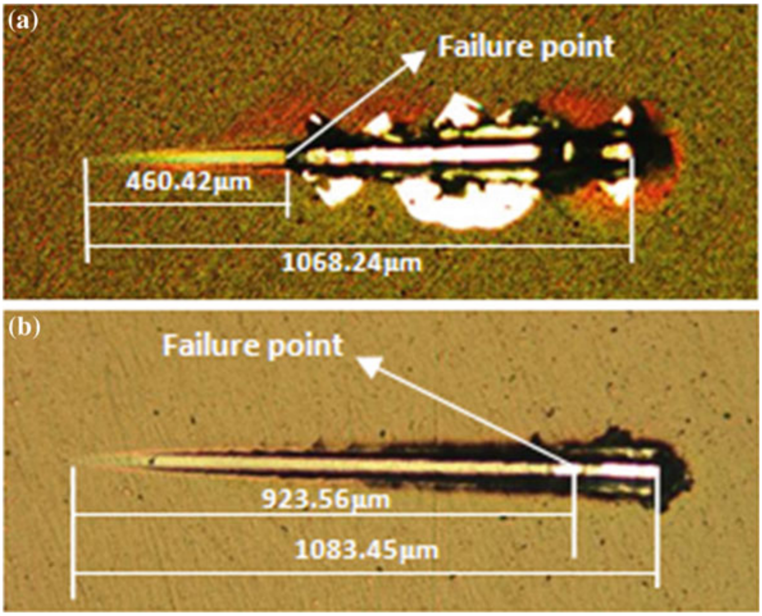


Fig. 3 Adhesion strength of Cr-CrN coated samples at different condition **a** weak adhesion (DC power 500 W, temperature 150 and nitrogen flow rate 3 %, scratch force 510 mN) and **b** strongest adhesion (DC power 300 W, temperature 250 and nitrogen flow rate 9 %, scratch force 2,200 mN)

Rule 1: if A is X_1 and B is Y_1 and C is Z_1 then D is W_1; else
Rule 2: if A is X_2 and B is Y_2 and C is Z_2 then D is W_2; else
Rule n: if A is X_n and B is Y_n and C is Z_n then D is W_n

X_n, Y_n, Z_n, and W_n are fuzzy subsets distinctive by their corresponding membership functions, λ_{Xn}, λ_{Yn}, λ_{Zn}, and λ_{Wn}, respectively.

Sixteen fuzzy rules were established based on the experimental conditions. By following the maximum-minimum compositional process, the fuzzy logic of these rules results in fuzzy output. Assuming that A, B, and C are the three input parameters of the fuzzy logic unit, the membership function of the fuzzy logic output can be stated as [15]:

$$\begin{aligned}\lambda_{\mathrm{W0}}(\mathrm{D}) = [&\lambda_{\mathrm{X1}}(\mathrm{A}) \wedge \lambda_{\mathrm{Y1}}(\mathrm{B}) \wedge \lambda_{\mathrm{Z1}}(\mathrm{C}) \wedge \lambda_{\mathrm{W1}}(\mathrm{D}) \vee \cdots \times \lambda_{\mathrm{Xn}}(\mathrm{A}) \wedge \lambda_{\mathrm{Yn}}(\mathrm{B}) \\ &\wedge \lambda_{\mathrm{Zn}}(\mathrm{C}) \wedge \lambda_{\mathrm{Wn}}(\mathrm{D})]\end{aligned} \tag{1}$$

where, $\vee$ is the maximum and $\wedge$ is the minimum operation. There are different forms of membership functions such as triangular, trapezoidal, Gaussian, sigmoid, etc. In this present study, the Gaussian membership function for input parameters DC power, temperature and nitrogen flow rate, and triangular membership function for the output parameter of coating adhesion strength are presented. The Gaussian fuzzy membership function often used to characterize vague, linguistic terms is given by:

$$\lambda_{\mathrm{A}}^{\mathrm{n}}(\mathrm{X}) = \exp\left(\frac{-(\mathrm{c_n} - \mathrm{x})^2}{2\sigma_{\mathrm{n}}^2}\right) \tag{2}$$

where, $\mathrm{C_n}$ and σ_{n} are the center and width of the *n*th fuzzy set $\mathrm{A^n}$, respectively.

The triangular form membership function for output is definite by three parameters {a, b, c} as follows [13]:

$$\mathrm{f(x; a, b, c)} = \left\{\begin{array}{ll} 0, & x \leq a. \\ \frac{x-a}{b-a} & a \leq x \leq b. \\ \frac{c-x}{c-b} & b \leq x \leq c. \\ 0, & c \leq x. \end{array}\right\} \tag{3}$$

By utilizing minimum and maximum, an alternate statement for the following equation is:

$$\mathrm{f(x; a, b, c)} = \max\left(\min\left(\frac{x-a}{b-a}, \frac{c-x}{c-b}\right), 0\right) \tag{4}$$

where; a, b, c define the triangular fuzzy triplet and determine the x coordinates of the three corners of the underlying triangular membership function.

The numerical input–output values are connected by linguistic variables. This was attained by designing membership functions consisting of a set of fuzzy set values. Linguistic values such as LOW, MEDIUM, HIGH and VERY HIGH show the input variables' DC power, temperature, and nitrogen flow rate. The output numerical values (adhesion to substrate) are also closely connected in the same way, using membership functions such as LOW, AVERAGE, HIGH and HIGHER.

Gaussian membership functions for input parameters' DC power, temperature and nitrogen flow rate and output parameter adhesion strength using triangular membership function are used in this study. Lastly, a defuzzification process was carried out. Defuzzification is an imperative procedure in the theory of fuzzy sets, and modifies fuzzy set information into numeric information. This process, accompanied by fuzzification, is critical in designing fuzzy systems because both of these processes deliver a series of connections between the fuzzy set region and real-valued scalar region. Defuzzification with a form of centroid was designated, because it provides the possibility distribution's center area of the inference output and is a more frequently used defuzzification method for calculating the centroid of the area under the membership function [14, 15].

$$D_0 = \frac{\Sigma D\lambda_{W0}(D)}{\Sigma\lambda_{W0}(D)} \tag{5}$$

The non-fuzzy value D_0 gives the output value in numerical form. Figure 4a, b shows the predicted surface adhesion by fuzzy logic in relation to parameter change. Figure 4a manifests the predicted adhesion strength using fuzzy logic in relation to change nitrogen flow rate and temperature. It is presented that the strength of adhesion between substrate and coating increases by raising the nitrogen flow rate from 3 to 9 %. However, a higher nitrogen value leads to a decline in adhesion strength. Study of Fig. 4b suggests that coating adhesion strength is increased with increasing DC power from 200 to 400 W. From Fig. 4a, b it can be understood that temperature has less effect than the other two parameters. However, the best temperature value for achieving higher adhesion strength of CrN coating seems to be around 250 °C.

After the fuzzy rules were created, six new experimental tests from separate experiments were carried out, while the proposed fuzzy model was used to predict surface adhesion under the same conditions as in Table 2 for further investigation of fuzzy model suitability compared to experimental results. The comparison between experimental results and fuzzy model prediction values are depicted in Fig. 5. The figure conclude that the experimental and fuzzy results are in very close agreement to each other and hence the fuzzy logic method can be efficiently used for predicting the adhesion strength of Cr-CrN coating on substrate in magnetron sputter coating technique. Maximum model error was found to be less than 7 %, indicating that the fuzzy prediction model can be used to predict CrN coating adhesion strength by application of magnetron sputtering technique in change in parameters levels.

5 Fretting Fatigue Life Evaluation of Multilayer Cr-CrN Coated AL7075-T6

In order to investigate the fretting fatigue life of specimens coated with Cr-CrN at high surface adhesion, some experiments were carried out and the results are shown in Fig. 6a, b.

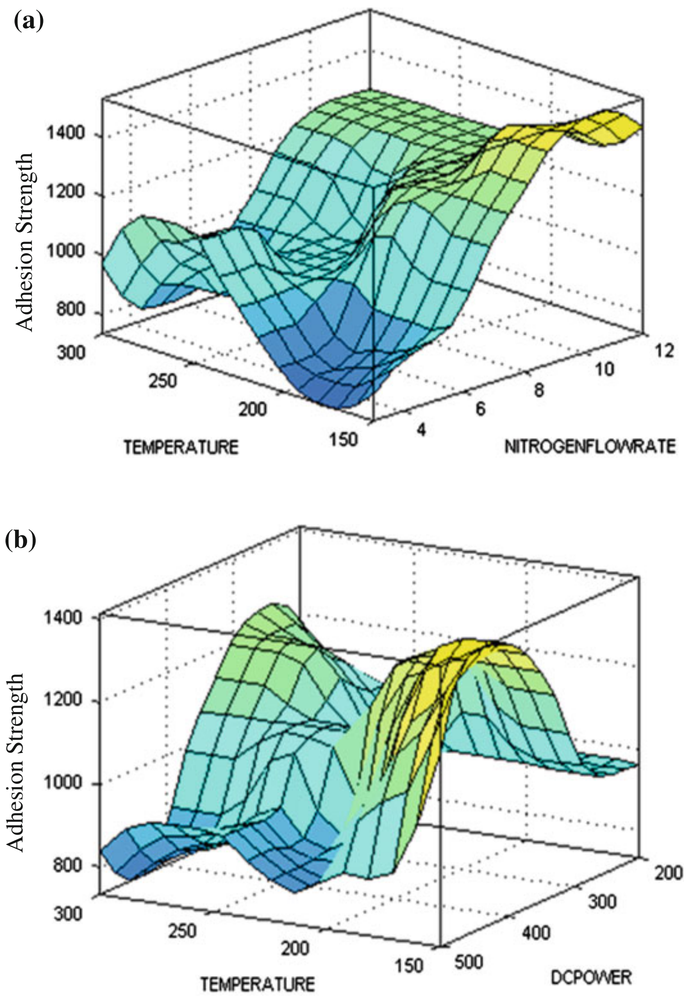


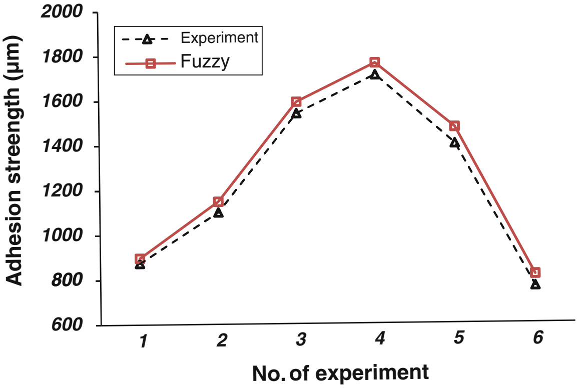
Fig. 4 The predicted adhesion strength by fuzzy logic in relation to parameters change. **a** Adhesion strength in relation to change nitrogen flow rate (*C*). **b** Adhesion strength in relation to change DC power (*A*) and temperature (*B*)

The experiments were conducted for a stress ratio of R = −1, 50 Hz, at a constant contact force of 100 Mpa and working stress amplitudes of 150–300 MPa. Each data point on the S/N curve represents the average of five specimens tested under identical conditions. Figure 6a shows a comparison of the number of cycles to failure versus bending stress for plain fatigue and fretting fatigue of uncoated specimens. As shown in Fig. 6a, the fatigue life of uncoated specimens diminished with increasing bending stress. It is also evident that fretting has a deleterious effect on fatigue life. The S/N curve of fretting fatigue for

Table 2 Accuracy and error of the fuzzy logic model prediction

No. of experiment	Scratch force result (output) (mN)									
	Parameters (inputs)			Measured scratch force				Standard deviation (σ)	Predicted adhesion strength (fuzzy)	Error (%)
	A	B	C	1st	2nd	3rd	Average			
1	225	160	4	871	884	866	874	657,172	897	2.63
2	250	180	5	1,085	1,115	1,101	1,100	1,061,544	1,148	4.36
3	275	220	7	1,565	1,442	1,612	1,540	6,207,405	1,591	3.12
4	350	240	8	1,694	1,712	1,724	1,710	106,7708	1,763	3.10
5	375	260	10	1,416	1,410	1,381	1,402	1,323,584	1,475	5.21
6	450	280	11	748	756	785	763	137,659	815	6.81

Fig. 5 Comparison between experimental and fuzzy results



uncoated and Cr-CrN coated specimens with higher surface adhesion is illustrated in Fig. 6b. The fretting fatigue lives of Cr-CrN coated specimens were enhanced at both low and high cyclic fatigue as opposed to uncoated specimens. Figure 7a, b show typical microscopic image examples from cross sectional views of fractured uncoated and Cr-CrN coated specimens under fretting fatigue test. From the images it can be understood that the fractured surfaces consist of two areas: the fretting regions made by friction pads and a tensile district made by bending stress.

6 Discussion

A. *Adhesion strength*

The results clearly indicate that applying DC power (in the range of 200–400 W) improves coating adhesion properties. This can be attributed to the additional energy available to the growing film. Thus, high-energy atoms have greater mobility to find the surface's low energy sites for maximizing adhesion characteristics. It was observed that maximum coating adhesion was obtained with a

Fig. 6 S/N curve of fretting fatigue test. **a** S/N curve of plain fatigue and fretting fatigue for uncoated specimens. **b** S/N curve of fretting fatigue for uncoated and Cr-CrN coated with highest surface adhesion specimens

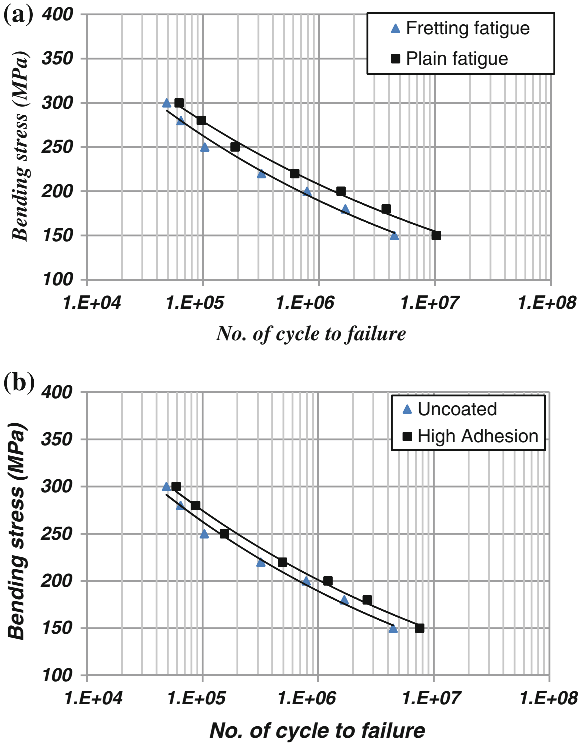
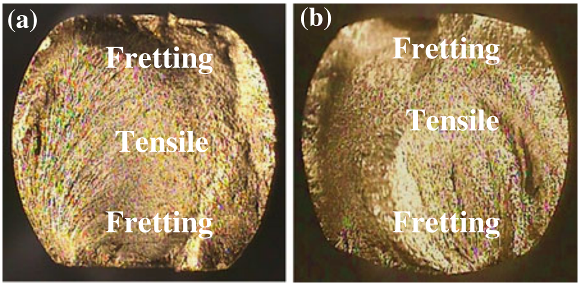


Fig. 7 Typically example of microscopic image from cross sectional view of fractured specimens under fretting fatigue test. **a** uncoated, **b** Cr-CrN coated



critical load of 2,200 mN. The best condition for scratch force of surface was achieved at maximum DC power (400). However, further increasing DC power would result in very high-energy bombardments, adding many defects to the growing film and lowering coating adhesion to a certain level [17]. Temperature has less effect on scratch force (adhesion); the best point for the highest surface adhesion is (250 °C).

Nitrogen is another important parameter playing a significant role in the effect on adhesion between substrate and coating. The detected enhancement in adhesion strength can be attributed to the increased strength of the nitrogen influencing the

Cr interlayer. Nitrogen gas does not significantly change the chemical nature of α-Cr, because no CrN compound is formed and nitrogen is dissolved in the α-Cr lattice. It is thus expected that nitrogen gas will influence the Cr interlayer. When the nitrogen rate increased from 3 to 9 %, the chromium nitride coating adhesion to the substrate increased. But, by further increasing nitrogen content, the interlayer became too strong and brittle to accommodate interfacial stresses, leading to an obvious reduction in adhesion strength. The nitrogen gas at 9 % N2 seemed to produce the optimum strength value for adhesion enhancement in the present deposition conditions [17, 18].

B. ***Fretting fatigue S/N curve***

The corresponding plain fatigue and fretting fatigue S/N curves at contact pressure of 100 MPa are displayed in Fig. 6a. Clearly, there is a significant reduction in fretting fatigue life compared to normal fatigue life, particularly at lower stresses. Decreased fretting fatigue life of uncoated specimens as opposed to plain fatigue is due to the existence of stress concentration in the interface area between the friction pads and substrate. The fretting fatigue crack formed in the region where the frictional shear stress on the contact surface locally concentrated. Thus, the decrease in fatigue life due to fretting damage is considered to be attributed to the increase in crack initiation life caused by the local stress concentration produced by fretting, and the acceleration of the initial crack propagation by fretting.

Figure 6b shows the S/N curve of fretting fatigue for uncoated and Cr-CrN coated AL7075-T6 with the highest surface adhesion. The fretting fatigue lives of coated specimens with high surface adhesion improved at both low and high cyclic fatigue. The values of surface hardness, adhesion and roughness of Cr-CrN coated specimens with the best adhesion are 630 HV, 2,150 mN and 0.055 μm respectively. The surface hardness of Cr-CrN with high adhesion coating is 630 HV, which seems to be sufficient to endure under the contact pressure of the fretting pads (346 HV).

Figure 7 represents fracture of a Cr-CrN coated specimen with high adhesion (2,200 mN). The fretting fatigue life of this Cr-CrN coated specimen is more than an uncoated specimen's, which is attributed to sufficient hardness, higher elasticity and good adhesion that caused the coating to endure the contact pressure under the fretting pads. Another advantage of the Cr-CrN coating with high adhesion is the suitable roughness (0.055 μm) which effectively influences the increase in fretting fatigue life.

A comparison between fretting fatigue life of uncoated and Cr-CrN coated specimens indicate that the lives of coated specimens increased more at high cyclic fatigue, which is attributed to the coating's sufficiently high stability under the fretting pads during the fretting fatigue test. The fretting fatigue results of the coated specimen with higher adhesion strength imply that Cr-CrN multilayer coating on AL7075-T6 alloy by magnetron sputtering technique can improve the fretting fatigue lives of AL7075-T6 by 70 and 22 % at high and low cyclic fatigue, respectively.

7 Conclusion

In this research work, Cr-CrN thin film was coated on AL7075-T6 alloy at different coating parameter conditions utilizing PVD magnetron sputtering technique. The influence of different coating parameters on adhesion strength to substrate was investigated through a fuzzy logic model. The experimental and fuzzy logic model results are in very close assent. The fretting fatigue lives of samples coated with Cr-CrN with the best adhesion strengths were evaluated at different bending stresses. The S/N curve indicates that thin film Cr-CrN multilayer coating improved the fretting fatigue life of AL7075-T6 alloy by 70 and 22 % at high and low cyclic fatigue respectively.

Acknowledgment This research was funded by the UMRG grant with number: RP021C-13AET from the University of Malaya, Malaysia.

References

1. K. Genel, The effect of pitting on the bending fatigue performance of high-strength aluminum alloy. Scripta Mater. **57**, 297–300 (2007)
2. P.S. Pao, S.J. Gill, C.R. Feng, On fatigue crack initiation from corrosion pits in 7075-T7351 aluminum alloy. Scripta Mater. **43**, 391–396 (2000)
3. H. Lee, S. Mall, Fretting behavior of shot peened Ti-6Al-4 V under slip controlled mode. Wear **260**, 642–651 (2006)
4. G.H. Majzoobi, J. Nemati, A.J. NovinRooz, G.H. Farrahi, Modification of fretting fatigue behavior of AL7075-T6 alloy by the application of titanium coating using IBED technique and shot peening. Tribol. Int. **42**, 121–129 (2009)
5. E. Zalnezhad, A.D.A. Sarhan, M. Hamdi, Investigating the fretting fatigue life of thin film titanium nitride coated aerospace Al7075-T6 alloy, **559**, 436–446 (2013)
6. L.C. Lietch, H. Lee, S. Mall, Fretting fatigue behavior of Ti-6Al-4 V, under seawater environment. Mater. Sci. Eng., A **403**, 281–289 (2005)
7. G.H. Majzoobi, M. Jaleh, Duplex surface treatments on AL7075-T6 alloy against fretting fatigue behavior by application of titanium coating plus nitriding. Mater. Sci. Eng. A **452–453**, 673–681 (2007)
8. S. Ortmann, A. Savan, Y. Gerbig, H. Haefke, In-process structuring of CrN coatings and its influence on friction in dry and lubricated sliding. Wear **254**, 1099–1105 (2003)
9. B. Sresomroeng, V. Premanond, P. Kaewtatip, A. Khantachawana, A. Kurosawa, N. Koga, Performance of CrN radical nitrided tools on deep drawing of advanced high strength steel. Surf. Coat. Technol. **205**, 4198–4204 (2011)
10. R. Rebole, A. Martinez, R. Rodriguez, G.G. Fuentes, E. Spain, N. Watson, J.C.A. Batista, J. Housden, F. Montala, L.J. Carreras, T.J. Tate, Microstructural and tribological investigations of CrN coated, wet-stripped and recoated functional substrates used for cutting and forming tools. Thin Solid Films **469–470**, 466–471 (2004)
11. G. Bertrand, H. Mahdjoub, C.A. Meunier, Study of the corrosion behavior and protective quality of sputtered chromium nitride coatings. Surf. Coat. Technol. **126**, 199–209 (2000)
12. E. Ufuah, T.H. Tashok, Behavior of stiffened steel plates subjected to accidental loadings. Eng. Lett. **21**(2), 95–100 (2013)
13. E. Zalnezhad, A.D.A. Sarhan, M. Hamdi, Adhesion strength predicting of Cr/CrN coated AL7075 using fuzzy logic system for fretting fatigue life enhancement, in *Lecture Notes in*

Engineering and Computer Science: Proceedings of the World Congress on Engineering and Computer Science 2013, WCECS 2013, pp. 589–595, San Francisco, USA, 23–25 Oct 2013

14. C. K. On, J. Teo, Artificial neural controller synthesis in autonomous mobile cognition. IAENG Int. J. Comput. Sci. **36**(4), 240–252 (IJCS_36_4_01) (2009)
15. A. Soleimani, Z. Kobti, Event-driven fuzzy paradigm for emotion generation dynamics, in *Lecture Notes in Engineering and Computer Science: Proceedings of the World Congress on Engineering and Computer Science 2013, WCECS 2013*, pp.168–173, San Francisco, USA, 23–25 Oct 2013
16. ISO Standard, Metallic materials—rotating bar bending fatigue testing, ISO International (2010)
17. C.I. Nkeki, C.R. Nwozo, Optimal investment under inflation protection and optimal portfolios with stochastic cash flows strategy. IAENG Int. J. Appl. Math. **43**(2), 54–63 (2013)
18. B. Latha, V.S. Senthilkumar, Modeling and analysis of surface roughness parameters in drilling GFRP composites using fuzzy logic. Mater. Manuf. Processes **25**, 817–827 (2010)

Chapter 23
Measuring the Usability of the Interface of the Saudi Digital Library

Abdulrahman Nasser Alasem

Abstract With the rapid expansion of the use of web-based resources for education purposes, the usability of a digital library has come to be considered as an important element in the achievement of the full potential of any DL project. This paper applied a questionnaire-based usability test as the main method of measuring the usability of the interface of the Saudi Digital Library (SDL). Based on related studies, a set of sixteen items covering four axes: efficiency, effectiveness, aesthetic appearance and learnability were developed to evaluate the interface of the SDL. Twenty-two undergraduate students from the department of Information Studies in IMAMU participated in completing the Likert scale questionnaire. The main finding of the study indicated that the level of usability of the SDL's interface was not acceptable, in particular in terms of aesthetic appearance. Moreover, it appears that problems facing other Internet applications in Saudi Arabia will continue to influence the development of digital library projects.

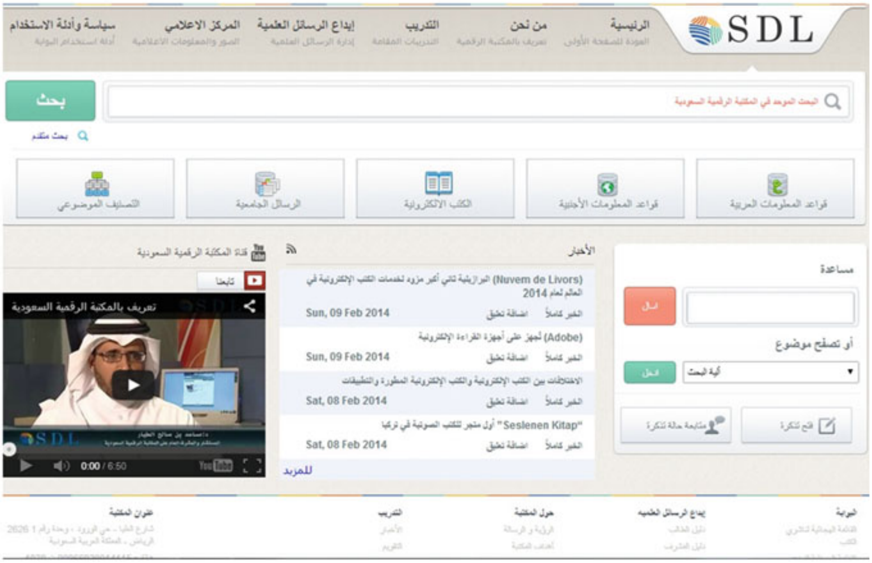
Keywords Digital library · Interface design · Questionnaire-based usability test · Saudi digital library · Usability · Users studies

1 Introduction

Higher education in Saudi Arabia is expanding significantly because of the considerable government investment in human resource development, particularly higher education reform. Prior to 2002, there were only eight government

A. N. Alasem (✉)
Information Studies Department, College of Computer and Information Sciences,
Al Imam Mohammad Ibn Saud Islamic University (IMBSIU), P. O. Box 5701,
Riyadh 11432, Saudi Arabia
e-mail: a.alasem@ccis.imamu.edu.sa

H. K. Kim et al. (eds.), *Transactions on Engineering Technologies*,
DOI: 10.1007/978-94-017-9115-1_23,



(Nuvem de Livors) البرازيلية ثاني أكبر مزود لخدمات الكتب الإلكترونية في العالم لعام 2014
Sun, 09 Feb 2014 الخبر كاملاً اضافة تعليق

(Adobe) تُجهز على أجهزة القراءة الإلكترونية
Sun, 09 Feb 2014 الخبر كاملاً اضافة تعليق

الاختلافات بين الكتب الإلكترونية والكتب الإلكترونية المطورة والتطبيقات
Sat, 08 Feb 2014 الخبر كاملاً اضافة تعليق

"Seslenen Kitap" أول متجر للكتب الصوتية في تركيا
Sat, 08 Feb 2014 الخبر كاملاً اضافة تعليق

Fig. 1 The interface of the Saudi digital library

universities and two private universities located in the main cities; however, today, there are twenty-five government universities and eleven private universities dispersed across the country [1, 2]. As a result of this development, in 2010, the Saudi Digital Library (SDL) was established by the Ministry of Higher Education (MOHE). According to the library's website, the sophisticated digital library aims to support the educational process and to meet the needs of researchers, students, and professionals in higher education. It has more than 24 000 full e-books in various scientific specializations. In addition, the SDL has subscribed to about 300 local, regional, and international publishers [3, 4] (Fig. 1).

However, a significant issue associated with electronic projects that has been addressed by many authors is the ease of using the systems. Maurer argued that if a system is difficult to use, users will react to it in one of several ways: first, they will not use it at all; second, they will carry out their tasks elsewhere; third, they will use it as little as possible, and; finally, they will spend time and seek support to learn how to use it [5]. The issue of ease of use for electronic projects has become more significant in developing countries, including Saudi Arabia. Salem [6] argues that the most important factors causing the failure of electronic initiatives in these countries are a lack of experience with large ICT projects and inadequate awareness of website design issues. This lack of awareness of website design, including usability, accessibility, and information architecture, among designers has been confirmed by many studies across several domains in Saudi Arabia (e.g., eGovernment, eCommerce) as a reason for the failure of the majority of their websites, which suffered from serious web-design problems that did not encourage people to use them [7, 8]. Another issue that needs to be addressed here is the fact that the SDL is the only digital library where academics can access electronic information resources for educational purposes, as it is the single point of access to national and international electronic resources. Therefore, users' experiences in using technology play a considerable part in the usage of systems. This has been identified as a significant element that needs to be considered when designing a website [9].

This study, therefore, attempts to fill this gap in knowledge by testing the usability of the SDL's interface using a questionnaire-based usability test technique to clarify how Saudi university students see the library's website in terms of usability. The study is organized into five sections. Section two is a review of the literature, focusing on studies using questionnaire-based usability tests. Sect. 3 describes the research methodology. The analysis and results are discussed in Sect. 4. Finally, Sect. 5 summarizes the study.

2 Literature Review

The rapid development of information and communication technology (ICT) continues to have a significant impact on all sectors of society. The library and information industry is no exception, with terms such as electronic library, digital library, hybrid library, and virtual library being used interchangeably to represent the same concept of collection digital objects that can be accessed over the Internet, together with methods of selecting, organizing and managing these objects [10]. Ease of access and ease of use are core characteristics of any online electronic project and reflect the main objective of establishing such projects.

Since the 1990s, much research and development has been carried out in the field of digital libraries, starting with the creation of specialist digital libraries (e.g., education, health care, and learning). The usability of digital libraries, which refers broadly to the user's experiences and satisfaction of using a digital library, has been the focus of the majority of these studies. Usability is important because it encourages use of a system by those who are connected to the Internet and attracts users who are not yet connected. According to Chowdhury et al. [11], the focal point of a digital library is its website, so the literature on the usability of digital libraries is extensive and rich [12]. Moreover, the usability attributes axes proposed by the International Organization for Standardization (ISO) or by the research community since the 1980s to guide the measurement of computer systems are still evolving and changing. For instance, in the ISO 9126-1:2001, which has been revised to the ISO/IEC 25010:2011, 'understandability' and 'attractiveness' have been renamed 'appropriateness' and 'interface aesthetics', and error protection, and accessibility have been added [13]. The research community also uses a number of different usability attributes; for instance, Nielson [14] proposed learnability, efficiency, memoability, errors, and satisfaction, whereas Tsakonas and Papatheodorou [15] specified learnability, ease of use, aesthetic appearance, navigation, and terminology. This has given the usability research community the ability to identify sub-attributes according to the context, the study objective, and the perspective of the researcher. In addition, the literature shows that digital libraries have been evaluated from various perspectives, including interface design and features, search facilities and process, text size and color, graphics and background color, labeling and language, navigation and information organization, and content and service [16–19].

In addition, the different usability test techniques either involve users (e.g. questionnaires, thinking aloud, and observation) or do not involve users (such as heuristics, cognitive, walkthrough, and action analysis); however, they have several things in common [20, 21]. Questionnaire-based usability test techniques are widely used for testing the usability of digital library website designs. Conyer [22] states that the advantages of this approach include its being inexpensive and easy to conduct. In addition, it is easy to quantify. Moreover, Joo et al. [23] highlighted the advantages of using a questionnaire-based usability test, which is to gather data directly from the expected users of the system. Moreover, it has been argued that it would be preferable to conduct an inspection and a formal usability test before the system became available for users, whereas other methods such as the use of a questionnaire would be used after the system became available to users [24].

The purpose of questionnaire-based usability test is to learn from the actual users about their ability to complete specific tasks successfully, how long they need to finalize the tasks and their level of satisfaction, to identify which areas need to be improved and to analyze the overall performance to see whether the usability objective has been met or not [25]. However, the questionnaire-based usability test, as with other kinds of empirical evaluation, has the limitation that it does not observe or measure actual performance; rather it tends to seek users' opinions and their experiences [26]. Maurer [5] argues that working with actual users is the only way to know whether the system is usable or not.

An example of studies of digital libraries that utilized a questionnaire-based usability test as the main tool in gathering data is the study of CUNY+ conducted by Oulanov and Pajarillo [27]. A questionnaire with 20 measurement items was developed around five axes (affect, efficiency, adaptability, control and helpfulness) was developed to compare the usability of text-based and web-based CUNY+ and tested with 10 users. Further, Soohyung et al. [28] conducted a usability test using a questionnaire to evaluate the UWM academic library website. Their questionnaire combined 28 measurement items developed around three axes, effectiveness, efficiency and learnability, and was tested with 147 actual users. Buchanan and Salako [20] developed a questionnaire to be used to test the usability of digital library around six axes: effectiveness/efficiency, aesthetic appearance, terminology, navigation/learnability, relevance and reliability/currency. Joo [29] in a questionnaire with 14 measurement items covering the four dimensions of efficiency, effectiveness, satisfaction and learnability.

Through reviewing the literature, it is clear that there are disagreements on the way in which the questionnaire is used in terms of what the key criteria are

Table 1 The four axes with their items

Axis	Item code	Items indicator
Efficiency	EFY1	Generally, the SDL is easy to use
	EFY2	The SDL responded without errors
	EFY3	The SDL responded quickly
	EFY4	Overall, the SDL is well designed in terms of helping users find what they want
Effectiveness	EFT1	Information located on the homepage is easy found
	EFT2	Generally, tasks can be completed easily
	EFT3	Overall, I'm satisfied with Saudi SDL
	EFT4	The search facility provided in the SDL is effective
Aesthetic appearance	AES1	The menu functions is well organized
	AES2	Color, graphics, and icons have been used appropriately
	AES3	Text type and font size are readable
	AES4	Text color and background color are consistent
Learnability	LER1	The SDL's interface is easy to lean
	LER2	The terminologies used are understandable
	LER3	Steps to complete tasks are clear and understandable
	LER4	SDL has an appropriate help function

for evaluating the usability of a digital library. As can be seen, different axes along with different measurement items have been used. This could be due to the fact that to each country has its own characteristics in terms of cultural, social, Internet technology infrastructure and usage, as well as the objectives of the study. However, even with the differences in the use of a questionnaire to test usability, there is a consensus about the core functions of how to develop the instrument.

3 Methodology

As mentioned previously, although a questionnaire-based usability test has certain limitations, one of its main advantages is that it reflects of the opinions of actual users rather than those of usability experts. The questionnaire was developed after reviewing and analyzing several digital library usability studies that used questionnaires to collect data [20, 27–30]. Four axes of usability were identified: efficiency, which refers to the effort to complete tasks; effectiveness, which refers to the ability of users to complete tasks; aesthetic appearance, or the consistency of the interface design as a whole; and learnability, which refers to how easy and quickly users can use the system. Sixteen questions were drawn from these four axes, and responses were recorded using Likert scales with end points ranging

Table 2 General information of respondents

Gender	Average age	Education status	Daily internet frequency use	Average computer experience
Male	24 years	Undergraduate	6 h	10 years

from (1) strongly disagree to (5) strongly agree. Table 1 shows the four axes with their items.

4 Results and Discussion

Twenty-two students from the department of Information Studies at IMAMU participated in this study. Table 2 shows the general information of participants.

Internal reliability of the instrument used in this study was examined with Cronbach's Alpha (α) to determine the consistency of the items listed in the four Axis on the questionnaire distributed. The statistics showed that the value was (0.743) for the sixteen items, which was acceptable. On the other hand, the internal validity, which is concerned about the verification that items listed in the four Axis are based on the study's objective, were considered. Consequently, the sixteen items questionnaire was developed and reviewed based on the related literature, as well as its applicability to the SDL. The external validity, which is concerned about to what extend the condition of the test is somewhat similar to the real world [31] was of concern and the test was carried out in the college's lab where the students usually use it to access the SDL. Table 3 shows the results of the study.

For the sixteen items, the descriptive statistics of frequency, percentage and mean were investigated. The above table present the results of analysis of the questionnaire data. Overall, the mean value of participants for all sixteen items was 3.04. and the mean value for the four axes were 3.35, 3.10, 2.34 and 3.38 for efficiency, effectiveness, aesthetic appearance and learnability, respectively. Although the results of the analysis in general were neutral, items regarding aesthetic appearance were found to be negative, with a mean value of 2.34, indicating that the website failed to provide an appropriate good aesthetic appearance. In this regard, the majority of the respondents commented that the text size used (11px) is too small and this caused a readability issue. In addition, in some parts of the website there is no consistency between the text color and background color used.

Moreover, this can be noticed in the fourth item (effectiveness), which was concerned with search functionality, with a mean value of 2.64. On the other hand, only three items across the four axes showed a positive attitude with a mean value of 3.41 or above. These were 3.73, 3.41 and 3.95, for 'the SDL responded quickly',

Table 3 Questionnaire results

Measurement item	Strongly disagree, n (%)	Disagree, n (%)	Natural, n (%)	Agree, n (%)	Strongly agree, n (%)	Mean
EFY1	3 (13.6)	4 (18.2)	3 (13.6)	7 (31.8)	5 (22.7)	3.32
EFT2	0 (0.0)	5 (22.7)	7 (31.8)	7 (31.8)	3 (13.6)	3.36
EFY3	0 (0.0)	3 (13.6)	7 (31.8)	5 (22.7)	7 (31.8)	3.73
EFY4	2 (9.1)	7 31.8)	6 (27.3)	3 (13.6)	4 (18.2)	3.00
EFT1	4 (18.2)	3 13.6)	4 (18.2)	7 ((31.8)	4 (18.2)	3.18
EFT2	0 (0.0)	4 (18.2)	7 (31.8)	9 (40.9)	2 (9.1)	3.41
EFT3	0 (0.0)	7 (31.8)	6 (27.3)	7 (31.8)	2 (9.1)	3.18
EFT4	4 (18.2)	7 (31.8)	6 (27.3)	3 (13.6)	2 ((9.1)	2.64
AES1	4 (18.2)	2 (22.7)	3 (13.6)	4 (18.2)	6 27.3)	3.14
AES2	7 (31.8)	7 (31.8)	5 (22.7)	3 (13.6)	0 (0.0)	2.18
AES3	8 (36.4)	8 (36.4)	3 (13.6)	3 (13.6)	0 (0.0)	2.05
AES4	5 (22.7)	5 (22.7)	5 (22.7)	5 (22.7)	5 (22.7)	2.00
LER1	2 (9.1)	7 (31.8)	3 (13.6)	6 (27.3)	4 (18.2)	3.14
LER2	0 (0.0)	0 (0.0)	9 (40.9)	5 (22.7)	8 (36.4)	3.95
LER3	0 (0.0)	5 (22.7)	8 (36.4)	7 (31.8)	2 (9.1)	3.27
LER4	0 (0.0)	6 (27.3)	9 (40.9)	4 (18.2)	3 (13.6)	3.18

'generally, tasks can be completed easily' and 'the terminologies used on the SDL's interface are understandable', respectively.

5 Conclusion

The usability evaluation of SDL using a questionnaire-based usability test technique showed that issues were raised in this study that support those shown in previous studies investigating official Saudi Arabian websites in various domains. The level of usability practice was not acceptable, particularly for the aesthetic appearance axis. This leads the issue that currently, the key point is not only to have a website, but how effective the website is in terms of usability, accessibility and findability. Addressing these issues may reduce other factors that can limit the use SDL, as in the case of Saudi Arabia familiarity with online activities is generally low. This is because of the absence of other Internet applications. Hence, considering usability issues is significant to bridge the new concept of the digital divide, which is a skills divide mainly regarding use of the Internet as a tool for pursuit of a better standard of living [32].

Finally, although this study developed the measurement instrument after reviewing relevant literature on the usability of digital library, it would be preferable for the instrument to have been reviewed by an expert in the area of digital libraries in order to reflect a better understanding of the digital library's features. In addition, the sampling used in this study did not include other expected users of the system, such as female students and faculty members. Thus, more studies are needed in order to obtain a better understanding of Saudi digital library users' views on the usability of the system.

References

1. MOHE, Government and private universities (2009). Available http://www.mohe.gov.sa/ar/studyinside/Government-Universities/Pages/default.aspx. Accessed 4 Feb 2014
2. S. Mahnoud, Development of higher education in Saudi Arabia. High. Educ. **15**(1–2), 17–23 (1986)
3. MOHE, Establishing the Saudi digital library (2010). Available http://www.mohe.gov.sa/ar/news/Pages/30_october_2_2010.aspx. Accessed 4 Feb 2014
4. Saudi Digital Library (2011). About us, available http://portal.sdl.edu.sa/english/. Accessed 3 Feb 2014
5. D. Maurer, What is usability (2006). Available http://www.steptwo.com.au/papers/kmc_whatisusability/pdf/KMC_WhatIsUsability.pdf. Accessed 5 Feb 2014
6. F. Salem, Exploring e-government barriers in the Arab States. Belfer Center for Science and International Affairs (2006). Available http://belfercenter.ksg.harvard.edu/files/DSGeGOVBRIEF_egov_fadi.pdf. Accessed 5 Feb 2014
7. A. Eidaroos, S. Probets, J. Dearnley, Heuristic evaluation for e-government websites in Saudi Arabia (2009). Available https://dspace.lboro.ac.uk/dspacejspui/bitstream/2134/5779/1/Heuristic%20Evaluation%20for%20e-Government%20Websites%20in%20Saudi%20Arabia_final.pdf. Accessed 11 Feb 2014
8. A. Abanumy, A. Al-Badim, P. Mayhew, E-government website accessibility: in-depth evaluation of Saudi Arabia and Oman. Electron. J. e-Gov. **3**(3), 99–106 (2005). Available http://www.ejeg.com/issue/download.html?idArticle=48. Accessed 11 Feb 2014
9. E. Duncker, Y. Theng, N. Mohd-Nasir, Cultural usability in digital libraries. Bull. Am. Soc. Inf. Sci. **26**(4) (2000). Available: http://www.asis.org/Bulletin/May-00/duncker__et_al.html Accessed 8 Feb 2014
10. M. Nazim, Digital library: contents and services. 3rd convention planner, Assam University (2005). Available http://eprints.rclis.org/10933/. Accessed 10–11 Nov 2005
11. G. Chowdhury, P.F. Burton, D. McMenemy, A. Poulter, *Librarianship: An Introduction* (Facet Publishing, London, 2008)
12. J. Gonzalo et al. (eds.), ECDL 2006, LNCS, vol. 4172, pp. 208–219 (2006)
13. BSI, Systems and software engineering—systems and software quality requirements and evaluation (SQuaRE)—system and software quality models (2011). Available: http://janus.uclan.ac.uk/pagray/BS-ISO-IEC%2025010%202011%20quality%20requirements%20models.pdf
14. J. Nielsen, Usability 101: Introduction to Usability (2012). Available http://www.nngroup.com/articles/usability-101-introduction-to-usability/. Accessed 8 Feb 2014
15. G. Tsakonas, C. Papatheodorou, Analysing and evaluating usefulness and usability in electronic information services. J. Inf. Sci. **32**(5), 400–419 (2006)
16. G. Chowdhury, From digital libraries to digital preservation research: the importance of users and context. J. Documentation, **66**(2), 207–223 (2010)
17. A. Blandford, G. Buchanan, Usability of digital libraries: a source of creative tensions with technical developments (2003). In IEEE-CS technical committee on digital libraries' on-line newsletter. Available http://www.ieeetcdlorg/Bulletin/current/blandford/blandford.htm. Accessed 3 Feb 2014
18. Y. Manolopoulos et al. (eds.), PCI 2001, LNCS, vol. 2563, pp. 217–231 (2003)
19. H. Hartson, P. Shivakumar, M. Quinones, Usability inspection of digital libraries: a case study. Int. J. Digit. Libr. **4**(2), 108–123 (2004)

20. S. Buchanan, A. Salako, Evaluating the usability and usefulness of a digital library. Libr. Rev. **58**(9), 638–651 (2009)
21. J. Sauro, What's the difference between a heuristic evaluation and a cognitive walkthrough? (2011). Available http://www.measuringusability.com/blog/he-cw.php. Accessed 10 Feb 2014
22. M. Conyer, User and usability testing - how it should be undertaken? Aust. J. Educ. Technol. **11** (2). pp. 38–51 (1995). Available http://www.ascilite.org.au/ajet/ajet11/conyer.html
23. S. Joo, S. Lin, K. Lu, A usability evaluation model for academic library websites: efficiency, effectiveness and learnability. J. Libr. Inf. Stud. **9**(2), 11–26 (2011)
24. D. Farkas, Evaluation and usability testing (2013). Available http://faculty.washington.edu/farkas/HCDE%20407-2013/Evaluation%20and%20Usability%20Testing.pdf Accessed 3 Feb 2014
25. Usability.gov (no date). Usability testing. Available http://www.usability.gov/how-to-and-tools/methods/usability-testing.html. Accessed 4 Feb 2014
26. N. Avouris, N. Tselios, C. Fidas, E. Papachristos, in *Website Evaluation: A Usability-Based Perspective*, ed. by Y. Manolopoulos, S. Evripidou, A. Kakas, Advances in Informatics SE: 15, vol. 2563. (Springer, Berlin, 2003) pp. 217–231
27. A. Oulanov, E.J.Y. Pajarillo, CUNY+ Web: usability study of the web-based GUI version of the bibliographic database of the City University of New York (CUNY). Electron. Libr. **20**(6), 481–487 (2002)
28. J. Soohyung, S. Lin, K. Lu, A usability evaluation model for academic library websites: efficiency, effectiveness and learnability. J. Libr. Inf. Stud. **9**(2), 11–26 (2011)
29. S. Joo, Measuring the usability of academic digital library. Electron. Libr. **29**(4), 523–537 (2010)
30. A. Alasem, Evaluating the usability of Saudi digital library's interface (SDL), in *Lecture notes in engineering and computer science: Proceedings of the World Congress on engineering and Computer Science 2013, WCECS 2013*, pp. 178–181, San Francisco, USA, 23–25 Oct 2013
31. M. Hughes, Reliability and dependability in usability testing (2011). Available http://www.uxmatters.com/mt/archives/2011/06/reliability-and-dependability-in-usability-testing.php. Accessed 10 Feb 2014
32. F. Bélanger, L. Carter, The impact of the digital divide on e-government use. Commun. ACM **52**(4), 132–135 (2009)

Chapter 24
Predictive Models for Undergraduate Student Retention Using Machine Learning Algorithms

Ji-Wu Jia and Manohar Mareboyana

Abstract In this paper, we have presented some results of undergraduate student retention using machine learning and wavelet decomposition algorithms for classifying the student data. We have also made some improvements to the classification algorithms such as Decision tree, Support Vector Machines (SVM), and neural networks supported by Weka software toolkit. The experiments revealed that the main factors that influence student retention in the Historically Black Colleges and Universities (HBCU) are the cumulative grade point average (GPA) and total credit hours (TCH) taken. The target functions derived from the bare minimum decision tree and SVM algorithms were further revised to create a two-layer neural network and a regression to predict the retention. These new models improved the classification accuracy. Furthermore, we utilized wavelet decomposition and achieved better results.

Keywords Decision tree · Machine learning · Neural network · Signal processing · Student retention · Support vector machines

1 Introduction

This paper studies the HBCU undergraduate student retention [1]. We explore the effectiveness of machine learning techniques to determine factors that influence student retention at an HBCU and create retention predictive models [2–8, 9].

J.-W. Jia (✉)
Bowie State University, 14000 Jericho Park Road, Bowie, MD 20715, USA
e-mail: jjia@bowiestate.edu

M. Mareboyana
Department of Computer Science, Bowie State University, 14000 Jericho Park Road, Bowie, MD 20715, USA
e-mail: MMareboyana@bowiestate.edu

H. K. Kim et al. (eds.), *Transactions on Engineering Technologies*,
DOI: 10.1007/978-94-017-9115-1_24,

Table 1 List of data set attributes

Number	Name	Description	Type
1	GPA	The last cumulative GPA while student enrolled	Number
2	TCH	The max total credit hours taken while student enrolled	Number
3	School	School that student enrolled in Fall 2006	Text
4	Plan	Academic program that student enrolled in Fall 2006	Text
5	Distance	Commuting distance of the student	Number
6	Gender	Student gender	Text
7	Age	Student age	Number
8	Race	Student race	Text
9	FINAID	The amount of financial aid that student awarded in Fall 2006	Number
10	SAT I Math	Student SAT I Math score	Number
11	SAT I Verb	Student SAT I Verbal score	Number
12	Retention	If student graduated or enrolled in Fall 2011 then yes, else no	Text

Table 2 Training data set numeric attributes

Naive Bayes	No retention		Retention	
Attribute name	Mean	Std. Dev.	Mean	Std. Dev.
GPA	1.9371	±0.8913	2.8864	±0.4276
TCH	50.9574	±35.9515	149.3327	±23.8385
Distance	5.8922	±9.6504	3.3919	±6.942
Age	18.6056	±1.5723	18.4133	±0.7819
FINAID	8165.95	±6924.45	8706.66	±7101.55
SAT I math	425.6	±57.48	425.77	±59.7
SAT I verb	442.27	±50.63	441.66	±54.63

In general, learning algorithms attempt to maximize classification accuracy (percentage of instances classified correctly) to obtain a correct solution of high quality (length, efficiency) [10].

We started collecting data from the HBCU Fall 2006 full-time and first-time undergraduate students, and tracked these students' activities in the following 6 years from Fall 2006 to Fall 2011. The data was queried from the Campus Solution database. The 6-year training data set size is 771 instances with 12 attributes shown in Table 1. The HBCU undergraduate 6 years retention rate 44.9 % was derived from the 6-year training data set [2, 11]. The HBCU 6-year training data set numeric attributes and statistics are shown in Table 2.

We classified the data under two groups—"Retention"—students who were retained in the HBCU and "No Retention"—students who were not retained in the HBCU.

Firstly, we used Weka to classify the cohorts' 6-year training data set using different machine learning algorithms with a goal to maximize classification accuracy (percentage of instances classified correctly). The models derived by J48 decision tree, Simple Logistic, Naïve Bayes and JRip algorithms gave classification accuracies of about 94.03 %.

Secondly, we pruned the J48 decision tree to get the bare minimum decision tree, and we found the learning rule for the HBCU undergraduate student 6-year retention. Then we created a neural network model for predicting retention, the model's accuracy was 94.16 %. We improved the model's performance from 94.16 to 94.42 %.

In addition, we used SVM algorithm and created a regression that proved the selection of the two major factors that affect the retention, which are cumulative GPA and total credit hours taken.

Furthermore, we extended the 6-year training classification to seven student academic level classifications, which are 6-year, 5-year, 4-year, 3-year, 2-year, 1-year, and 0-year classification. We created retention models for each level. These models are validated by each independent corresponding test data sets. The predictive accuracy of 6-year neural network model is 93.05 %. We utilized wavelet decomposition.

In the following sections, we describe the methodology and algorithms.

2 Methodology

2.1 Weka J48 Decision Tree

J48 decision tree is an implementation of the C4.5 algorithm in the WEKA. C4.5 is an algorithm used to generate a decision tree developed by Ross Quinlan. C4.5 is an extension of Quinlan's earlier ID3 algorithm. C4.5 builds decision trees from a set of training data in the same way as ID3, using the concept of information entropy. At each node of the tree, C4.5 chooses the attribute of the data that most effectively splits its set of samples into subsets enriched in one class or on other. The splitting criterion is the normalized information gain (difference in entropy). The attribute with the highest normalized information gain is chosen as a leaf to make the decision [12, 13]. The information entropy and information gain are defined below [10].

$$Entropy(S) = \sum_{j} -p_j \log_2 p_j \tag{1}$$

where p_j is the fraction of j type examples in S.

$$Gain(S,A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v) \tag{2}$$

where Values(A) is the set of all possible values for attribute A, and S_v is the subset of S for which attribute A has value v (i.e., $S_v = \{s \in S | A(s) = v\}$).

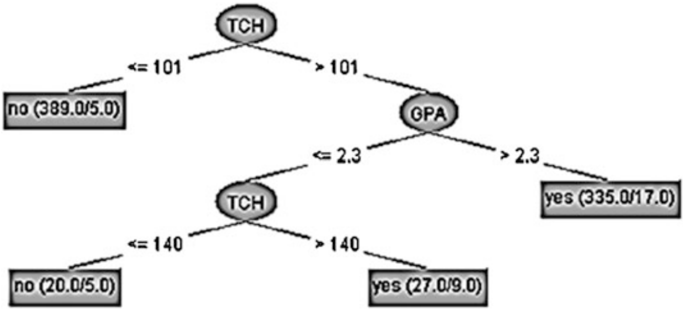


Fig. 1 J48 decision tree

We applied J48 decision tree to the 6-year training data set. The J48 decision tree algorithm classified 94.03 % of the instances correctly and 5.97 % of the instances incorrectly. The Weka created J48 decision tree for the 6-year training data set is shown in Fig. 1. The numbers in (parentheses) at the end of each leaf tell us the number of instances in this leaf. If one or more leaves were not pure (=all of the same class), the number of missing classified instances would also be given, after a slash (/).

2.2 *Weka J48 Decision Tree Pruning*

The goal of a decision tree learning algorithm is to have the resulting tree to be as small as possible to avoid overfitting to the training set [10, 13–15]. However, finding a minimal decision tree (nodes, leaves, or depth) is an NP-hard optimization problem [10]. We created an algorithm to further prune the Weka J48 decision tree to bare minimum decision tree. The algorithm is given below:

Step 1: Prune the deepest node and assign a new leaf as the prediction of a solution to the problem under consideration.
Step 2: Calculate the new tree's estimated accuracy.
Step 3: If the new tree's estimated accuracy is improved, then the pruning is successful; otherwise stop, and exit.

We pruned the Fig. 1 Weka J48 decision tree using above algorithm below:

Step 1: Pruned the parent of two leaves: no(20/5) and yes(27/9) and assigned the new leaf as "no retention (47/23)"
Step 2: Calculated the new tree's estimated accuracy

True (data) to true (pruned tree) $= (335 - 17) = 318$
False (data) to false (pruned tree) $= ((389 + 47) - ((27 - 9) + 5 + 5))$ $= 408$

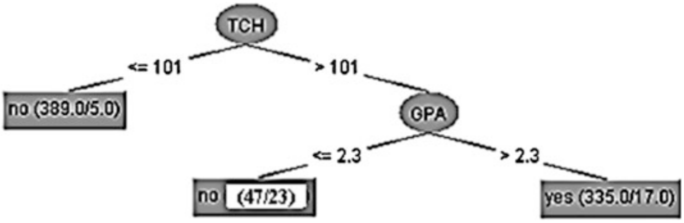


Fig. 2 Bare minimum decision tree

Pruned tree correct = 318 + 408 = 726
False (data) to true (pruned tree) = 17
True (data) to false (pruned tree) = ((27 − 9) + 5 + 5) = 28
Pruned tree incorrect = 17 + 28 = 45
Pruned tree accuracy = (726/(726 + 45)) * 100 % = 94.16 % > 94.03 % of the Weka J48 decision tree accuracy, therefore, the pruning is successful and the heuristic bare minimum decision tree is shown in Fig. 2.

The target function depicted in a decision tree can be represented as a first order logic rule by following each path in the tree from the root to a leaf and creating a rule with the conjunction of tests along the path as an antecedent and the leaf label as the consequent [10]. The target function in the bare minimum decision tree (Fig. 2) can be expressed as a first order logic rule as given in (3).

$$((TCH > 101) \cap (GPA > 2.3)) \rightarrow Retention \tag{3}$$

2.3 Six-Year Neural Network Model

We used (3) derived from the bare minimal decision tree to build a two-layer Neural Network that can predict the HBCU undergraduate student 6-year retention as illustrated in Fig. 3.

For example if a student's last cumulative GPA is 2.4 and total credit hours taken is 102, then the two-layer neural network can determine whether the student will remain in school using the following calculation:

$$\begin{aligned}
&\textit{First Layer}\\
&2.4 * 0.43478 = 1.04 > 1 \rightarrow 1\\
&102 * 0.0099 = 1.01 > 1 \rightarrow 1\\
&\textit{Second Layer}\\
&1 * 0.50001 + 1 * 0.5 = 1.00001 > 1 \rightarrow 1(Retention)
\end{aligned}$$

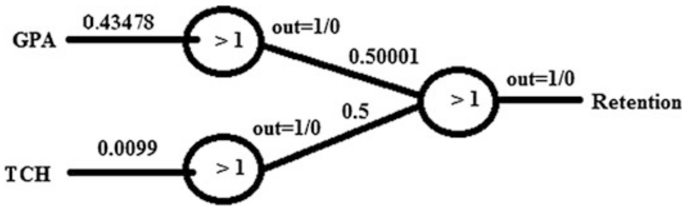


Fig. 3 Six-year training neural network model (GPA—input, TCH—input)

2.4 Improved Model Accuracy

We also created an algorithm to improve the two-layer neural network model's accuracy shown below:

Step 1: Test neural network model using the 6-year training data set
Step 2: Put the 6-year training data set and neural network model output values to an array
Step 3: Sort the array by the model output values
Step 4: On the boundary of 1/0 of the model output values adjust the GPA and TCH weight with $W_0 \pm \Delta\omega$
Step 5: Calculate the new model's estimated accuracy
Step 6: If the new model's estimated accuracy is improved then the adjustment is successful; otherwise stop, and exit.

We used above algorithm and adjusted the weight of input GPA from 0.43478 to 0.437255, which improved the 6-year neural network model's accuracy shown in Fig. 4.

The improved 6-year neural network shown in Fig. 4 was tested against the 771 training data set. The results showed that 728 are true and 43 are false, and the model's accuracy is 94.42 %. The detail is shown in the following:

True (data) to true (model) = 321
False (data) to false (model) = 407
Model correct = 321 + 407 = 728
False (data) to true (model) = 18
True (data) to false (model) = 25
Model incorrect = 18 + 25 = 43
Model accuracy = (728/(728 + 43)) * 100 % = 94.42 %.
Model in-sample error = (43/771) * 100 % = 5.58 %.

2.5 SVM Classification

Support vector machine is a supervised learning algorithm and it has the three following properties [16]:

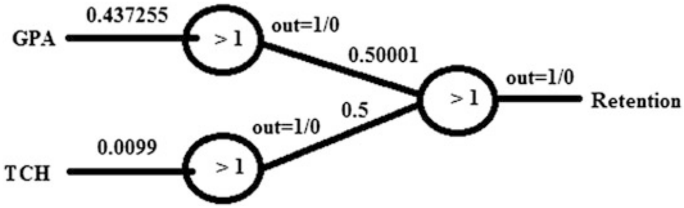


Fig. 4 Improved 6-year neural network model

(1) SVM constructs a maximum margin separator—a decision boundary with the largest possible distance to example points.

(2) SVM creates a linear separating hyperplane, but it has the ability to embed the data into a higher-dimensional space, using the so-called Kernel Trick. This means the hypothesis space is greatly expanded over methods that use strictly linear representations.

(3) SVM is a nonparametric method. It retains training examples, and potentially needs to store them all. In practice, it often ends up retaining only a small fraction of the number of examples; sometimes as few as a small constant times the number of dimensions. The SVM combines the advantages of nonparametric and parametric models: they have the flexibility to represent complex functions, but they are resistant to overfitting.

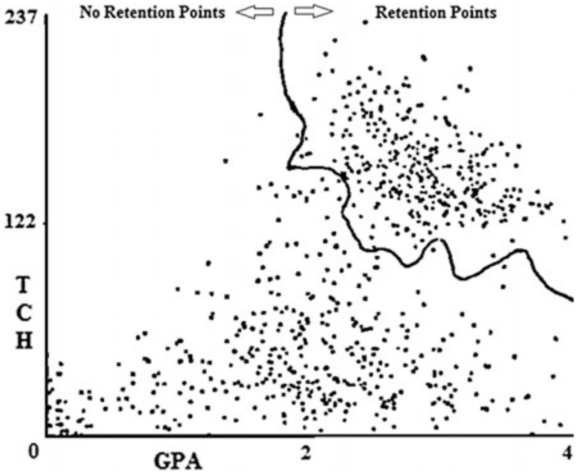
To prove the validity of our model, that the last cumulative GPA and total credit hours (TCH) taken are the major factors which affect the HBCU undergraduate student retention, we use the 6-year training data set to model the retention by SVM algorithm. These points are shown in the normal space with a curve boundary in Fig. 5. The points above the curve correspond to “retention” and the ones below the curve correspond to “no retention.” The x-axis is 6-year cumulative GPA and the y-axis is 6-year total credit hours taken.

We mapped the data from the normal space x into a z space using the following transformed function $\phi(x)$ (Kernel function) [17–19].

$$x = \begin{bmatrix} 1 \\ GPA \\ TCH \end{bmatrix}$$
$$z = \phi(x) = \begin{bmatrix} 1 \\ GPA^2 \\ TCH^2 \end{bmatrix} \tag{4}$$

In the z space, the separating curve is changed to a line, and we created the retention regression as (5).

Fig. 5 The normal data set space



$$\frac{21857.5}{19.5} GPA^2 + TCH^2 - 21857.5 > 0 \tag{5}$$

The model has been tested and the model's accuracy is 93.64 %.

We also applied the improvement performance algorithm to the above model and created the following (6):

$$\frac{20800}{21} GPA^2 + TCH^2 - 20800 > 0 \tag{6}$$

The new model's accuracy is improved to 94.29 %.

2.6 Retention Using Wavelet Decomposition on GPA

We applied linear smoothing to the retention discrete GPA signals. The algorithm of linear smoothing is shown below [20].

$$\begin{aligned} y_1 &= \frac{x_1 + x_2 + x_3}{3} \\ y_2 &= \frac{x_2 + x_3 + x_4}{3} \\ &\ldots\ldots \end{aligned} \tag{7}$$

where $x_1, x_2, x_3 \ldots$ are the original GPA data, and $y_1, y_2, y_3 \ldots$ are the new GPA data.

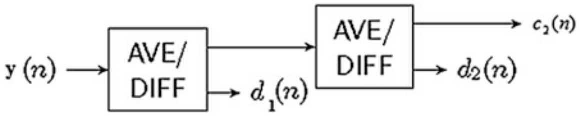


Fig. 6 The Haar retention GPA processes

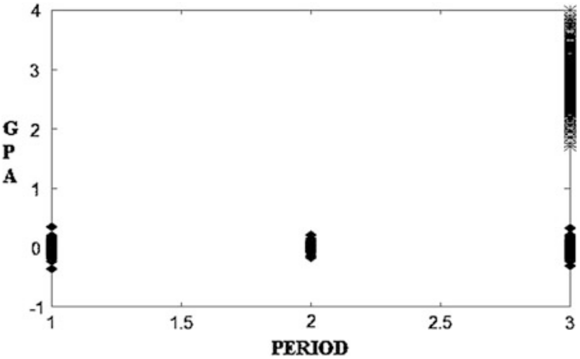


Fig. 7 Haar retention GPA representation

After the retention discrete GPA signals have been smoothed by linear filters, we used Haar transform processing. Haar transform' decompositions can be written as below [21].

$$\begin{aligned} c(n) &= 0.5 * y(2n) + 0.5 * y(2n+1) \\ d(n) &= 0.5 * y(2n) - 0.5 * y(2n+1) \end{aligned} \tag{8}$$

where c(n) is average of the pairs of discrete GPA signals, and d(n) are their differences.

The retention waves' Haar processes are shown in Fig. 6.

Where $d_1(n)$ is the first level decomposition GPA difference, $d_2(n)$ is the second level decomposition GPA difference, and $c_2(n)$ is the second level decomposition GPA average, and the retention GPA representation is shown in Fig. 7. The average GPA points are shown as star on the top, and the difference GPA points are shown as rhombus on the bottom [20, 21].

By using the results from the second level Haar transform over the entire student population, we computed the average GPA and the difference for retention students.

From the results of Haar transform processing, we can say that the average GPA for the HBCU undergraduate student retention should be 2.8597, and the average difference should be −0.023307.

Table 3 Training data set accuracies

Accuracy (%)							
Number of year	0-year	1-year	2-year	3-year	4-year	5-year	6-year
J48	59.66	66.54	81.19	85.86	89.49	91.44	94.03
Simple logistic	65.24	68.61	80.16	85.21	90.66	92.09	94.03
Naïve Bayes	60.57	67.06	78.47	84.96	88.72	91.31	93.39
JRip	59.27	67.06	80.16	84.57	90.92	91.70	94.03

Table 4 Retention models by year

Accuracy (%)							
Number of year	0-year	1-year	2-year	3-year	4-year	5-year	6-year
J48	59.66	66.54	81.19	85.86	89.49	91.44	94.03
Neural network	59.92	68.35	81.71	86.64	90.53	92.48	94.42
Improved	0.26	1.81	0.52	0.78	1.04	1.04	0.39

3 Results

3.1 Training Data Sets Results

Based on the HBCU 6-year (2006–2011) training data set, we also collected training data sets corresponding to 5-year (2006–2010), 4-year (2006–2009), 3-year (2006–2008), 2-year (2006–2007), 1-year (2006), and 0-year (which used high school GPA to replace the undergraduate academic data) periods for the Fall 2006 cohort students [2, 11, 22]. We applied several Weka algorithms, J48 decision tree, Simple Logistic, Naïve Bayes, and JRip on the seven training data sets. The results are shown in Table 3.

We applied the pruning algorithm to the seven J48 decision trees, performance improvement algorithms on predictive models, and then we created seven different retention neural networks models by the seven student academic training data sets. The new results are shown in Table 4.

The HBCU undergraduate student retention results by seven student academic levels are shown in Table 5.

3.2 Test Data Sets Results

After the HBCU retention models are created by the seven student academic levels training data sets, we further collected test data sets from the Fall 2007 full-time and first-time undergraduate student, and tracked these students' activities in the following 6 years from Fall 2007 to Fall 2012. The data was also queried from the

Table 5 The training retention results by year

Number of year	0-year	1-year	2-year	3-year	4-year	5-year	6-year
GPA		2.391	2.04	2.443	2.333	2.292	2.3
TCH			36	60	81	106	101
Distance (mile)	6.9	5.2					
HS GPA	2.28						

Table 6 Test data set numeric attributes

Naïve Bayes	No retention		Retention	
Attribute name	Mean	Std. Dev.	Mean	Std. Dev.
GPA	1.8364	±0.9129	2.9063	±0.4445
TCH	38.8145	±27.4391	125.7639	±18.2481
Distance	6.9231	±10.8022	4.67	±8.2694
Age	18.1685	±0.5397	18.0756	±0.3778
FINAID	9337.6	±7080.19	9886.25	±6765.68
SAT I math	390.75	±134.27	406.15	±120.36
SAT I verb	405.52	±136.80	420.1	±117.72

Table 7 Test data set accuracies

Accuracy (%)							
Number of year	0-year	1-year	2-year	3-year	4-year	5-year	6-year
J48	52.2	62.9	80.2	85.1	88.1	91.7	93.8
Simple logistic	59.3	63.9	76.6	83.4	88.9	92.9	93.8
Naïve Bayes	57.7	64.5	74.8	83.5	88.2	91.6	93.5
JRip	59.2	65.9	78.9	82.9	88.7	91.8	94.3

Campus Solution database. The 6-year test data set size is 820 instances with 12 attributes, same as training data set attributes, which are shown in Table 1. The HBCU undergraduate 6 years retention rate 43.5 % was derived from the 6-year test data set. We also collected test data sets for 5-year (2007–2011), 4-year (2007–2010), 3-year (2007–2009), 2-year (2007–2008), 1-year (2007), and 0-year (which used high school GPA to replace the undergraduate academic data) periods of the Fall 2007 cohort students. The 6-year test data set numeric attributes and statistics are shown in Table 6.

We applied the same Weka algorithms on the seven test data sets. The results are presented Table 7.

The HBCU undergraduate student retention test data set' results by year are shown in Table 8.

Table 8 The test data set retention results

Number of year	0-year	1-year	2-year	3-year	4-year	5-year	6-year
GPA		2.545	2.2	2.111	2.3	2.381	
TCH			34	49	69	76	99
HS GPA	2.78						

Table 9 Six-year predictive results

Predictive value	Correct	Error	Total
Retention	314	14	328
No-retention	449	43	492
Total	763	57	820
Percentage (%)	93.05	6.95	100

Table 10 Models' predictive accuracies

Accuracy (%)							
Number of year	0-year	1-year	2-year	3-year	4-year	5-year	6-year
Model	59.92	68.35	81.71	86.64	90.53	92.48	94.42
Predicted test data	55.85	64.88	79.88	82.44	86.71	87.93	93.05
Difference	4.07	3.47	1.83	4.20	3.82	4.55	1.37

3.3 Retention Models' Validation

We used the 6-year training neural network model shown in Fig. 4 to predict the 6-year test data set and the model predicted correct students are 763, and errors are 57. The model's predicted accuracy is 93.05 % and the model's out-of-sample error is 6.95 % shown in Table 9. The out-of-sample error (6.95 %) is close to the model in-sample error (5.58 %), which was calculated earlier. We validated our models by the seven student academic levels of test data sets.

The summary of the predictive accuracies for the seven neural networks is shown in Table 10.

The summary of the predictive errors for the seven neural networks is shown in Table 11.

We used the following equation to validate the seven predictive models [17].

$$E_{out}(g) - E_{in}(g) < = \sqrt{\frac{1}{2N}\ln\frac{2M}{\delta}} \tag{9}$$

Table 11 Models' predictive errors

Error (%)							
Number of year	0-year	1-year	2-year	3-year	4-year	5-year	6-year
In-sample error	40.08	31.65	18.29	13.36	9.47	7.52	5.58
Out-of-sample error	44.15	35.12	20.12	17.56	13.29	12.07	6.95
Difference	4.07	3.47	1.83	4.20	3.82	4.55	1.37

where $E_{out}(g)$ is the model's out-of-sample error, $E_{in}(g)$ is the model's in-sample error, N is the size of training data set, and M is the size of test data set, δ is tolerance.

The size of training data set = 771, the size of test data set = 820, and the tolerance $\delta = 0.05$, then the Eq. (9) became (10).

$$E_{out}(g) - E_{in}(g) < = \sqrt{\frac{1}{2N} \ln \frac{2M}{\delta}} \approx \sqrt{\frac{1}{2 * 771} \ln \frac{2 * 820}{0.05}} = 8.21\,\% \qquad (10)$$

Based on Eq. (10) the difference in errors $(E_{out}(g) - E_{in}(g))$ should be less than 8.21 %. The differences in errors shown in Table 11 are less than 8.21 %, so the seven retention predictive models are validated by the seven test data sets.

4 Conclusions

The goal of a decision tree learning algorithm is to have the resulting tree to be as small as possible, per Occam's razor [10]. The Weka J48 decision tree is not a minimum decision tree. We further pruned it, improved the estimated accuracy, and simplified the learning rules for the HBCU undergraduate student retention. This is an effective way to find the most important factors that affect the retention and then to build the simplifying retention models. Occam's razor is the machine learning principle, where the "razor" is meant to trim down the explanation to the bare minimum that is consistent with the data. The simplest model that fits the data set is also the most plausible.

After the retention model was created, we used learning feedback based performance improvement algorithm to improve the neural networks models' accuracy.

We studied the HBCU undergraduate student retention in the 6 years period and split the 6 years to seven student academic levels. We classified and created retention models for each level, and then we validated the models by seven independent corresponding test data sets. The 6-year retention model's out-of-sample error is 6.95 %, which is close to the in-sample error 5.58 %.

The SVM is currently the most popular approach for retention supervised learning. For the nonlinear SVM boundary, we used transformed function (Kernel function) to change the normal space x to a z space for linear separation, and then we created a retention regression. This is an effective way to directly create a retention regression without using any machine learning tool such as Weka. The SVM retention model used two significant attributes GPA and TCH, and the model's accuracy was improved to 94.29 %.

The retention analysis using Haar transform used GPA only as opposed to the other methods used in this paper. In the Haar decomposition approach the average and difference of total credit hours (TCH) has no significance in retention. From the results of Haar transform processing, we can say that the average GPA for the HBCU undergraduate student retention should be 2.8597, and the average difference should be −0.023307.

References

1. D.B. Stone, African-American males in computer science – examining the pipeline for clogs, The School of Engineering and Applied Science of the George Washington University, Thesis, 2008
2. C.H. Yu, S. Digangi, A. Jannasch-pennell, C. Kaprolet, A data mining approach for identifying predictors of student retention from sophomore to junior year. J. Data Sci. **8**, 307–325 (2010) (ISSN 1680-743X)
3. S.K. Yadav, B. Bharadwaj, S. Pal, Mining educational data to predict student's retention: a comparative study. Int. J. Comput. Sci. Inf. Secur. **10**(2), 113–117 (2012) (ISSN 1947-5500)
4. A. Nandeshwara, T. Menziesb, A. Nelson, Learning patterns of university student retention. Expert Syst. Appl. **38**(2), 14984–14996 (2011)
5. S.A. Kumar, M.V.N., Implication of classification techniques in predicting student's recital. Int. J. Data Min. Knowl. Manage. Process (IJDKP), 1(5), 2011
6. D. Kabakchieva, Student performance prediction by using data mining classification algorithms. Int. J. Comput. Sci. Manage. Res. **1**(4), 686–690 (2012)
7. S. Lin, Data mining for student retention management, in *The Consortium for Computing Sciences in Colleges* (2012)
8. S. Singh, V. Kumar, Classification of student's data using data mining techniques for training and placement department in technical education, Int. J. Comput. Sci. Netw. (IJCSN) 1(4) (2012)
9. J. -W. Jia, M. Mareboyana, Machine learning algorithms and predictive models for undergraduate student retention. Lecture notes in engineering and computer science, in *Proceedings of the World Congress on Engineering and Computer Science 2013, WCECS* 2013, San Francisco, pp 222–227, 23–25 Oct, 2013
10. T. Mitchell, *Machine Learning* (McGraw Hill, 1997). ISBN 0070428077
11. S.L. Hagedorn, How to define retention: a new look at an old problem, in *College Student Retention: Formula for Student Success*, ed. by Alan Seidman (Praeger Publishers, Westport, 2005)
12. M.H. Dunham, *Data Mining Introductory and Advanced* pp 1–124, (Prentice Hall, Upper Saddle River 2003). ISBN 0-13-088892-3
13. E. Frank, Pruning Decision Trees and Lists, Department of Computer Science, University of Waikato, Thesis, 2000
14. F. Esposito, D. Malerba, G. Semeraro, A comparative analysis of methods for pruning decision tree. IEEE Trans. Pattern Anal. Mach. Intell. **19**(5), 476–491 (1997)

15. D.D. Patil, V.M. Wadhai, J.A. Gokhale, Evaluation of decision tree pruning algorithms for complexity and classification accuracy. Int. J. Comput. Appl. **11**(2), 23–30 (2010) (0975–8887)
16. S. Russell, P. Norvig, *Artificial Intelligence, a Modern Approach*, 3rd edn. (Pearson, 2010). ISBN-13: 978-0-13-604259-4, ISBN-10: 0-13-604259-7
17. Y.S. Abu-Mostafa, M. Magdon-Lsmail, H. Lin, *Learning from Data* (AMLbook, 2012). ISBN 10:1-60049-006-9
18. N. Stanevski, D. Tsvetkov, Using support vector machines as a binary classifier, in *International Conference on Computer Systems and Technologies – CompSys Tech'* (2005)
19. S. Sembiring, M. Zarlis, D. Hartama, E. Wani, Prediction of student academic performance by an application of data mining techniques, in *International Conference on Management and Artificial Intelligence IPEDR*, vol. 6 (2011)
20. S.D. Stearns, D.R. Hush, *Digital Signal Processing with Examples in MATLAB*, 2nd edn. (CRC Press, Boca Raton, 2011)
21. I.W. Selesnick, *Wavelet Transforms – A Quick Study*, Physics Today magazine (2007)
22. R. Alkhasawneh, R. Hobson, Modeling student retention in science and engineering disciplines using neural networks, in *IEEE Global Engineering Education Conference* (EDUCON) (2011), pp. 660–663

Chapter 25
Mobile Adaptive Tutoring System with Formative Assessment

Hisashi Yokota

Abstract Smartphones and tablets are becoming major learning tools in higher education. But developing educational software for smart phones and tablets based on iOS, Android, and Windows operating systems are time consuming task. Since smart phones and tablets are equipped with the web browsers, instead of developing educational software for all different operating systems, it is only necessary to develop educational web application. This way, all the effort for developing educational software can be concentrated on how to evaluate learner's level of understanding. In this paper, how to generate a formative assessment using educator's knowledge structure map and how to implement it as an web application is shown.

Keywords Adaptive tutoring system · Educator's knowledge structure map · Formative assessment · Knowledge score · Relative distance · Web application

1 Introduction

One of the most commonly used communication devices for which college students own in 2014 is either a smartphone or a tablet. This means that for those of us used to develop educational learning system for PC have to reconsider about developing new educational learning system for smartphones or tablets. There are iOS based smart phones and tablets, Android based smart phones and tablets, and Windows based smart phones and tablets. Thus there are essentially three different operating systems to study to create an educational learning system. This becomes a burden for a college professor who is willing to develop an educational learning software for students.

One solution to this problem is to develop an educational learning web application. For all of these devices have web browsers to run web application. Therefore,

H. Yokota (✉)
Department of Mathematics, College of Engineering, Shibaura Institute of Technology,
307 Fukasaku, Minuma-ku, Saitama 337-8570, Japan
e-mail: hyokota@shibaura-it.ac.jp

H. K. Kim et al. (eds.), *Transactions on Engineering Technologies*,
DOI: 10.1007/978-94-017-9115-1_25,

any adaptive tutoring systems runnable on web browser can become adaptive tutoring systems for mobile learning. Furthermore, many integrated development environments for an adaptive tutoring system runnable on web browser are available. This suggests us that instead of developing an adaptive tutoring system for each of communication devices with different operating systems, it is more practical to develop an adaptive tutoring system runnable on web browser.

To develop an adaptive tutoring system runnable on web browser, ASP.NET web application (visual c#) is adopted as an integrated development environment. Since the functions such as automatic question generation, automatic determination of correctness of a learner's inputted expression, calculation of the relative distance which are essential part of forming an adaptive tutoring system are implemented in the software called JCALC [7], it is natural decision for us to choose ASP.NET web application. Note that those functions developed for JCALC can be put into dynamic link library so that they can be reused for web application.

It is noted in [3] that any educational software needs to give a quick feedback to encourage a learner to study more. To develop an adaptive tutoring system with formative assessment runnable on web browser, it is necessary for the system to be equipped with the ability to diagnose a level of understanding of each learner by examining their inputted answer. Furthermore, it is necessary for the system to be equipped with the ability to choose the most appropriate formative assessment.

In order to equip with such abilities, educator's knowledge structured map is studied in detail and utilized to diagnose a level of understanding for each student. Then by tracing the edges of educator's knowledge structured map with evaluated value, the learner's concept map is created as a subset map. Furthermore, by preparing a formative assessment for an each edge of the learner's concept map, it becomes possible to develop a mobile adaptive mathematics tutoring system with formative assessment.

Finally, it is shown that how the mobile adaptive tutoring system with formative assessment is implemented into our web application called webCalc.

2 How to Build Experienced Mathematics Educator's Knowledge Structure

2.1 Examining Experienced Mathematics Educator's Knowledge Structure

In an article [7], how 10 experienced mathematics educators in my school answered the question like the followings are examined:

(1) If a student writes $2(x^2+3x)^3(2x+3)$ as the answer to the question of "Find the derivative of $(x^2+3x)^4$". How do you assess the student's level of understanding?

(2) If a student writes $-\sin(3x+1)$ as the answer to the question of "Find the derivative of $\cos(3x+1)$". How do you assess the student's level of understanding?

(3) If a student writes $-2/(x+1)^2$ as the answer to the question of "Find the derivative of $(x-1)/(x+1)$". How do you assess the student's level of understanding?

(4) If a student writes $xe^x + x + c$ as the answer to the question of "Evaluate $\int xe^x dx$". How do you assess the student's level of understanding?

(5) If a student writes $\det\begin{pmatrix} -1 & 0 \\ 1 & 2 \end{pmatrix}$ as the answer to the question of "Find the (1, 2) minor of $\begin{pmatrix} 1 & 3 & 2 \\ 2 & -1 & 0 \\ 0 & 1 & 2 \end{pmatrix}$" How do you asses the student's level of understanding?

(6) If a student writes $\det\begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$ as the answer to the question of "Find the (1, 2) cofactor of $\begin{pmatrix} 1 & 3 & 2 \\ 2 & -1 & 0 \\ 0 & 1 & 2 \end{pmatrix}$". How do you asses the student's level of understanding?

(7) If a student writes $\det\begin{pmatrix} -1 & 0 \\ 1 & 2 \end{pmatrix}$ as the answer to the question of "Find the cofactor expansion of $\begin{pmatrix} 1 & 3 & 2 \\ 2 & -1 & 0 \\ 0 & 1 & 2 \end{pmatrix}$ along the 1st row". How do you asses the student's level of understanding.

For the question (1), most of my colleagues agreed that this student knows how to differentiate the cosine function. But he/she probably does not know how to apply the chain rule. This information suggested us that if two different values are obtained when two different expressions are evaluated, comparing with the correct value, it may be possible to tell which student has better understanding for simple differentiation and differentiation with chain rule. Furthermore, this information suggested us that it is important to tell the student to go over the chain rule.

For the question (2), most of my colleagues agreed that this student knows the derivative of the sine function is the cosine function. But this student has no idea about the chain rule. This information suggested us that if two different values are associated with two different expressions, by checking the values associated with the expressions, it becomes possible to tell this student to learn how to apply the chain rule as a feedback.

For the question (3), most of my colleague agreed that this student knows how to differentiate the quotient of polynomials. But this student somehow memorized the quotient rule in the wrong way. This suggest us that it is important to tell the student to memorize the quotient rule correctly as a feedback.

For the question (4), most of my colleague agreed that this student knows about the integration by parts. But the person did not apply the integration by parts correctly. So, the person's knowledge about integration by parts is not enough. This suggests us that it is possible to tell where the student made mistake. Furthermore, by knowing where the mistake has occurred, it is possible to suggest the student which part of the solution is wrong.

For the question (5), most of my colleagues responded by saying that this student knows that the minor of a matrix is given by the determinant. But the person does not know how to find it. This suggest us to tell the student to study the determinant as a feed-up.

For the question (6), most of my colleagues responded by saying that this student may know a little bit about cofactor. But forgetting a sign means that the person's knowledge about cofactor is not enough. This suggest us to ask the student to solve the same type of questions again.

For the question (7), most of my colleague responded by saying that this student has no idea about cofactor expansion. This suggest us that the student must go over the section of cofactor expansion.

2.2 What Is Needed to Reproduce Experienced Mathematics Educator's Response

The idea of reproducing the experienced mathematics educators' responses on web browser by using the experienced mathematics educators' concept map is studied. To explain this idea, consider the following question and answer, and educator's response:

- A student writes $2(x^2 + 3x)^3(2x + 3)$ as the answer to the question of "Find the derivative of $(x^2 + 3x)^4$".
- This student knows how to differentiate the cosine function. But this person probably does not know how to apply the chain rule.

To produce the comments similar to the experienced mathematics educator's responses on web browser, the developing software must be equipped with the ability to read the students' input. Then the developing software must be equipped with the ability to distinguish the right answer from the wrong one. Furthermore, the developing software must be equipped with the ability to tell the student's level of understanding from the student's input.

To achieve the ability to read the student's solution, the string comparison method is one way to do it. But as explained in [8], $\tan x - \sec x + c$ and $2\tan\frac{x}{2}/(1 + \tan\frac{x}{2}) + c$ give rise to the same derivatives. Thus, the string comparison is not good enough. To overcome this problem, one-point criterion is developed in [6].

To achieve the ability to tell the right answer from the wrong one, it is known [6] that the one-point criterion is capable of distinguishing the two different expression.

To achieve the ability to tell the student's level of understanding from the student's input, the first approach was the relative distance defined in [6]. It is capable of finding the simple mistakes such as wrong constant multiple and forgetting constant. Then in [9], educator's knowledge structure was studied and realized that attaching some values on the edges of knowledge structure map may give the ability to infer the student's understanding.

3 Generating Formative Assessment

3.1 Knowledge Score

Now to implement the concept map into webCalc, the "knowledge score" will be introduced. The marks such as 0.2 and 0.3 on edges of the Fig. 1 are called "knowledge score" which indicates that the basic knowledge needed to obtain a correct map. Note that the knowledge scores on the top row adds up to 1, and the knowledge score added vertically adds up to the one of top scores. In other words, to be able to differentiate a composite function, the knowledge about composite function consists of 20 %, the knowledge about the chain rule consists of 50 %, and the knowledge about the basic rules of differentiation consists of 30 %. These percentages are derived by adding the necessary knowledge needed to acquire before completing the top row knowledge.

Using [2, 5], the explained knowledge structure map can be expressed in the Fig. 1 is shown in [8].

3.2 Generating Formative Assessments

How the experienced mathematics educators create the formative assessment is studied. To create the formative assessments, the experienced mathematics educators read the student's solution $2(x^2+3x)^3(2x+3)$. Then the learner's solution to the correct answer is compared. After that whether the derivative of power function is essentially in the right form or not is checked. From this examination, the feedback part of the formative assessment becomes like the following: "You know how to differentiate the power function". Furthermore, the chain rule is applied correctly. This is one example of how the experienced mathematics educators create the formative assessment.

Suppose this time that the question given is "find a derivative of $(x^2+3x)^4$" and the student's solution is $4(2x+3)^3$. Then anyone with calculus teaching

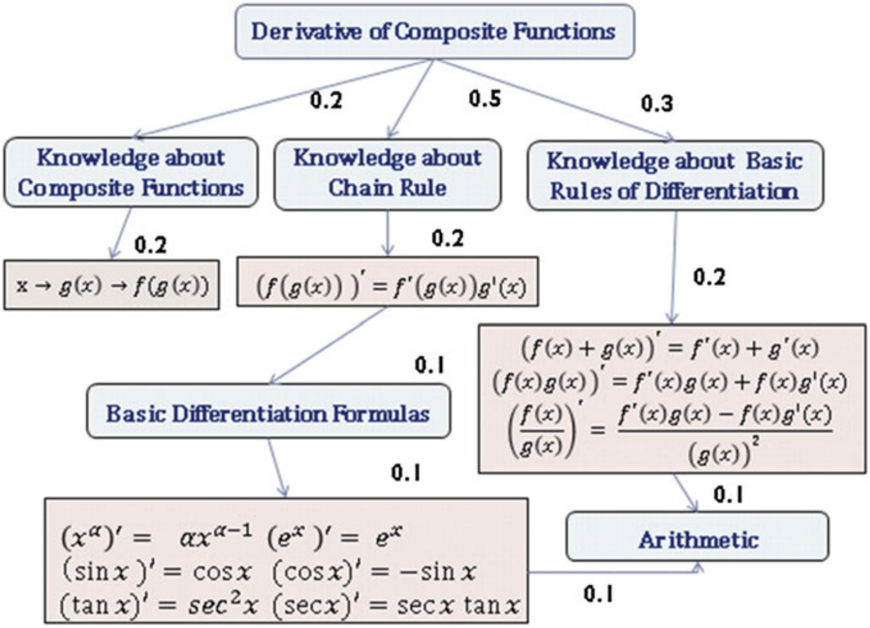
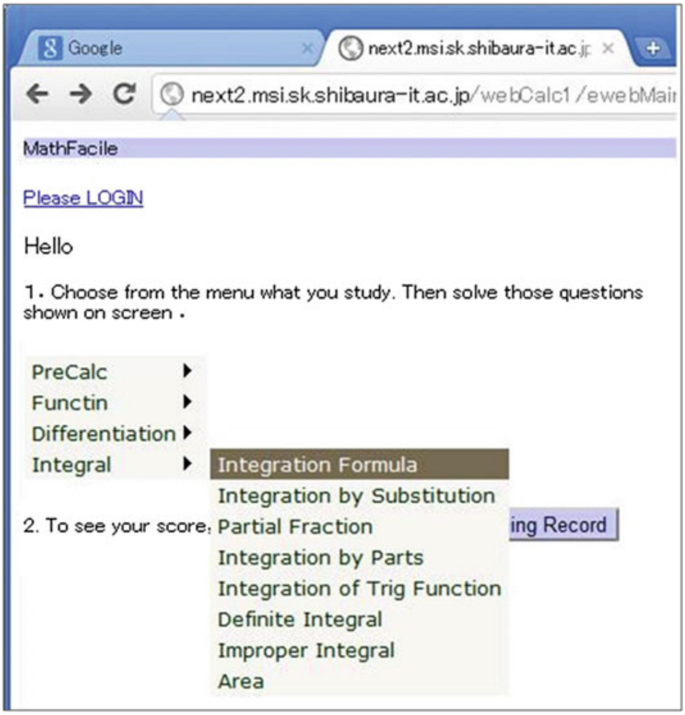

Derivative of Composite Functions
0.2
0.5
0.3
Knowledge about Composite Functions
Knowledge about Chain Rule
Knowledge about Basic Rules of Differentiation
0.2
$x \to g(x) \to f(g(x))$
0.2
$(f(g(x)))' = f'(g(x))g'(x)$
0.2
$(f(x)+g(x))' = f'(x)+g'(x)$
$(f(x)g(x))' = f'(x)g(x)+f(x)g'(x)$
$\left(\frac{f(x)}{g(x)}\right)' = \frac{f'(x)g(x)-f(x)g'(x)}{(g(x))^2}$
0.1
Basic Differentiation Formulas
0.1
$(x^\alpha)' = \alpha x^{\alpha-1}$ $(e^x)' = e^x$
$(\sin x)' = \cos x$ $(\cos x)' = -\sin x$
$(\tan x)' = sec^2 x$ $(\sec x)' = \sec x \tan x$
0.1
Arithmetic
0.1

Fig. 1 Knowledge structure map with knowledge score

experience would say that this student did not master the rule of derivative of composite functions. It is because the derivative of the inside function is taken before the derivative of the power function. This time it is not easy to design the system to judge the same way as the experienced educator. For learners inputs vary many ways and it is impossible to cover all.

To go on, why adaptive system is useful should be reconfirmed. Well known effective educational model for less prepared learners is one-on-one tutoring [1]. Why one-on-one is effective can be explained by many experiments. The behavior of human tutors is studied in [6] to implement into an adaptive tutoring system. There the importance of formative assessment is not mentioned. But the adaptive tutoring system created certainly includes formative assessment.

To generate the similar comments as the experienced mathematics educators, the value of the learner's input is evaluated using one-point criterion. Then the value of $2(x^2+3x)^3(2x+3)$ is different from the value of the machine generated solution $4(x^2+3x)^3(2x+3)$. Then the relative distance is calculated to give 2 which is simple integer. Thus, the machine generates the formative assessment like "You must be careful". Furthermore, the formative assessment like the following is given "You know how to differentiate the composite function and how to apply the chain rule. But somehow made a mistake multiplying by 2 instead of multiplying by 4".



Hello

1. Choose from the menu what you study. Then solve those questions shown on screen .

2. To see your score,

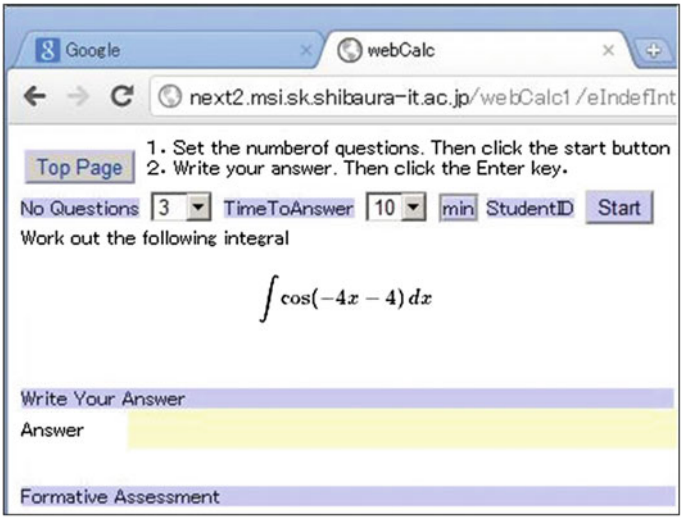
Fig. 2 Web application webCalc

Now notice that every experienced mathematics educator used the chain rule and the derivative of the power function to produce the formative assessment. Thus, every experienced mathematics educators' knowledge structure is very similar. Even though the knowledge of an individual expert consists of both a cognitive element—the individual's viewpoints and beliefs, and a technical element—the individual's context specific skills and abilities [4], experienced mathematics educators' knowledge structure can be used as the basic knowledge structure about how to solve problems.

4 Implementing Formative Assessment

In the next few figures, how a mobile adaptive tutoring system with formative assessment behaves is shown. webCalc contains not only the materials in one variable for which students study in regular calculus course but the materials in pre-calculus course.

The Fig. 2 shows that the top page of webCalc. Here the learner can login to set his/her login name. Then the learner can choose the subject to study. In the following figure, the subject of Integral and item Integration Formula is chosen.

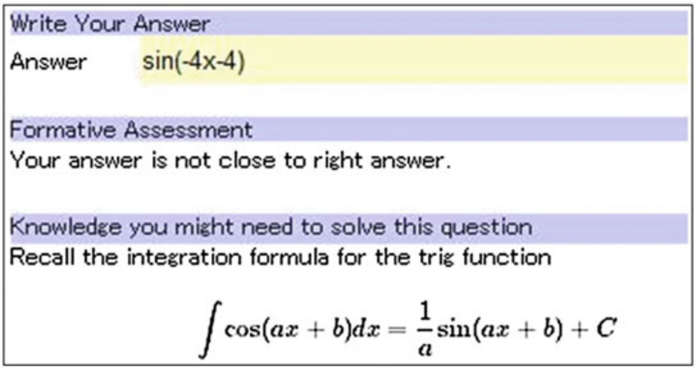


1. Set the numberof questions. Then click the start button
2. Write your answer. Then click the Enter key.

Work out the following integral

$$\int \cos(-4x-4)\,dx$$

Fig. 3 Snap shot of webCalc



Answer sin(-4x-4)

Your answer is not close to right answer.

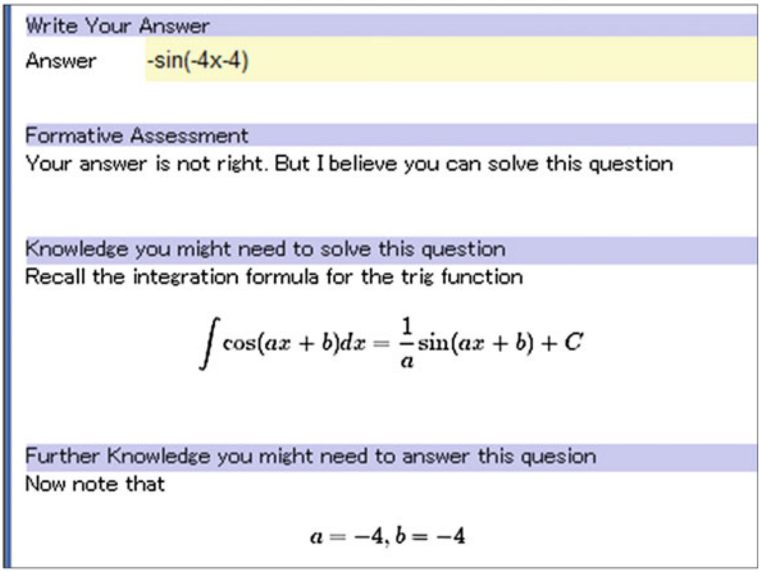
Recall the integration formula for the trig function

$$\int \cos(ax+b)dx = \frac{1}{a}\sin(ax+b)+C$$

Fig. 4 Formative assessment

When the learner pushes the Start button, the question will be shown on web browser. Notice that the question is automatically generated and the mathematical expression is given in nice form. How to generate automatic question is discussed in [6] (Fig. 3).

Suppose that the learner writes $\sin(-4x - 4)$ as in the Fig. 4. Then webCalc responds to give the formative assessment such as “your answer is no close to right answer” and then shows the knowledge needed to solve this question like “Recall the integration formula for the trig function” to let the learner notice his/her mistakes.



Write Your Answer

Answer -sin(-4x-4)

Formative Assessment

Your answer is not right. But I believe you can solve this question

Knowledge you might need to solve this question

Recall the integration formula for the trig function

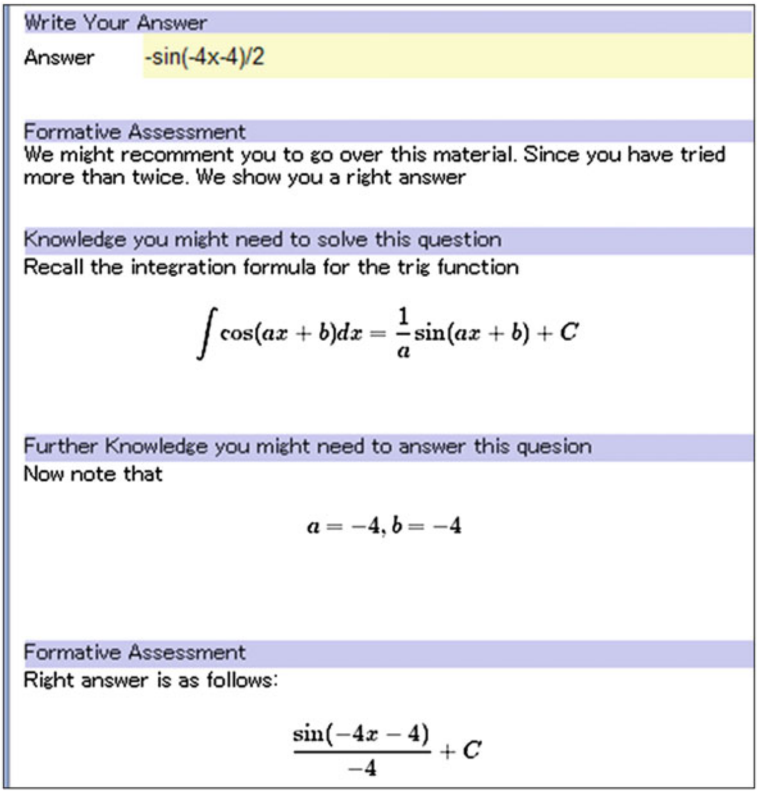
$$\int \cos(ax+b)dx = \frac{1}{a}\sin(ax+b) + C$$

Further Knowledge you might need to answer this quesion

Now note that

$$a = -4, b = -4$$

Fig. 5 Another formative assessment



Write Your Answer

Answer -sin(-4x-4)/2

Formative Assessment

We might recomment you to go over this material. Since you have tried more than twice. We show you a right answer

Knowledge you might need to solve this question

Recall the integration formula for the trig function

$$\int \cos(ax+b)dx = \frac{1}{a}\sin(ax+b) + C$$

Further Knowledge you might need to answer this quesion

Now note that

$$a = -4, b = -4$$

Formative Assessment

Right answer is as follows:

$$\frac{\sin(-4x-4)}{-4} + C$$

Fig. 6 Some other formative assessment

The Fig. 5 shows the different situation. The solution written here is –sin (−4x − 4) instead of sin(−4x − 4). Then webCalc calculates the value of this function and the machine generated function $-\sin(-4x-4)/4$. Furthermore, webCalc adds the knowledge score and responds by saying that "Your answer is not right. But I believe you can solve this question".

In Fig. 6, the answer written is –sin(−4x − 4)/2 which is only differ by the ration of 2 from the correct answer. Then webCalc responds by saying "We might recommend you to go over this material. Since you have tried more than twice. We show you a right answer".

5 Conclusion

The importance of mobile adaptive tutoring system is getting more obvious. Yet it is not easy for educators to develop their own mobile learning system adjusted to their students. One of the major reasons why adaptive tutoring system with summative assessment is losing interest from many educators is the cost performance. The method shown here does not require any marketed software to develop an adaptive tutoring system with formative assessment. Using this method, it is also possible to transform any PC software into mobile learning software.

The system can be used by anyone to access the following URL: http://next2.msi.sk.shibaura-it.ac.jp/webCalc1/ewebMainForm.aspx.

Acknowledgment This work was supported by JSPS KAKENHI Grant Number 25330415. We thank all colleagues and students participated in this project and suggested useful ideas to develop a mathematical intelligent tutoring system with formative assessment.

References

1. M. Alavi, D.E. Leidner, Knowledge management and knowledge management systems: conceptual foundations and research issues. MIS Q. Rev. **25**(1), 107–136 (2001)
2. R.G. Ainsworth, Turning potential school dropouts into graduates: the case for school-based one-to-one tutoring. Research Report 95—07. National Commission for Employment Policy, 35 (1995)
3. P. Bana, Artificial intelligence in educational software: has its time come? Br. J. Educ. Technol. **30**(1), 79–81 (1999)
4. J. Keyes, Where's the ''expert'' in expert systems. AI Expert **5**(3), 61–64 (1990)
5. J. Rentsch, T. Heffner, Group Organ. Manage. **19**(4), 450–474 (1994)
6. H. Yokota, An adaptive tutoring system for Calculus learning, in *Proceedings of The World Congress on Engineering and Computer Science 2009, WCECS 2011*, San Francisco, USA, 20–22 Oct 2009, pp. 640–645
7. H. Yokota, On a development an adaptive tutoring system utilizing educator's knowledge structure, in *Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering and Computer Science 2011, WCECS 2011*, San Francisco, USA, 19–21 Oct 2011, pp. 260–264

8. H. Yokota, On mathematical software equipped with adaptive tutor system. IAENG Trans. Eng. Technol. Lect. Notes Electr. Eng. **170**, 229–238 (2013)

9. H. Yokota, On developing an adaptive tutoring system with formative assessment for mobile learning, in *Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering and Computer Science 2013, WCECS 2013*, San Francisco, USA, 23–25 Oct 2013, pp. 174–177

Chapter 26
A Mentorship Framework for WIL in Software Development

Mosima Anna Masethe and Hlaudi Daniel Masethe

Abstract The major difficulty faced by potential graduates in the software development diploma in South African universities of technology is lack of placement for work integrated learning. Most companies do not have anyone with time to properly mentor a potential graduate. The student challenge is transition from the academic environment to the workplace without graduate attributes, such as employability skills, kaizen theory and productivity theory. The universities on their own cannot prepare a student for the real world of work without engaging external industry partners. The research study provides a mentorship framework to operationalize simulated work integrated learning to incorporate concrete experience, reflective observation, abstract conceptualization and active experimentation in writing applications according to industry outcomes. The research results present a case study of the Tshwane University of Technology in the implementation of the mentorship framework for simulated work integrated learning.

Keywords Attributes · Economy · Learning · Mentorship · Simulation · Technology

1 Introduction

South African University graduates are facing a critical time in history, where there is a shift in the economy and demographic factors influencing competition and pressure amounting on those employed and those looking for employment

M. A. Masethe (✉)
Department of Software Engineering, Tshwane University of Technology eMalahleni Campus, eMalahleni 1035, South Africa
e-mail: masethema@tut.ac.za

H. D. Masethe
Department of Software Engineering, Tshwane University of Technology, Pretoria 0001, South Africa
e-mail: masethehd@tut.ac.za

H. K. Kim et al. (eds.), *Transactions on Engineering Technologies*,
DOI: 10.1007/978-94-017-9115-1_26,

[1, 20]. WIL creates understanding of the work place culture and drives career development and the career choice pathways [2, 20]. It is critical for universities to actively examine and re-curriculate to educate a competitive workforce required for the 21st century in which mentorship models will play a role in simulated work integrated learning. WIL programs have provided training and career development opportunities at previous times of economic recession and served as a license to employment sectors, however, we still have a moderate percentage of students that are not employed or placed [1, 20]. WIL gives students opportunity to actually put their learning into practice in an approved workstation, while students are able to engage and create meaningful knowledge in applied professional setting [2, 20]. However, opportunities for industry placements are limited and moderate in South Africa, especially in the computing sector [2, 20]. The Universities of Technology in South Africa has all introduced WIL in their curricular, and this trend is potentially problematic since work placement is a heavy drain on scarce resources. The WIL activities requires access to quality learning environments, preparations and support, and time investments by potential employers and mentors [3, 20]. The potential employers in South African industries identified that they need graduates with better problem solving skills and overwhelming work experience, which graduates lack [4, 20].

2 WIL Challenges at Tshwane University of Technology

Since most universities of technology have embarked in the implementation of the work integrated program, the students compete for the same position. The challenge is to get an increase in the total number of companies, willing to offer more placements opportunities. Rather than just sharing the same available placements opportunities, as they have been [5]. The responsibility of creating opportunities for students is the sole responsibility of the government and the industry investing in the country, and not the universities [5]. The government, through National Skills Accord, came with the New Growth Path, as a strategy to involve business, organised labour, and community constituency to agree on partnership to achieve placement opportunities, create more internships, and expand on the level of training [6].

The university has a number of common challenges, which includes:

- The skills and experience needed by academic staff to embark in securing placement
- The students that are not capable to advertise their acquired knowledge
- Initiating academic business partnership
- Initiating a memorandum of understanding with industries
- Stakeholder management and analysis
- Common understanding between industry and academic objectives

- The funding framework for work integrated learning, apart from government subsidies linked to the credit value and CESM categories of the subject
- Capacity development for the academic staff to visits industry, and industry placements
- Resources for the site visits and the fieldwork
- Industry simply do not have anyone to mentor the students properly

3 Employability Skills

The researchers [5] raised a concern about the work readiness of graduates that lack generic employability skills, industry is satisfied with discipline specific skills of graduates, but employability skills are under developed. The employability skills are defined as an intelligence to apply knowledge in the work environment [5]. The Tshwane University of Technology signed a memorandum of understanding with Japan International Cooperation Agency (JICA), in order to give a training workshop on Employability Improvement skills to all the staff and the students. The training on the 3i concepts focus on Improvement, Implementation, and Innovation. The workshop concentrate on graduate attributes, such as productivity theory, project management, kaizen theory, value of innovation in economic paradigm, problem thinking skills, intellectual property rights, work readiness, lifelong learning and enterprise collaboration.

It is recommended that Universities of Technology can emphasize WIL program to participate or play a role in advancing employability skills [5]. WIL is also defined as a strategy for enhancing national productivity and addressing growing skill shortages while providing income support for students while building their careers [5].

4 Mentorship

The universities cannot prepare a software engineer for the real world on their own, WIL is needed to complete the student education in the discipline. Unfortunately most companies simply do not have anyone with the time to properly mentor a potential graduate [7]. The best way of learning and integrating software engineering into the workplace is by learning from experience of the other professionals through mentoring processes [8]. Simulated WIL with a mentor as an external partner is described as an innovative approach to combine formal learning at the university and industrial experience in order to prepare software development graduate [9]. Mentorship promotes professional advancement, higher salaries, and higher career satisfaction among potential students willing to learn [8]. Through the mentorship framework, students are able to improve theoretical knowledge and apply in a workplace [8]. The WIL coordinators serve as an internal mentor, and represent the ICT Faculty at the university. They are responsible to direct the students in their industrial work and serve as their academic advisors [9]. The WIL coordinators ratio to student is 1:150 per department, contrary to CHE report [10] which is 1:88 and serve as management interface until the student are ready for industrial environment [9].

The key element in simulated WIL is that each student will be allocated both an internal academic partner and external workplace partners/mentors, in order to solve an authenticated workplace problem, identified either by the student through the community, or provided by the external partners [10]. The student is also provided with orientation to the workplace and the particular problem to solve, while adequate time is given for consultation with the respective clients, or supervisor.

The WIL coordinator provide guidance in the following areas: Technical and non-technical support, where and how to obtain relevant information during the project life span, work ethics and organizational behaviour, verbal and written communication, soft skills, leadership and direction on how to proceed with the project. Whereas the external mentors from industry play a key role in reviewing the student work according to the industry specifications and the outcomes, quality assure processes in software engineering [9].

The student performance is done weekly by the WIL coordinator, with a numeric grade assigned to work done, with expectations as per the software engineering contract ratings. Students also participate in regular peer review evaluations using a 360° self-evaluation forms that track progress and performance in the areas of leadership, communication, and professional behaviour [9].

5 Simulation as a WIL Program

Simulation is a method identified in the running of work integrated learning programs, others being project work, work based placement and virtual WIL [11, 20]. Simulation is a method for preparation and knowledge that can be applied in a number of disciplines [12, 20]. An effective workplace simulation needs to be quality assured and authenticated by the relevant professional body or industry in software development or engineering [13, 20]. Simulation, in this research is defined as representation of reality, as a model of events or processes or items that exists [11, 20]. The recent developments in Information Technology (IT) led to an increase in number of simulations in training conducted within WIL programs, with the most predominant example being simulated work environment and role—plays [11, 20]. Simulation is a technique to substitute and intensify real experiences with guided ones that evoke substantial aspects of real world in a fully interactive way [12, 20].

The universities of technology needs to devise workplace simulations within the context of the curriculum through their advisory committees, by involving the relevant business in their fields [13]. While simulation is a representation of the workplace, the researchers acknowledge that effective learning takes place in a workplace, however, simulation engage students in an educationally purposeful experiences which are self-directed and experiential [5]. The notion of simulated work integrated learning as defined in this study, also incorporate project based learning, as learning involves real projects located in the world of work, which involve an academic supervision and workplace supervisor or mentor [10].

The example of simulated work environment at Tshwane University includes a scientific computer laboratory sponsored by Software Engineering Department, while the role players include Microsoft South Africa as an academic business partner [20] and IBM South Africa [20]. In this case simulated work environment is a reality that focuses on work-place interactions, employer involvement. It is acknowledged that it is not possible to include workplace issues without the external workplace partners [10]. However, the program tackles relevant workplace concerns by introducing the following principles:

- The WIL coordinator attends all academic colloquiums, industry conferences e.g. GOVTECH, ICT INDABA, etc., organised by external partners,
- Subscribe and reads professional journals while encouraging students to do the same
- Invite potential employers to give guest lecturers, including recent graduates, and prominent professionals
- The WIL coordinator arranges site visits for the students and staff in the workplace
- Invite industry representatives or professionals to give awards and participate in project assessment and offer awards to recognise student achievements.
- Industry is invited to give workshops and training for certification on matters related to the discipline (e.g. business process management, training on cognos express, software testing, cloud computing, DB2 certificate, etc.,)

The external partners, such as mentors are invited to suggest workplace problems, scenarios or provide case studies, while acting as the clients to provide authenticated world of practice [10]. The department includes participation from industry in South Africa in their simulated work environment, where mentors from industry play a major role. The success of the simulations is predicted on factors that possess authenticity, planning, structure and support for the team based learning [11, 20]. Simulated work integrated learning cannot occur without external industry partners who represent the knowledge field in the discipline of software engineering. The effectiveness of the framework proposed depend entirely on the commitment of both the academic and professional partner [10].

The simulated WIL program shares its roots in the clinical model, normally employed by the health sciences. This framework requires undergraduate students in software engineering field to practice under a guidance of a software professional from industry and the academic mentor in an authenticated setting unique for software development [9].

The selection of the student is done through an internal advertisement at the university, the WIL coordinator gather the student's resumes and academic transcript for review. The selection is made from the candidates that pass the interview process. The selected candidates do not receive any stipend. However, they can earn a stipend by the community projects that they propose for the external stakeholders like Non-Government Organization, or Non-Profit Organization, if the project is approved by Vodacom. SITA also gives a stipend through Diakanyo project, provided students can propose a system that can help improve service delivery.

The aspects of the external mentoring makes the program unique, and student build competencies and relationship with external entities [9].

6 Research Design

The research is a Case Study within the Tshwane University of Technology with an implementation of the mentorship model, for the simulated aspects of WIL that enables students to experience workplace within an educational framework [20]. The initiatives on simulated work integrated learning are defined within a campus setting that emulates important objectives within the workplace [14, 20]. The software engineering students within the computing sector are taken through an environment and training to design certified mobile applications with a mentor from industry with a goal to develop mobile entrepreneurs within a South African context [20].

The students design and upload the applications to the market place, as part of the training exercise [20]. After mastery and branding of the applications, students will now presents innovative ideas which will lead to a registered patent and a non-disclosure agreement in consultations [20]. The proposed model describes a partnership between students, employers and the university with specified responsibilities for all stakeholders [14, 20]. A simulated work environment allows students to experience a workplace within the educational framework, the proposed model take an approach from an employer involvement with a focus on work place interactions [13, 20]. The purpose of the model is to improve employability and entrepreneurship of the students, and embrace ability to retrieve relevant information, improve communication and presentation skills, planning and problem solving, and improve social development and interaction in a workplace [14, 20].

7 Industry Exposure

Industry exposure is a third year core module in most national diplomas in IT, in most South African Universities of Technology [20]. The module is divided into two sections, Industry Exposure A which is basically aimed to prepare a student for the workplace, while Industry Exposure B brings a student into the workplace environment for authentic experience in industry [20]. The work integrated learning components count 60 credits towards the software development diploma in most universities in South Africa. Students cannot graduate without certifying the requirements for the expected core modules [20]. It's a challenge on its own, as industry is unable to absorb all students, and the only solutions most universities are providing are simulation environment, which cannot be authenticated by industry expects [20]. Software Engineering discipline is a scarce skill in the parlance of the Department of Labour and the Sectoral Education and Training Authorities (SETA) in South Africa, since we have scarcity of the qualified and experienced engineers, and those available do not meet the employment criteria [4]. The researcher [4] confirms that the demand in the software engineering discipline exceeds supply. South Africa comprises of 44.5 % skilled workers, which is far low as compared to other countries like Brazil with 77.3 %, China with 69.1, India with 55 % and Poland with 79.9 % [4].

In order to actively address workforce and education challenges at the Department of Software Engineering, at the Tshwane University of Technology, a mentorship framework was proposed which build upon five stages of work integrated learning: preparations, placement, monitoring, assessment and feedback/reflections [20]. The proposed mentorship framework aims to provide understanding of how a computing professional in software engineering works in practice and equip students with entrepreneurship skills and practical skills to apply in a real world situations [2, 20]. It is most important for software engineering students to understand, accept latest methods and practices in the design of intricate computing systems [15, 20]. The framework includes the improvement of graduate attributes and opportunities to engage with the professional mentor at the relevant industry for the software engineering discipline [13]. The students are encouraged by the entry into the job market or work environment and a career [5]. The proposed framework intends to promote lifetime learning and self-development process [5]. The study operationalised work integrated learning by simulation involving four stages that incorporate concrete experience, reflective observation, abstract conceptualization and active experimentation in writing applications according to industry outcomes [5].

The Tshwane University of Technology adopted a flowchart model on Fig. 1 that depicts five stages of work integrated learning on quality management of work placement [10, 20].

The adoption of the model is met with a number of challenges in bridging the gap between industry and world of work which are: Inappropriate placements of students in the industry, reported to be a time consuming exercise, an insufficient resources to support industry visits and monitoring, poor industry Liaison and Poor stakeholder management.

The industry demands that an investment in learning must be transformed into industrious outcomes that progress the organisation towards defined strategic goals [16, 20]. WIL has potentials, although limited to achieve productive outcomes since it allows students to contextualise study content within the socio-cultural and

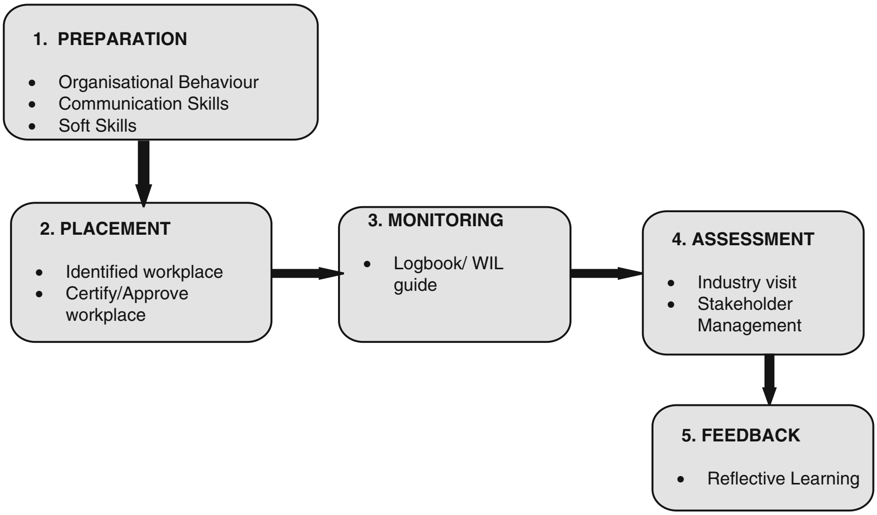


Fig. 1 Quality management work placement learning model

serviceable environment of the workplace [16, 20]. The model above relates challenges of developing industry partnership which is problematic and demands rearrangement of knowledge power relations between academia and industry. Partnership requires academic curriculum to be aligned with work tasks which offers experience for student to focus on outcomes that serve that industry expectations [16, 20]. The alignment is complicated and challenging for universities of technology, as vendor relationship has underlying expectations [16, 20]. Partnership with industry for WIL implies that academics do no longer dominate curriculum matters, but advocate content and theoretical knowledge base, while they rely heavily on the tacit knowledge of the organisations [20].

8 Mentorship Career Model

The proposed model is based upon developing a constructivists approach based on student's active involvement in problem solving and critical thinking based on learning mobile application relevant and engaging to the 21st century [17, 20]. Students construct their own mobile application, upload to the market place and test their ideas based on the number of downloads to their applications. The model introduces the following principles: student undergo training with experienced mentor for a concrete experience, reflect and re-examine through hands on projects, formulate innovative concepts and test in new environments [18, 20]. The model adopts techniques by the Newcastle model which allows students to

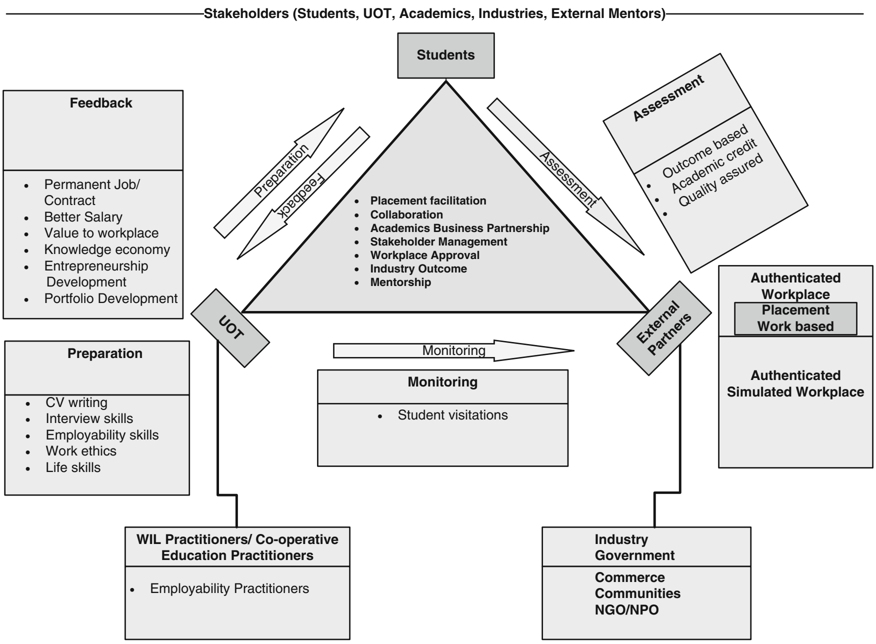


Fig. 2 Mentorship framework for simulated work integrated learning in mobile applications

construct their knowledge, allow holistic and integrated learning, while engaged in experience-based focus [17, 20].

The model enhance its relevance and connections for service delivery issues and problems faced by the communities in South Africa, by introducing an entrepreneurships to solve real-world problems which is a source of genuine creativity and academic importance [19, 20]. The framework encourages all stakeholders to participate with similar motivation, objectives, and a common understanding of the purpose of WIL based on employer objectives, academic objectives, generic skills within the context of the field and specific skills in the discipline [13, 20].

The above mentioned framework on Fig. 2 demonstrates an approach using academic business partnership, where academic manager serve as a mentor and develops a partnership which is based on industry expectations and pre-defined outcomes [20]. The industry assigns a mentor at the university workstation for simulated work integrated learning with pre-defined goals and SMART objectives, which are aligned to the industry requirements, with all required activities [20]. Upon an understanding on the desired outcomes, student register for the module, attends inductions and workshops, experience hands on workplace environment, and develop certified application for the market place as per the industry desired outcomes [20]. The certified applications by the students build a success portfolio

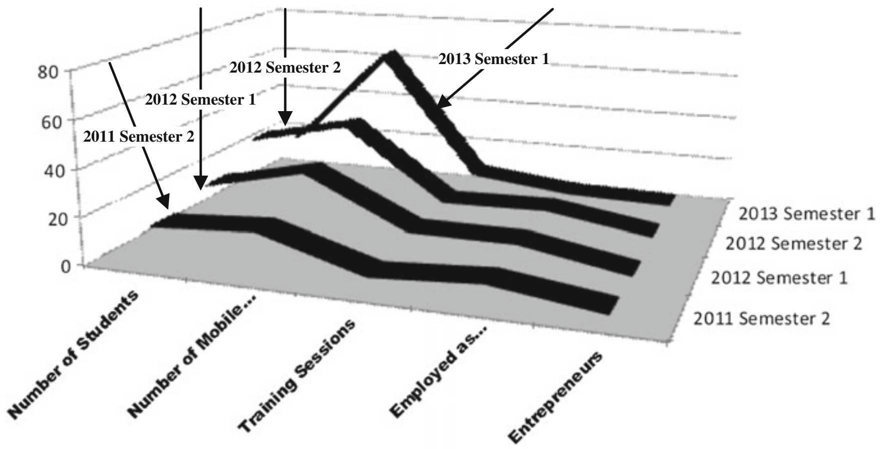


Fig. 3 The student demographics since 2011 semester 2

to prove to employers that students have technical skills relevant to design and deploy software applications. The applications designed by student in this research are certified by Microsoft using industry certification workflow that maintains application quality. The application review result comes through by email notification.

Students will then create innovative ideas for solving service delivery issues for the community, which engages partnership framework for the faculty to relate to the community [19, 20]. Knowledge is required as a point of entry into the software engineering career, an a need arise to provide students with more than just a method and academic problems, but allow time in industry working on real software development projects, and the framework is able to provide that insight [15, 20]. The simulated work environment enhance skills required by industry [12, 20], such as: problem solving skills, decision making skills, communications skills, team based competencies, technical competencies, and functional expertise [20].

The research study shows that the simulated work environment through the proposed framework improves learning, it is effective in developing skills, and help the student prepare to deal with unanticipated industry challenges and events, and increase confidence.

The mentorship framework is validated on the success stories of the students since the adoption. Figure 3 below shows the student demographics since 2011 semester 2 at the Tshwane University of Technology.

The student got an opportunity to experience a simulated work environment within the campus that demonstrates design problem complexities and richness of a mobile development from industry perspective in creating new mobile products [15, 20].

9 Conclusions and Future Work

Mentorship framework in a simulated work environment has opened a new educational paradigm in the application of software engineering. The evidence through the practice can be put into action by WIL practitioners, where challenges in WIL placements arise. The key to the success of the framework is academic business partnership, with integrated learning outcomes decided by industry and academia. The success of simulated work integrated learning activities contributes to maintainable employees for the future economic growth of the country.

The future work to be looked upon may be the software quality management and mathematical modelling for WIL.

Acknowledgment This work was supported by Tshwane University of Technology, South Africa.

References

1. K. Betts, M. Lewis, A. Dressler, L. Svensson, Optimizing learning simulation to support a quinary career development model. Asia Pac. J. Coop. Educ. **10**(2), 99–119 (2009)
2. A. Goold, N. Augar, Using virtual meeting spaces for work integrated learning, in *Proceedings Ascilite Auckland* (2009), pp. 367–371
3. J. Orrell, Work-integrated learning programmes: management and educational quality Janice Orrell*, in *Proceedings of the Australian Universities Quality forum* (2004), pp. 1–5
4. R.C. Daniels, *Skills Shortages in South Africa: A Literature Review* (Development Policy Research Unit, Cape Town, 2007)
5. B. Mclennan, S. Keating, ALTC NAGCAS national symposium work-integrated learning (WIL), in *Australian Universities: The Challenges of Mainstreaming WIL* (2008)
6. E. Patel, Nationals Skills Accord, Pretoria (2011)
7. J. James, Seven tips on mentoring entry-level developers, *techrepublic* (2014). Available http://www.techrepublic.com/blog/software-engineer/seven-tips-on-mentoring/. Accessed 1 Mar 2014
8. M.A. Omazic, D. Bla ekovic, Z. Baracskai, Virtual mentorship program: a development of young professionals and scientists, in *2008 International Conference on Computer Science and Software Engineering*. IEEE Computer Society (2008), pp. 331–336
9. S.L. Hazen and S.T. Freua, *An integrate Embedded Software Engineering Program and Frontiers in Education Conference*, IEEE (1997), pp. 862–866
10. CHE, *Work-Integrated Learning: Good Practice Guide*, Pretoria (2011)
11. H. Vo-Tran, S. Pittayachawan, S. Reynolds, Learning and Teaching Investment Fund 2010, *academia.edu* (2010), pp. 1–32
12. F. Lateef, Simulation-based learning: just like the real thing. J Emerg Trauma Shock **3**(4), 348–352 (2010)
13. G. Patrick, C-j. Peach, D. Pocknee, C. Webb, F. Fletcher, M. Pretto, *A National Scoping Study*. Queensland University of Technology: The WIL [Work Integrated Learning] report: a national scoping study [Australian Learning and Teaching Council (ALTC) Final report], Brisbane, p. 111 (2009)
14. I. Abeysekera, Issues relating to designing a work-integrated learning program in an undergraduate accounting degree program and its implications for the curriculum. Asia Pac. J. Coop. Educ. **7**(1), 7–15 (2006)

15. G. Regev, D.C. Gause, A. Wegmann, Experiential learning approach for requirements engineering education. Requirements Eng. **14**(4), 269–287 (2009)
16. S. Choy, B. Delahaye, Partnerships between universities and workplaces: some challenges for work-integrated learning. Stud. Continuing Educ. **33**(2), 157–172 (2011)
17. J. Gibbons, M. Gray, An integrated and experience-based approach to social work education: the Newcastle model. Soc. Work Educ. **21**(5), 529–549 (2002)
18. S. Ferns, K. Moore, Assessing student outcomes in fieldwork placements: an overview of current practice. Asia Pac. J. Coop. Educ. **13**(4), 207–224 (2012)
19. H. Ishisaka, S. Sohng, Teaching notes: partnership for integrated community-based learning: a social work community—campus collaboration. J. Soc. Work Educ. **40**(2), 321–337 (2004)
20. M.A. Masethe, H.D. Masethe, Mentorship model for simulated work integrated learning using windows phone, in *Lecture Notes in Engineering and Computer Science: Proceedings of the World Congress on Engineering and Computer Science 2013, WCECS 2013*, San Francisco, USA, 23–25 Oct 2013, pp. 212–215

Chapter 27
CLAHE Inspired Segmentation of Dermoscopic Images Using Mixture of Methods

Damilola A. Okuboyejo, Oludayo O. Olugbara and Solomon A. Odunaike

Abstract The overarching objective of this study is to segment lesion areas of the surrounding healthy skin. The localization of the actual lesion area is an important step towards the automation of a diagnostic system for discriminating between malignant and benign lesions. We have applied a combination of methods, including intensity equalization, thresholding, morphological operation and GrabCut algorithm to segment the lesion area in a dermoscopic image. The result shows that the approach used in the study is effective in localizing lesion pixels in a dermoscopic image. This would aid the selection of discriminating features for the classification of malignancy of a given dermoscopic image.

Keywords Automation · CLAHE · Diagnosis · Lesions · Melanoma · Segmentation

1 Introduction

Since 1663 when the first observation of vessels in respect to nail matrix was examined with a microscope by Kohlhaus [1], the field of Dermatology has since advanced through the advent of first Binocular Dermtoscope in 1916 by Zeiss [1]. The introduction of pattern analysis by Pehamberger et al. [2] has inspired the

D. A. Okuboyejo (✉) · S. A. Odunaike
Department of Software Engineering, Tshwane University of Technology,
Pretoria, South Africa
e-mail: okuboyejoda@tut.ac.za

S. A. Odunaike
e-mail: odunaikesa@tut.ac.za

O. O. Olugbara
Department of Information Technology, Durban University of Technology,
Durban, South Africa
e-mail: oludayoo@dut.ac.za

H. K. Kim et al. (eds.), *Transactions on Engineering Technologies*,
DOI: 10.1007/978-94-017-9115-1_27,

acclaimed ABCD-rule: **A**symmetry, **B**order, **C**olor, **D**iameter which was proposed by Stolz et al. [3]. The ABCD-rule has now become the standard that is widely used by most Dermatologists for screening malignancy of melanocytic lesions.

The use of Dermoscope, which is sometimes called Dermatoscope by Dermatologists has recorded a great success [4]. The heavy dependencies on medical practitioners such as Dermatologists for diagnosing medical images is however a major concern for underserved areas where such experts might not be readily available or non-responsive for urgent medical need [5]. Even for areas where Dermatologists are available, their diagnosis interpretation of results obtained from image acquisition devices such as Dermoscope has been characterized with subjectivity and sometimes poor reproducibility [6, 7]. For the past two decades, medical image analysis has seen the application of image processing techniques for providing machine interpretation of medical images for fostering a better objective decision.

Computer-Aided Diagnostic (CAD) systems have greatly contributed positively at each stage of medical image analysis to ensure faster and reproducible diagnosis [5, 8, 9]. In relation to skin disease diagnosis, the automated ordered-steps involved typically include

- Image formation and preprocessing,
- Segmentation of lesion area,
- Extraction and selection of discriminating features,
- Classification of the lesion based on selected morphological features.

This work extends our previous study [5] where we discussed an approach towards automating the diagnosis of skin disease using CAD systems. In this study, we propose a Fast Image Segmentation (FIS) method based on Contrast-Limited Adaptive Histogram Equalization (CLAHE) [10] and Thresholding capable of segmenting a lesion area from surrounding healthy skin. The rest of the paper is tersely structured as follows: Sect. 2 discusses relevant works and the importance of our study. Section 3 highlights the techniques used in this study and we concluded the chapter in Sect. 4.

2 Relevant Works

The recognition of lesion areas within a dermoscopic image data has been characterized with several difficulties, mainly because of the lopsided structure that are exhibited by many of the available medical images [11, 12]. In addition, the smooth transition between the lesion areas and surrounding healthy skin makes it hard to effectively segment lesion area from surrounding skin [11, 13]. Thresholding is one of the major techniques that are widely used for segmenting a lesion image based on color difference between the lesion area and surrounding skin. A particular threshold value is selected based on histogram data of a medical image. This threshold value can then be used to express the gray scale lesion image in a two homogenous mode (black and white). The white pixels of the binary image provide the necessary information for performing the segmentation task. One great merit of thresholding is that the method is computationally inexpensive and thus requires less computation time compared to other segmentation techniques such as region growing.

A number of thresholding approaches have been proposed in the literature for localizing lesion areas. A three-stage segmentation technique including segmentation initialization, processing and refinement was proposed by Cavalcanti and Scharcanski [14]. In the same paper, Otsu [15] intensity thresholding methods was used for the initialization task for obtaining a preliminary segmentation before later applying Tsumura et al. [16] color space projection technique to maximize the separability of the lesion pixels from the non-lesion counterpart. The skin tumor segmentation using dynamic programming was introduced by Abbas et al. [11] by combining thresholding method with edge-based dynamic programming in Commission Internationale de l'Eclairage (CIE) L*a*b color space.

Considering the difficulty in the detection of accurate threshold, Rahman et al. [17] used an iterative threshold clustering to segment lesion image data. Similar iterative technique using Iterative Self-Organizing Data Analysis Technique (ISODATA) algorithm was exemplified by Schaefer et al. [18] for determining optimal threshold value that can be used for effective segmentation. Supot [19] used fuzzy c-means clustering to determine acceptable threshold for lesion segmentation. The type-2 fuzzy based thresholding was used by Yuksel and Borlu [20] to ensure lesion localization is not affected by the fuzziness that is usually encountered at the border between lesion area and surrounding skin. In a bid to achieve an optimal segmentation result, Ganster et al. [13] employed the usage of both thresholding and 3D color clustering. The fusion of thresholding methods using Sarsa reinforcement algorithm assisted the study conducted by Ebrahimi and Pourghassem [21] to detect precise threshold value for localizing the lesion part of dermoscopic images. Similar fusion technique was used by Celebi et al. [22] where a set of different thresholding algorithms were ensembled. A Tetra level segmentation of different lesion was performed by Humayun et al. [23] to test effectiveness of multi-level thresholding.

While considerable good results has been achieved in the literature using various thresholding based segmentation techniques, the variation in image intensity at hand in lesion images sometimes makes current approaches to produce non-optimal segmentation results. In addition, the outcome of thresholding could be unfavorably affected by the shadow areas within the image data. These shadow areas due to their dark properties might be confused as being part of the lesion data. In this chapter, we propose a Fast Image Segmentation (FIS) procedure using mixture of methods involving the application of CLAHE, thresholding technique and GrabCut [24].

3 Materials and Methods

3.1 Data Set

A selected subset of dermoscopic images from the database provided by Dermatology Society of South Africa (DSSA) is used. In total, 294 images comprising of 105 Melanoma (Superficial Spreading Melanoma and Acral Lentiginous Melanoma) and 189 benign lesion images were used in our study. Each image is of dimension 640 × 480 pixels and 24-bit depth.

3.2 System Design

3.2.1 Initialization Phase

In the initialization phase, the RGB color space input is subjected to some pre-processing (see Sect. 3.3) to remove unwanted artefacts (see Table 1). The system then generates the Gray level of the pre-processed image. CLAHE [10] algorithm is then used to normalize the intensity of the gray image. It is important to state here that preprocessing is particularly important before normalizing the intensity of the image histogram to avoid noise being part of the equalized pixels.

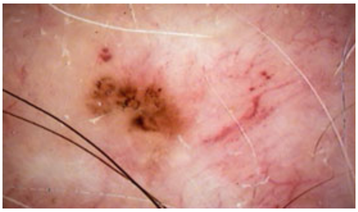
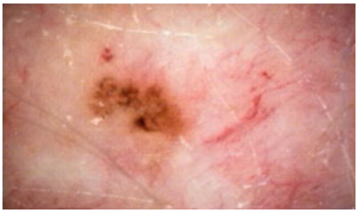
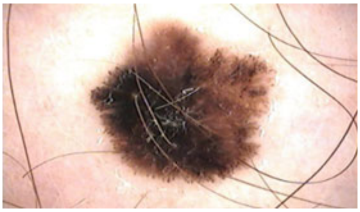
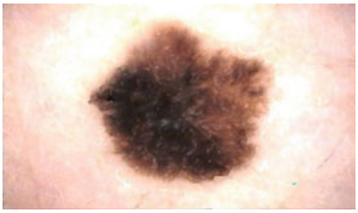
3.2.2 First Level Segmentation

To avoid data loss, the normalized image is first converted from grayscale to RGB before converting to CIE Lab*a*b color space. The image is quickly filtered on a disk size of 3 to smooth the image. To avoid data loss, the system convert the filtered image from CIE Lab*a*b color space to gray. In addition, good morphological operation result was achieved by operating on the gray level of an image. Thresholding was performed on the resultant image to produce a binary mask. The mask was then eroded with a 3 × 3 matrix kernel. The system then opened and closed the eroded image with a 9 × 9 matrix kernel. Median filtering was then applied to produce a segmentation mask which was eventually used to automatically crop out the image lesion area from the original source image.

3.2.3 Second Level Segmentation

GrabCut Algorithm (see Sect. 3.6) was used as a form of second segmentation process. The effect of the First Level is very crucial for the success of the second level segmentation. GrabCut can sometimes perform poorly in instances where smooth transition exists between foreground and background pixels. The first level

Table 1 Image cleansing

Source image	Image after pre-processing
	
	

segmentation approach using mixture of threshold and morphological operation assisted in estimated probable foreground. As shown in Table 3, there exist white pixels around the segmented image after undergoing first level segmentation. These white pixels are very helpful in ensuring GrabCut perform accurate second level segmentation to localize the lesion area from surrounding healthy skin.

3.3 Preprocessing

The artefacts such as hair shafts and ruler marking make segmentation of a lesion very difficult. These artefacts are generally regarded as noise. In this study, artefacts were removed by first detecting line segments representing hair pixels and ruler markings. We then replace the recognized noise with neighborhood pixels with the closest match to the lesion pixel under consideration (see Table 1).

3.4 Intensity Normalization

Many lesion images tend to have some pixels that are confined to some specific range of intensity values rather than having pixels from all regions of the image. The Histogram Equalization (HE) algorithm could easily be used to improve the contrast of an image by stretching the intensity of the image histogram.

The HE algorithm computes histogram of a given image and uses the information to transform all pixels of the image. This approach usually performs well if the distribution of the pixel values is identical across the image. However, unfavorable effect of this approach is that regions within the image that has significant darker or lighter pixels with respect to other pixels within the image would not be satisfactorily enhanced. The Adaptive Histogram Equalization (AHE) algorithm proposed by Ketcham et al. [25] addresses this challenge by using a transformation function derived from the neighborhood pixels to transform each pixel of the image.

A major challenge with the use of traditional AHE is that the method sometimes over amplifies noise in the region having relatively small intensity range. This study uses Contrast-Limited Adaptive Histogram Equalization (CLAHE) [10] in order to limit the noise amplification. The procedure used by CLAHE algorithm involves partitioning of a given grayscale image into contextual regions and then equalizes the histogram of each region. The CLAHE applies contrast limiting technique to each neighborhood grid point within a particular region from which a transformation is derived. Noise amplification is reduced by clipping image histogram at a predefined value just before computing the Cumulative Distribution Function (CDF) for each grid point.

Let

δ_i represents image to be histogram equalized using CLAHE
$\mathcal{X}$ represents total number of image pixels
$\mathcal{X}_o$ represents the number of occurrences of grayscale level o within the image
Γ represents total number of grayscale levels within the image (typically 256)
P_{δ_i} represents the probability of gray level o appearing in the image δ_i

$$\mathrm{P}_{\delta_i}(o) = \frac{x_o}{x} \tag{1}$$

Let CDF_{δ_i} represents the Cumulative Distribution Function of δ_i

$$CDF_{\delta_i}(o) = \int_0^o \mathrm{P}_{\delta_i}(o) \tag{2}$$

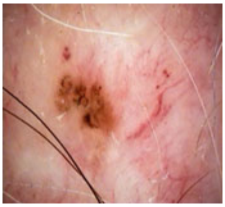
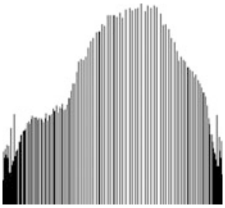
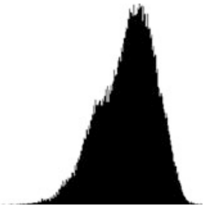
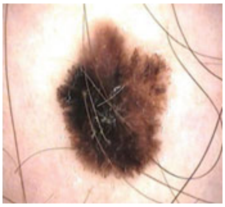
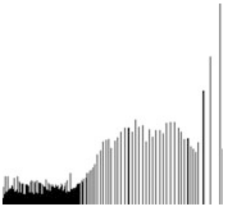
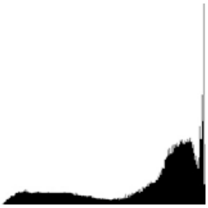
The Histogram Equalization (HE) of above CDF can be computed as:

$$HE_{CDF} = \sim \left[\left(\frac{CDF_{\delta_i} - CDF_{\min}}{CDF_{\max} - CDF_{\min}} \right) \times (\Gamma - 1) \right] \tag{3}$$

3.4.1 CLAHE Algorithm

1. Use a window size of 8 with a clip-limit of 1
2. Define grid points (separated by window size) on the image, beginning from Point (0,0)

Table 2 Intensity normalization

Source image	Equalization using HE	Equalization using CLAHE
		
		

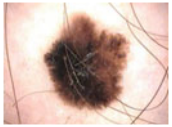
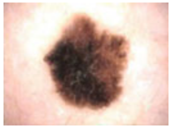
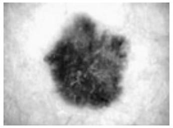
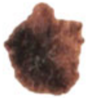
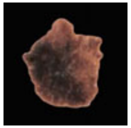
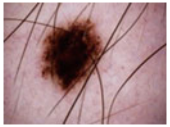
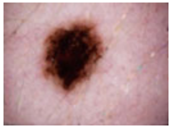
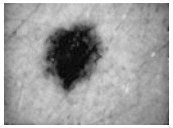
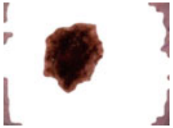
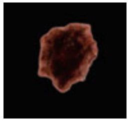
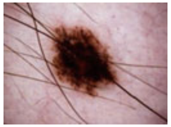
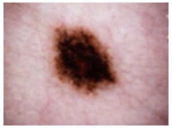
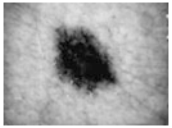
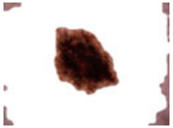
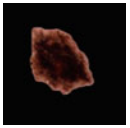
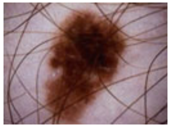
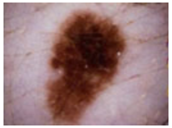
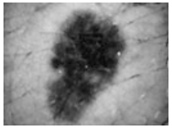
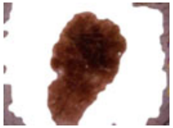
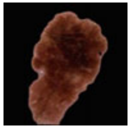
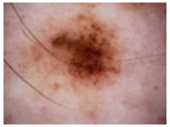
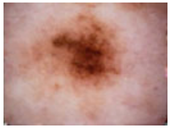
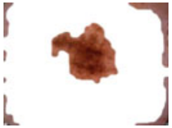
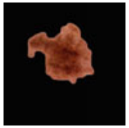
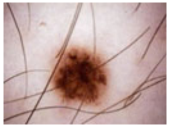
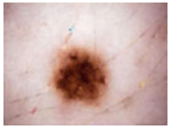
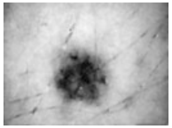
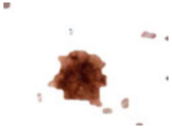
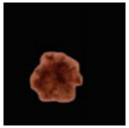
3. For each grid point in a given region
   - (a) define area = window size
   - (b) center area at grid point
   - (c) compute histogram for region around grid point
   - (d) clip the histogram above the clip-limit to define a new histogram
   - (e) define *CDF* using new histogram
4. For each pixel
   - (a) find the four closest neighboring grid points surrounding the pixel
   - (b) use the pixel intensity value to find the mapping of the pixel at the four grid points based on their *CDF*
   - (c) interpolate among the *CDF* values to get the intensity mapping at the current pixel location
   - (d) map the intensity to the intensity range values
   - (e) insert the intensity in the output image

Table 2 compares effect of equalization between CLAHE and HE technique to demonstrate the effectiveness of contrast limiting technique.

3.5 Thresholding and Morphological Operation

The estimation of the foreground pixel colors could sometimes be difficult for cases where smooth transition exist between lesion areas and the surrounding skin. As a first level segmentation, we applied mixture of morphological operation with thresholding. This step is particularly important before application of separation algorithm such as GrabCut as a second level segmentation for accurate localization

Table 3 Lesion segmentation

Source image	Preprocessed image	Image after using CLAHE	Thresholding and morphology	Image after using GrabCut
				
				
				
				
				
				

of lesion areas. A Median filtering with kernel size of 3 was used to smooth the image in order to ensure we get a better segmentation output. A Threshold method was then applied on the filtered image to generate a binary image. We eroded the resulting silhouette image with kernel size of 3, later opened the image with kernel size of 9 and closed the image with kernel size of 9. An iterative median blurring was again carried out on the image using disk size of 3 to produce a mask. This mask was then used to segment the lesion area from non-lesion ones.

3.6 GrabCut Segmentation

This study employed the usage of GrabCut [24] algorithm to perform a second level segmentation. GrabCut is a segmentation method that is based on GraphCut [26] technique. The GrabCut algorithm combines graph cut implementation with statistical modeling to achieve a favorable 2D segmentation. It uses Orhard-Bouman [27] color clustering algorithm to model foreground and background pixels as a Gaussian Mixture Model (GMM). The Detailed information regarding the implementation of GrabCut was carried out by Rother et al. [24].

3.6.1 GrabCut Algorithm Overview

1. (a) select a rectangle to create initial tri-map of the lesion area, (b) pixels outside the rectangle would be judged as known background pixels whereas pixels inside the rectangle are classified as unknown
2. classify all unknown pixels as probable foreground
3. using the probable foreground pixel data, apply Orchard-Bouman color clustering algorithm to model initial foreground and initial background as a Gaussian Mixture Models (GMMs)
4. (a) assign each pixel in the initial foreground to the most likely Gaussian component in the foreground GMM, (b) assign each pixel in the initial background to the most likely Gaussian component in the background GMM
5. discard GMMs in step 3 and learn new GMMs from pixels in step 4
6. (a) compute a graph and use graph cut to find new initial foreground and new initial background class, (b) model the new initial foreground and background as GMMs
7. repeat steps 4 to step 6 until the classification converges

4 Results and Conclusion

We have used ensemble of segmentation techniques on a total of 294 dermoscopic images. The preprocessing of the image is particularly important before either of the segmentation levels previously discussed is performed. The preprocessing used in our study ensured hair shafts and ruler markings are removed to avoid the segmentation process picking up noise pixels as valid lesion pixels. While there are many variation of AHE, the application of CLAHE was particularly effective for the objective of this study for reducing the contrast amplification of noise during histogram equalization.

It might appear sole usage of GrabCut algorithm would have been effective in the segmentation of lesion areas from surrounding skin. In our study, we discovered it performs poorly in instances where smooth transition exists between foreground and background pixels. Our first level segmentation approach using a combination of threshold and morphological operation assisted in estimated probable foreground. The actual lesion areas were then easily localized from resultant image using GrabCut algorithm. We have highlighted in Table 3, the resultant effect of major techniques applied in our study to a given dermoscopic image. We were able to achieve an average speed of 5188 ms on a heavily loaded virtual machine running Ubuntu 12.0.4.3 LTS, with a base memory of 3096 MB and 2 processors.

In this study, we have applied a mixture of methods including thresholding and GrabCut color clustering to accurately localize lesion areas from surrounding healthy skin. This is essential for an effective selection of discriminating features, which in turn could be used for classifying melanocytic lesions as either malignant or benign.

References

1. M.C. Gereli, Melanoma Dermoskopik Tanisinda (2006), http://www.istanbulsaglik.gov.tr/w/tez/pdf/deri_zuhrvei/dr_muge_celebi_gereli.pdf
2. H. Pehamberger, A. Steiner, K. Wolff, In vivo Epiluminiscence microscopy of pigmented skin lesion: pattern analysis of pigmented skin lesions. J. Am. Acad Dermatol. **17**, 571–583 (1987)
3. W. Stolz, A. Riemann, A. Cognetta et. al., ABCD rule of dermoscopy: a new practical method for early recognition of malignant melanoma. Eur. J. Dermatol. **4**, 521–527 (1994)
4. H. Kittler, Dermoscopy of pigmented skin lesions. G.Ital. Dermatol. Venereol. **139**(6), 541–546 (2004)
5. D.A. Okuboyejo, O.O. Olugbara, S.A. Odunaike, Automating skin disease diagnosis using image classification, Lecture Notes in Engineering and Computer Science. in *Proceedings of the World Congress on Engineering and Computer Science 2013, WCECS, 2013*, 23–25 Oct 2013, (USA, San Franscisco, 2013), pp. 850–854
6. P. Rubegni, G. Cevenini, M. Burroni, P. Roberto, G.D. Eva, P. Sbano, C. Moracco, P. Luzi, P. Tosi, P. Barbini, L. Andreassi, Automated diagnosis of pigmented skin lesion. Int. J. Cancer **101**, 576–580 (2002)
7. C. Rosendahl, P. Tschandl, A. Cameron, H. Kittler, Diagnostic accuracy of dermatoscopy for melanocytic and nonmelanocytic pigmented lesions. J. Am. Acad Dermatol. **64**, 1068–1073 (2001). doi: 10.1016/j.jaad.2010.03.039
8. I. Stanganelli, A. Brucale, L. Calori, R. Gori, A. Lovato, S. Magi, B. Kopf, R. Bacchilega, V. Rapisarda, A. Testori, A. Ascierto, E. Simeone, M. Ferri, Computer-aided diagnosis of melanocytic lesions, in *Proceedings of Anticancer Research*, 2005, pp. 4577–4582

9. A.K. Mittra, R. Parekh, Automated detection of skin disease using texture features. Int. J. Eng. Sci. Tech. **3**(6) (2011)
10. K. Zuiderveld, *Contrast Limited Adaptive Histogram Equalization*, (Academic Press, New York, 1994), pp. 474–485
11. Q. Abbas, M.E. Celebi, I.F. García, Skin tumor area extraction using an improved dynamic programming approach. Skin Res. Tech. **18**, 133–142 (2012)
12. R. Dobrescu, M. Dobrescu, S. Mocanu, D. Popescu, Medical images classification for skin cancer diagnosis based on combined texture and fractal analysis, in *Proceedings of WSEAS Transactions on Biology and Biomedicine, 2010* (2010)
13. H. Ganster, A. Pinz, R. Roher, E. Wilding, M. Binder, H. Kittler, Automated melanoma recognition. IEEE Trans. Med. Imaging **20**(3), 3 (2001)
14. P.G. Cavalcanti, J. Scharcanski, A coarse-to-fine approach for segmenting melanocytic skin lesions in standard camera images. Comput. Methods Programs Biomed. **112**, 684–693 (2013). doi: 10.1016/j.cmpb.2013.08.010
15. N. Otsu, A threshold selection method from gray-level histograms. IEEE Trans. Sys., Man., Cyber **9**(1), 62–66 (1979). doi: 10.1109/TMSC.1979.4310076
16. N. Tsumura, N. Ojima, K. Sato, M. Shiraishi, H. Shimizu, H. Nabeshima, S. Akazaki, K. Hori, Y. Miyake, Image-based skin color and texture analysis/synthesis by extracting hemoglobin and melanin information in the skin. in *ACMSIGGRAPH, SIGGRAPH'03,; 2003*; (ACM, New York, NY, USA, 2003), pp. 770–779
17. M.M. Rahman, P. Bhattacharya, B.C. Desai, A multiple expert-based melanoma recognition system for dermoscopic images of pigmented skin lesions, in *Proceedings of 8th IEEE International Conference on BioInformatics and BioEngineering (BIBE), 2008*, 8–10 Oct 2008, pp. 1–6
18. G. Schaefer, M.I. Rajab, M.E. Celebi, H. Iyatomi, Skin lesion extraction in dermoscopic images based on colour enhancement and iterative segmentation, in *Proceedings of ICIP* (2009), pp. 3361–3364
19. S. Supot, Border detection of skin lesion images based on fuzzy C-Means thresholding, in *Proceedings of 3rd International Conference on Genetic and Evolutionary Computing* (2009), pp. 777–780
20. M.E. Yuksel, M. Borlu, Accurate Segmentation of dermoscopic images by image thresholding based on type-2 fuzzy logic. IEEE Trans. Fuzzy Syst. **17**(4), 976–982 (2009)
21. S.M.S. Ebrahimi, H. Pourghassem, Lesion Detection in dermoscopy images using Sarsa reinforcement algorithm, in *Proceedings of 17th Iranian Conference of Biomedical Engineering* (*ICBME2010*), *2010*, 3–4 Nov 2010
22. M.E. Celebi, S. Hwang, H. Iyatomi, G. Schaefer, Robust Border detection in dermoscopy images using threshold fusion, in *Proceedings of IEEE 17th International Conference on Image Processing, 2010*, Hong Kong, 26–29 Sept 2010
23. J. Humayun, A.S. Malik, N. Kamel, Multilevel thresholding for segmentation of pigmented skin lesions, in *Proceedings of IEEE International Conference on Imaging Systems and Techniques* (*IST*) (2011)
24. C. Rother, V. Kolmogorov, A. Blake, GrabCut: interactive foreground extraction using iterated graph cuts. in *ACM SIGGRAPH; 2004*, (ACM, Los Angeles, California, 2004), Aug 2004, pp. 309–314
25. D.J. Ketcham, R.W. Lowe, J.W. Weber, Image enhancement techniques for cockpit displays, Technical report (1974)
26. Y. Boykov, M.P. Jolly, Interactive graph cuts for optimal boundary and region segmentation of objects in N-D images, in *Proceedings of IEEE International Conference on Computer Vision* (2001)
27. M. Orchard, C. Bouman, Color quantization of images. IEEE Trans. Sig. Process. **39**, 2677–2690 (1991)

Chapter 28
ISAR Imaging for Moving Targets with Large Rotational Angle

Jinjin Zhang, Xinggan Zhang and Yechao Bai

Abstract Wideband radar can get high range resolution. In order to obtain high azimuth resolution, a large rotation angle of moving targets relative to the radar line of sight is required. As the migration of scatterers in different range resolution cells is different, range alignment cannot compensate it completely. When the rotation angle of the moving target is large, the migration through resolution cell (MTRC) will occur obviously, which results in great image resolution degradation. A new ISAR imaging algorithm is proposed in this paper to solve the problem. In the proposed algorithm, the process of large-angle high-resolution imaging is divided into several small-angle low-resolution imaging processes. Then after image scaling, image rotation, image translation, inverse process of ISAR imaging, data splicing and re-ISAR imaging, high resolution image will be produced. Theoretical analysis and simulations show that the proposed algorithm can compensate the MTRC of scatterers and obtain high-resolution images.

Keywords High resolution imaging · Image registration · Image scaling · Inverse synthetic aperture · Migration through resolution cell · Wideband radar

J. Zhang · X. Zhang · Y. Bai (✉)
School of Electronic Science and Engineering, Nanjing University, Nanjing 210023,
People's Republic of China
e-mail: ychbai@nju.edu.cn

J. Zhang
e-mail: njuskyee@gmail.com

X. Zhang
e-mail: zhxg@nju.edu.cn

H. K. Kim et al. (eds.), *Transactions on Engineering Technologies*,
DOI: 10.1007/978-94-017-9115-1_28,

1 Introduction

Inverse Synthetic Aperture Radar (ISAR) can obtain two-dimensional images (the images of range and azimuth directions) of moving targets. Usually the range-Doppler (RD) algorithm is used for ISAR imaging [1]. To get high azimuth resolution images, it needs a large angle of the rotating target relative to the radar line of sight during the coherent accumulation time. However, too large rotational motion will cause the migration through resolution cell (MTRC) of part of scatterers, and the imaging quality based on the traditional RD algorithm will decline.

Currently, a number of papers have proposed methods to compensate the MTRC of scatterers. In [2], the traditional RD algorithm, the polar formatting algorithm (PFA) and the keystone algorithm are analyzed. PFA can compensate the MTRC of scatterers, while the range and cross range interpolation operation is required, which will affect the imaging precision and the computational efficiency. The Keystone algorithm [3] and the MTRC compensation algorithm (MTRCCA) [4] can both compensate the MTRC. Unfortunately, these methods are not applicable to the large rotational motion. In [5] the sub-image fusion algorithm (SIF) is proposed. The algorithm can compensate the large rotational motion in high-resolution ISAR imaging and reduce the image blurring. However, it is limited to the situation with equal range resolution and azimuth resolution.

Considering the above problems, we propose a new high resolution ISAR imaging algorithm [6]. Echo segmenting is performed first, and low-resolution images (sub-image) can be obtained from the segments. Then after image scaling, image registration including image rotation and translation, inverse process of ISAR imaging, data splicing and re-ISAR imaging, the synthetic high resolution image can be obtained finally. The proposed method just needs to estimate the angles between one of the sub-images and the others, which can greatly reduce the computation. Besides, it is suitable for any resolution imaging due to the new process of scaling and registration. The simulation results show that the proposed algorithm avoids the MTRC in the situation of large rotational motion, and the azimuth resolution is significantly improved.

2 Algorithm Description

The proposed algorithm has two main steps: echo segment imaging and high-resolution imaging. The segment imaging will be performed first to obtain sub-images. And then the high-resolution imaging will be performed which includes scaling, image rotation, image translation, inverse process of ISAR imaging, data splicing and ISAR imaging. The high resolution image can be obtained at last. The equivalent rotation angle is large in the proposed algorithm, and the MTRC does not exist.

2.1 Echo Segment Imaging

The relative motion between the radar and the target can be decomposed into translational motion and rotational motion. The translational motion should be accurately compensated in ISAR imaging, which is called motion compensation. The compensated signals can be treated as the echoes of a rotating scatterer with respect to the reference center as shown in Fig. 1.

Firstly, the echoes should be segmented to several equal parts. In order to obtain clear sub-images, the migration of the scatterers in each echo segment should not be through resolution cell. After segmenting, the rotation angle and the walking distance of the scatterers in each segment are small. In that condition, the traditional imaging methods can be used. The total number of the echoes is assumed to be N. The echoes are segmented to k parts. Then the number of echoes in each part is n, which can be written as

$$\mathrm{n} = \frac{N}{k} \tag{1}$$

In Fig. 1, point O is the reference point, and the scatterer $Q(x, y)$ rotates about O. During the coherent processing, the rotation angle of $Q(x, y)$ is defined as θ, and the angular velocity is represented by ω. T_a is the observing duration. R_0 is the range from the radar to the reference center. Suppose that the radar transmits a linear frequency modulated (LFM) signal, which can be written as

$$s(t) = rect(\frac{t}{T_p}) \exp\left[j2\pi(f_c t + \frac{\mathrm{u}}{2} t^2)\right] \tag{2}$$

f_c is the carrier frequency, u is the chirp rate, T_p denotes time width of the chirp pulse, the signal bandwidth is B, and $rect(\bullet)$ is the unit rectangular function. B can be written as

$$B = T_p \cdot u \tag{3}$$

After range compression, phase compensation and Keystone transformation, the echo segment data is transformed from time domain to range domain by applying the Fourier Transformation in the azimuth direction, which can be written as

$$s(t,f_d) = A \cdot \sin c[B(t-y)] \cdot \sin c\left[T_a\left(f_d - \frac{2\omega x}{\lambda}\right)\right] \exp\left(-j4\pi\frac{y}{\lambda}\right) \tag{4}$$

where A is the backward scattering amplitude (A can be viewed stationary during the coherent processing), and λ is the wavelength. From the above, the sub-images are produced.

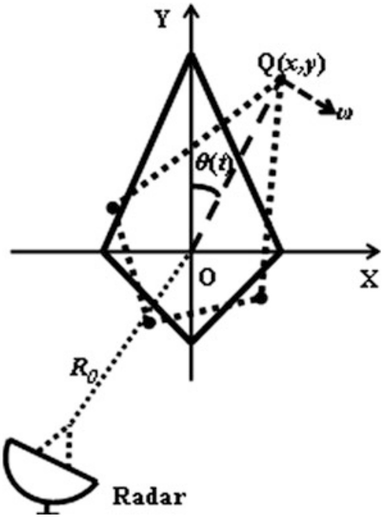


Fig. 1 ISAR imaging geometry

2.2 Sub-image Processing

We have got sub-images after echo segment imaging. Because of the different original time of the sub-images, the coordinates of the scatterers are different. The sub-images should be rotated and aligned to make the coordinates of the scatterers same which is called image registration. However, rotating directly with different range resolution and azimuth resolution will cause the image distortion. So scaling should be done before image rotating.

Assume that the actual distance denoted by the sampling unit in the range direction is δ_y, and f_s is the sampling frequency. We know that

$$\frac{1}{f_s} = \frac{2\delta_y}{c} \tag{5}$$

$$\delta_y = \frac{c}{2f_s} \tag{6}$$

Assume that the size of azimuth resolution cell is δ_x which can be written as

$$\delta_x = \frac{PRF}{n} \cdot \frac{\lambda}{2\omega} \tag{7}$$

where n is the number of echoes in each segment, and PRF is the pulse repetition frequency.

When δ_x doesn't equal δ_y, image scaling must be performed before image rotating which means adjusting the range resolution and the azimuth resolution to be the same by stretching the range image. All the sub-images can be rotated to the same coordinate after scaling. The angles between the first sub-image and the others should be estimated before rotating. Currently, some approaches [7, 8] have been introduced to estimate the rotational angular velocity. In this paper, we will not describe it in detail.

Suppose that the coordinate of a scatterer in the first sub-image is (x_0, y_0) and the coordinate in another sub-image is (u_0, v_0). When the rotation angle between them is β, the coordinate transformation relationship is

$$\begin{cases} x = u\cos(\beta) - v\sin(\beta) \\ y = v\cos(\beta) + u\sin(\beta) \end{cases} \tag{8}$$

The matrix size will change while rotating, so it is necessary to do the image translation and then cut the sub-images to the same size. The image translation is composited by rough translation and accuracy translation. The image moves integral sampling units in rough translation. Then fractional delay algorithm [9] is used for accuracy translation. The sub-images are cut to the same size after translation.

2.3 High Resolution Imaging

After image registration, the inverse process of ISAR imaging should be performed. Then by splicing the entire segment data and re-ISAR, the high resolution image is obtained. As all the echo segments contain the same motion information after processing, the data can be spliced to produce a high resolution image. The segment data in the azimuth direction should be transformed from range domain to time domain by IFFT (Inverse Fast Fourier Transform). After splicing all the segments data and transforming the spliced data by FFT (Fast Fourier Transform) in the azimuth direction, we can get high resolution ISAR image.

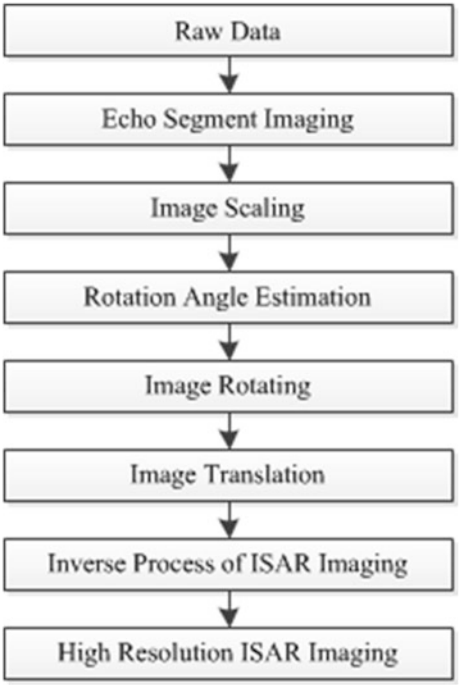
The algorithm flow chart is shown as Fig. 2.

3 Simulation and Analysis

The proposed algorithm divides the large rotation angle into several small ones. The length of the data after splicing increases to N ($N = n \times k$). The actual distance represented by the azimuth resolution cell of the proposed algorithm is $\delta_x^{'}$, which can be written as

$$\delta_x' = PRF \cdot \frac{\lambda}{2N\omega} \tag{9}$$

Fig. 2 The flow chart of the proposed algorithm


Raw Data
Echo Segment Imaging
Image Scaling
Rotation Angle Estimation
Image Rotating
Image Translation
Inverse Process of ISAR Imaging
High Resolution ISAR Imaging

According to (9), the actual distance represented by the azimuth resolution cell can be reduced to 1/k. The azimuth resolution is improved significantly. As the process involved in the interpolation, the resolution will be reduced to a certain extent, and actually it is difficult to achieve the theoretical resolution.

Assume that a moving target is composed with four points (A, B, C, D, E and F), with coordinates being A (−2 m, 0), B (−1.84 m, 0 m), C (1.84 m, 0), D (2 m, 0), E (−2 m, 0), F (0, 4 m) respectively, shown as Fig. 3.

The angular velocity of the target is ω, and $\omega = 0.15$ rad/s. The distance between the radar and the target is $R_0 = 4{,}000$ m. Radar parameters are set as follows: the pulse repetition frequency is 400 Hz, the bandwidth is 2 GHz, the center frequency is 10 GHz, the sampling frequency is 2 GHz. and the range resolution is 3/40 m. To avoid sub-image degradation caused by the rotational motion, the number of the echoes should be limited. Suppose that D_r and D_a are the maximum in range and azimuth directions of the target respectively. The following equations should be satisfied in RD algorithm.

$$\begin{cases} D_r < \frac{4\rho_x^2}{\lambda} \\ D_a < \frac{4\rho_x\rho_y}{\lambda} \end{cases} \tag{10}$$

As D_a is supposed to be 4, the number of echoes is limited to 100 according to Eqs. (7) and (10). We take 500 echoes for ISAR imaging and divide the echoes to

Fig. 3 Coordinates of the target

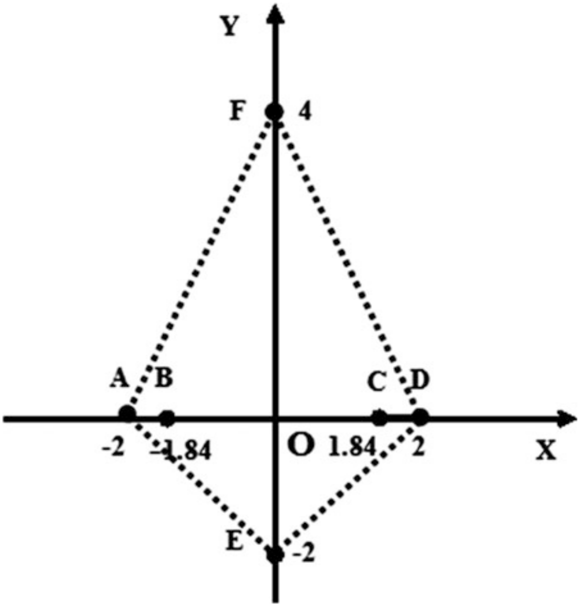
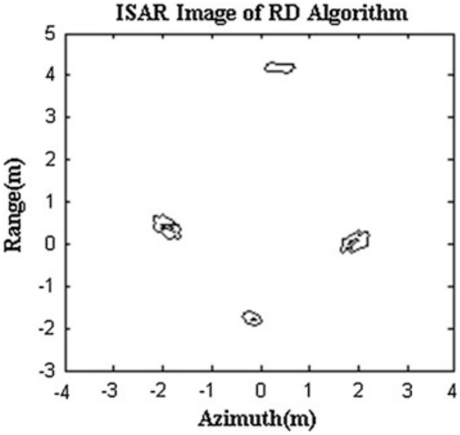
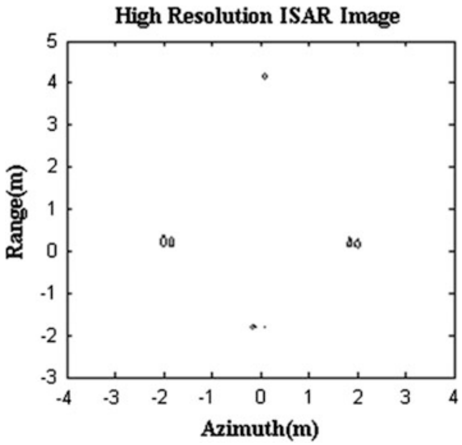


Fig. 4 ISAR image of RD algorithm



five equal parts. 100 echoes are used for producing one sub-image and the migration of the scatterers in each sub-image is not through resolution cell. According to (8), the azimuth resolution is 0.08 m. The distance between C and D is 0.16 m. So C and D can be separated theoretically. The imaging results of the traditional RD algorithm and the proposed algorithm are shown in Figs. 4 and 5.

Fig. 5 ISAR image of the proposed algorithm



4 Conclusion

To compensate the MTRC in the situation of large rotation angle, a new algorithm is proposed in this paper. The algorithm is utilized by dividing the large-angle high-resolution imaging process into multiple small-angle low-resolution images. Firstly, we should divided the echoes into several parts and do echo segment imaging. Then after image scaling, image rotation, image translation, inverse process of ISAR imaging, data splicing and re-ISAR imaging, high resolution image will be produced. It is verified by simulation results.

Acknowledgment This work was supported by Industry-academic Joint Technological Innovations Fund Project of Jiangsu Province (BY2012187).

References

1. C.C. Chen, H.C. Andrews, Target-motion-induced radar imaging. IEEE Trans. Aerosp. Electron. Syst. **1**, 2–14 (1980)
2. F. Zhou, X. Bai, M. Xing et al., Analysis of wide-angle radar imaging. IET Radar Sonar Navig. **5**(4), 449–457 (2011)
3. M. Xing, R. Wu, J. Lan et al., Migration through resolution cell compensation in ISAR imaging. IEEE Geosci. Remote Sens. Lett. **1**(2), 141–144 (2004)
4. G.Y. Lu, Z. Bao, Compensation of scatterer migration through resolution cell in inverse synthetic aperture radar imaging. IEE Proc. Radar Sonar Navig. **147**(2), 80–85 (2000)
5. H. Wang, Y. Liang, M. Xing et al., Subimage fusion for high-resolution ISAR imaging. IEEE 2010 3rd Int. Congr. Image Signal Process. (CISP) **5**, 2260–2264 (2010)
6. J. Zhang, X. Zhang, Y. Bai, A high resolution ISAR imaging method for wideband radar, in *Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering and Computer Science 2013, WCECS 2013*, pp. 550–553, San Francisco, USA, 23–25 Oct 2013

7. S.B. Peng, J. Xu, Y.N. Peng et al., Inverse synthetic aperture radar rotation velocity estimation based on phase slope difference of two prominent scatterers. IET Radar Sonar Navig. **5**(9), 1002–1009 (2011)
8. Y. Wang, Y. Jiang, A novel algorithm for estimating the rotation angle in ISAR imaging. IEEE Geosci. Remote Sens. Lett. **5**(4), 608–609 (2008)
9. Y. Wei, X. Zhang, Y. Bai et al., LFM pulse compression of wideband passive phased array based on true time delay, in *Proceedings of the 2012 International Conference on Control Engineering and Communication Technology*. IEEE Computer Society, pp. 436–439 (2012)

Chapter 29
Analysis of Acoustic Signal Based on Modified Ensemble Empirical Mode Decomposition

Sundeok Kwon and Sangjin Cho

Abstract Empirical mode decomposition (EMD) has been proposed recently as an adaptive time-frequency data analysis tool in the nonlinear and nonstationary signal processing research field. However, it has some drawbacks such as mode mixing, which is defined as a single intrinsic mode function (IMF) consisting of signals of widely disparate scales, and the prediction problem, which is intimately related to the end effects in the EMD. Although the ensemble EMD (EEMD), which overcomes these problems, can represent a major improvement of the EMD, this paper proposes a modified EEMD which includes the best IMF selection algorithm to analyze the acoustic signal. To evaluate the proposed method, it is applied to the vibration signal of an induction motor which has a bearing fault and musical percussion sound.

Keywords Acoustic signal processing · Ensemble empirical mode decomposition · Fault detection · Intrinsic mode function · IMF selection algorithm · Musical instrument

1 Introduction

Most popular signal processing techniques include time domain analysis, frequency domain techniques like spectral analysis and time-frequency domain methods such as short-time Fourier transform (STFT) or wavelet analysis. The main purpose of

S. Kwon
Industry-Academy Cooperation Foundation, Youngsan University, 99 Pilbong-gil,
Haeundae-gu, Busan 612-713, South Korea
e-mail: winder2000@naver.com

S. Cho (✉)
Automobile/Ship Electronics Convergence Center, University of Ulsan, 93 Daehak-ro,
Nam-Gu, Ulsan 680-749, South Korea
e-mail: sjcho75@ulsan.ac.kr

H. K. Kim et al. (eds.), *Transactions on Engineering Technologies*,
DOI: 10.1007/978-94-017-9115-1_29,

signal processing step in a fault diagnosis system, for example, is to reveal fault signatures from the measured quantities obtained from a motor in operation. For this purpose, time–frequency analysis tools are popular as they can provide both time and frequency resolution simultaneously. Most existing time–frequency analysis methods decompose the signal based on a priori bases with stationary assumption of the signal. But both of these techniques are not suitable enough for the analysis of fault signals as they can be non-linear and non-stationary at the same time. In contrast, empirical mode decomposition (EMD) is a signal decomposition method which decomposes the signal into some intrinsic mode functions (IMFs) based on the local characteristic time scale of the data [1]. These IMFs represent natural oscillatory modes embedded in the signal and works as the basis functions, which are derived from the signal itself, rather than any pre-determined kernel. Therefore, EMD is a data adaptive decomposition technique which overcomes limitations of other similar tool such as STFT or wavelet. The essence of the EMD method is that, it empirically determines the intrinsic oscillatory modes by the characteristic time scales within a signal and decomposes the signal accordingly. This excellent mode separation capability of EMD makes it an optimum choice for the analysis of natural phenomenon like vibration signal analysis. Every rotating part of a mechanical system contributes to the generation of vibration signal which is acquired through accelerometer. As a result any defect or abnormality in rotating behavior of a rotating part will cause to change contribution of that specific part, which ultimately modifies the normal vibration pattern and indicate the inception of fault of a rotating part. As EMD extract the intrinsic oscillatory modes from a signal, abnormal rotating behavior of a mechanical part can easily be identified by inspecting statistical property of these oscillatory modes. Therefore, EMD is a suitable signal analysis tool and it can be exploited for the development of a fault detection and diagnosis system. Yu et al. proposed the concept of EMD energy entropy and showed that its value for vibration signals differs in case for different bearing fault types [2]. In the proposed fault diagnosis method IMFs with dominant fault information were identified and their energy values constituted feature vector which was later utilized by back propagation artificial neural network to recognize fault pattern. Junsheng et al. proposed a fault diagnosis method for gear and bearing signals which utilized singular values of the matrices constituted by the IMFs as feature vectors for support vector machine classifier [3]. Peng et al. proposed an improved Hilbert–Huang transform combining wavelet packet transform, EMD and IMF selection technique to detect rubbing between stator and rotor of an induction motor [4]. In this case vibration signal is first decomposed into a set of narrowband signals which are further decomposed into IMFs by EMD and useful IMFs are selected by thresholding correlation coefficient between the IMFs and the original signal. Finally, rubbing symptoms are detected through the analysis of Hilbert spectrum of the selected IMFs.

The EMD can be also applied to the musical instrument research because it is important to analyze natural mode of vibration in the musical acoustics; it is utilized for the extraction of the resonance or vibrational mode of the percussion instruments. Cho applied EMD and ensemble EMD (EEMD) to the extraction of the vibrational mode of the percussion instrument called *Jing* [5, 6]. In [5], an algorithm was proposed to solve end effects problem and EEMD was used to extract features of non-harmonic characteristics of the Korean percussion instrument called *Kkwaenggwari* [7].

In this paper, the EEMD is used because the EEMD can represent a major improvement of the EMD. Therefore, this paper proposes a modified EEMD including best IMF selection algorithm and applies proposed method to analysis of the induction motor fault signal and musical signal.

2 Ensemble Empirical Mode Decomposition

The EEMD is one of noise-assisted data analysis (NADA) method and is proposed to overcome the scale separation problem, which the EMD has, without introducing a subjective intermittence test. This defines the true IMF components as the mean of an ensemble of trials, each consisting of the signal plus a white noise of finite amplitude [8]. In other words, EEMD performs iterative EMD process for the noise-added signal.

In the EMD approach, the data $x(t)$ is decomposed in terms of IMFs, c_j, i.e.,

$$x(t) = \sum_{j=1}^{n} c_j + r_n, \tag{1}$$

where r_n is the residue of data $x(t)$, after n number of IMFs are extracted. The EMD procedure is as follows:

Step 1. After identifying all local extrema, connect all these local maxima (minima) with a cubic spline as the upper (lower) envelope and calculate the local mean of the two envelopes.

Step 2. Obtain the first component h by taking the difference between the data and the local mean of the two envelopes.

Step 3. Treat h as the data and repeat steps 1 and 2 as many times as is required until certain stoppage criterion is satisfied.

Step 4. The final h is designated as c_j and a complete sifting process stops when the residue, r_n, becomes a monotonic function from which no more IMFs can be extracted.

To develop the EEMD, noise is introduced to the data as if separate observations were indeed being made as an analog to a physical experiment that could be repeated many times. Under this condition, the ith artificial observation will be

$$x_i(t) = x(t) + w_i(t) \tag{2}$$

where $w_i(t)$ is the added white noise treated as the possible random noise that would be encountered in the measurement process [8]. Therefore, EEMD procedure can be summarized as follows:

Step 1. After adding a white noise series to the data, perform the EMD as described above.
Step 2. Repeat step 1 again and again, but with different white noise series each time.
Step 3. Obtain the ensemble means of corresponding IMFs of the decompositions as the final result.

3 Proposed Method

The EEMD is the improved method that overcomes major drawbacks of the EMD such as mode mixing; the EEMD eliminated largely the mode mixing problem and preserve physical uniqueness of decomposition. These advantages notwithstanding, the EEMD still has some problems to be solved such as IMF selection. It is important to select IMFs representing a natural oscillation as well as a significant feature. Therefore, we propose a specific IMF selection considering characteristics of the acoustic signal.

The significant IMFs usually hold their unique characteristics. These characteristics worked as the basis of the proposed IMF selection process. The first characteristic is the energy variation of the specific oscillations. For instance, a faulty induction motor generates new oscillations or amplifies specific oscillations. These oscillations generally have higher energy in comparison to other oscillations which do not represent a fault situation. This phenomenon is observed in an excited body of the musical instrument. The body amplifies oscillations when the excitation oscillation matches the resonant mode of the body. As a consequence, these oscillatory modes are characterized by the energy variation. The second characteristic is the regularity of the spectral peak distribution of the IMFs. Usually, most of the fault signatures appear themselves as peak amplitude at several partials (or harmonics) of some fundamental frequency or show specific spectral components such as certain frequency and its subband. Due to the dyadic filter bank nature of EMD process, few of these partial peaks will be observed in the spectrum of lower index IMFs; whereas, fundamental fault frequency peak can be found in the higher index IMFs. Considering above two facts, an index named as Partial-Energy Constant (PEC) is calculated for each IMF. This higher value of PEC helps us to identify the IMFs with higher average power containing fault related frequency peaks. A low value of PEC indicates that the IMF may be of low power or contain many low amplitude partials. A summary of this IMF selection algorithm is as follows:

Step 1. Determine the threshold value defined as the mean of the energy values for each of the IMFs and select candidate IMFs having higher energy than threshold.

Step 2. For each of candidate IMFs, evaluate Fourier spectrum and find frequencies whose peaks show higher amplitude than the mean value for all detected peak amplitude in the spectrum.

Step 3. After calculating the total energy of frequencies found in step 2 and their partial (or harmonic) components, compute the PEC of each candidate IMF defined as the ratio between energy of the IMF and the total energy.

Step 4. Rearrange IMFs in descending order and first *m* IMFs are selected. *m* is the number of desired IMFs.

4 Simulation and Results

For the purpose of evaluating performance of the proposed method, two kinds of signal are used: one is the vibration signal acquired from a faulty induction motor and another is the musical sound of the *Jing*, the traditional Korean percussion instrument. To analyze these acoustic signals by utilizing EEMD, it is necessary to set parameters as follows: (1) Large ensemble number, which is the iteration number of the EMD for noise-added signal. As the number grows, the signal tends to emerge from the averaged IMF set and it can reduce the mode mixing problem. In this paper, the ensemble number is set for 1000. (2) Stoppage criterion, which determines the number of sifting process to produce an IMF. The *S*-number [9] is used and is set for 10. (3) Total number of IMFs, which is usually close to $\log_2 N$, where N is the number of data sample. In this paper, this is set to $\mathrm{fix}(\log_2 N)$ in which $\mathrm{fix}(X)$ rounds the elements of X to the nearest integers towards zero.

In the experiment, 0.5 kW, 60 Hz, 2-pole induction motor is used to produce the fault data under full load conditions and three accelerometers are used to measure vibrations in horizontal, vertical and axial directions. The sampling frequency of the data acquisition unit was 7.68 kHz. The maximum frequency of interest of the measured signals was 3 kHz [10]. In this paper, the fault signal, which is horizontal vibration signal for the bearing fault, is used and it contains 7680 number of samples. Therefore, number of IMF obtained after the EEMD is fixed to 12 ($\mathrm{fix}(\log_2 7680) = 12$). Among these 12 components, first 11 are IMFs and last one is residue. For the bearing fault at outer raceway, PEC values of different IMFs and associated parameters required for calculation of PEC according to the proposed IMF selection technique are shown in Table 1. Power spectra of the selected IMFs are shown in Fig. 1. According to Table 1, the first IMF has the highest PEC value, therefore, it is expected that the power spectrum of IMF 1 will show the signature of the bearing fault. The bearing fault was created on the outer raceway by a spalling, therefore, harmonics of ball pass frequency

Table 1 PECs for IMFs of bearing fault signal

IMF index	EI ($\times 10^{-2}$)	Threshold ($\times 10^{-2}$)	PEC ($\times 10^{-2}$)	PEC DO
1	9.7207	0.7355	26.1287	1
2	5.6341		14.6117	2
3	2.3830		5.9713	4
4	1.5158		4.1986	5
5	0.9979		2.0011	7
6	0.8928		0.6143	8
7	2.2983		6.7833	3
8	1.1458		2.6274	6
9	0.3489		–	–
10	0.1119		–	–
11	0.1354		–	–

EI energy of IMF, *DO* descending order

outer race (BPFO) is expected in the spectrum. The BPFO is observed in the spectrum of the IMF 7 and is 76 Hz. IMF 1 showing the highest PEC value contains 14th to 49th harmonics of BPFO and the 18th harmonic which is not observed in the IMF 1 is shown in the IMF 2. IMF 3 has 14th to 49th harmonics and other IMFs shown in Fig. 1 have one to four harmonics of BPFO. Thus, according to PEC value we can distinguish the significant IMFs with fault signature.

The musical sound used in this paper is performed by the professional player and is recorded by a digital recorder in the anechoic room. The recorder was set for 48 kHz sampling frequency and 16 bits quantization. From the *Jing* sound, 17 IMFs are totally extracted and four IMFs are selected by the proposed algorithm as shown in Table 2, whereas, in [7], six IMFs are selected empirically by the author. According to Fig. 2, IMF 1 and 2 contain high amplitude partials that they show harmonicity as described in Table 3 and the conspicuous component in IMF 3 and 4 is the fundamental oscillatory mode. This characteristic is similar to those of [11, 12]. In other words, the selected IMFs clearly represent the acoustic features of the *Jing*. In addition, the harmonicity of the *Jing* sound is attained through a unique tuning process which involves "hammering" the main plate of the *Jing*. This hammering process gives rise to thickness variation, surface inhomogeneity, and changes in the material property. These factors are responsible for shifting the inharmonic partials to harmonicity [11]. Therefore, the *Jing*s, manually manufactured by the master, have the similar but different their own eigen frequency and partials. In other words, selected IMFs are physically meaningful.

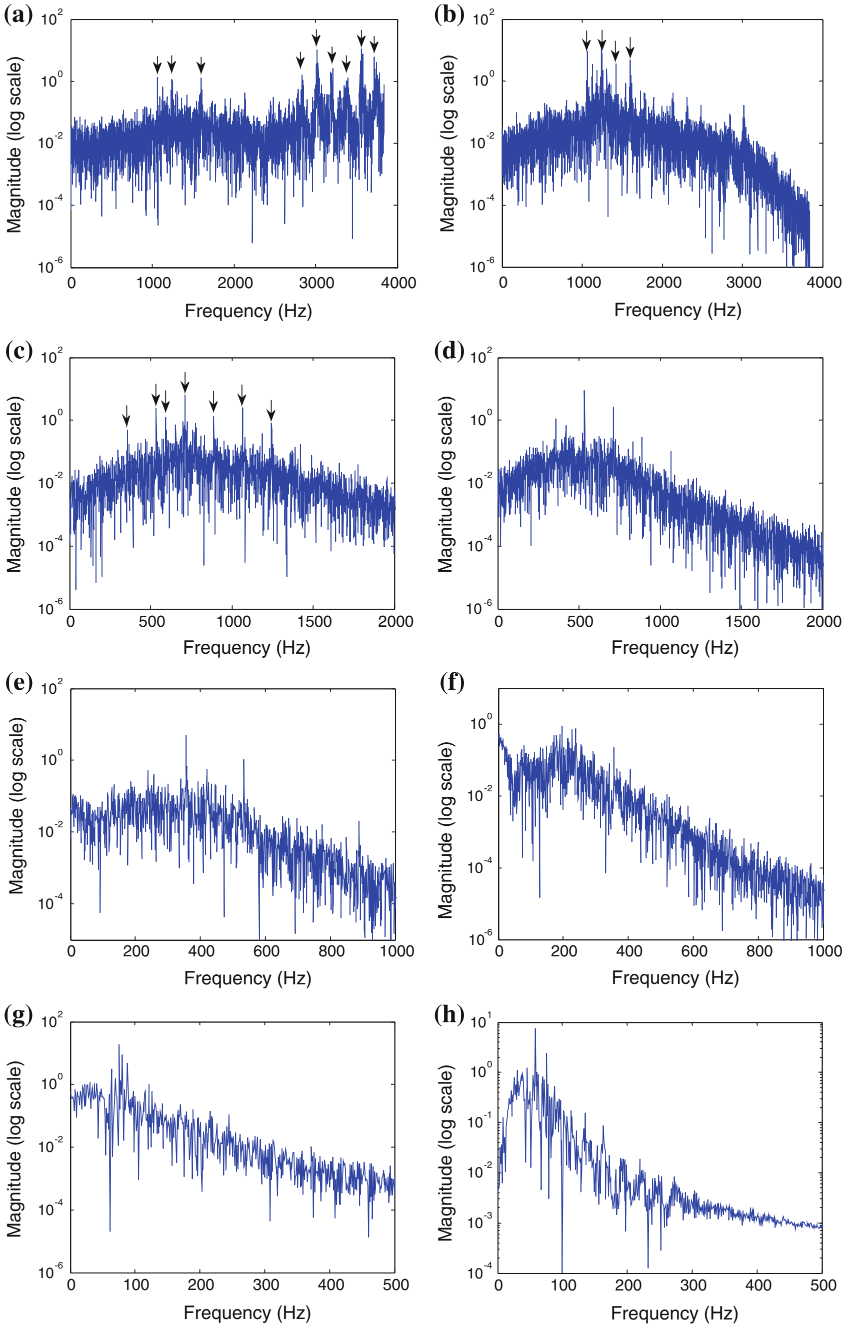


Fig. 1 Power spectra of the selected IMFs of bearing fault signal: **a** IMF 1, **b** IMF 2, **c** IMF 3, **d** IMF 4, **e** IMF 5, **f** IMF 6, **g** IMF 7, and **h** IMF 8

Table 2 PECs for IMFs of the *Jing* sound

IMF index	EI(×10^2)	Threshold	PEC	PEC DO
1	2.3741		0.4939	1
2	0.8886		0.1808	2
3	0.3304		0.0628	3
4	0.1637		0.0366	4
5	0.0109		–	–
6	0.0053		–	–
7	0.0018		–	–
8	0.0014		–	–
9	0.0013	1.3341	–	–
10	0.0012		–	–
11	0.0008		–	–
12	0.0001		–	–
13	0.0001		–	–
14	0.0000		–	–
15	0.0000		–	–
16	0.0000		–	–
17	0.0000		–	–

EI energy of IMF, *DO* descending order

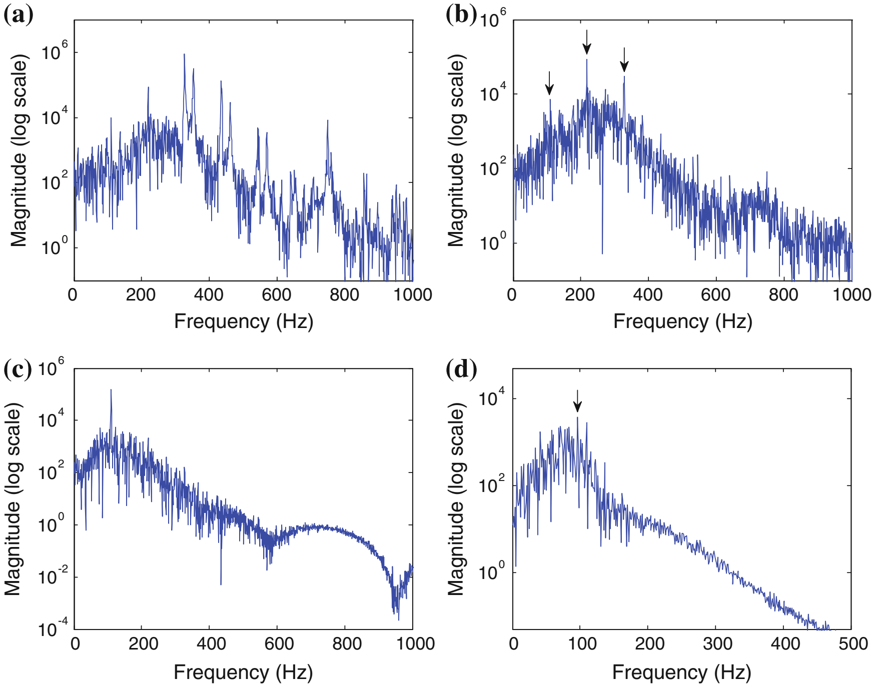


Fig. 2 Power spectra of the selected IMFs of the *Jing* sound: **a** IMF 1, **b** IMF 2, **c** IMF 3, and **d** IMF 4

Table 3 Comparison of eigen frequency (EF) and partial tone's frequency ratio (PFR) of the *Jing*

In [11]	EF (Hz)	112	176	196	228	304	340	416	456
	PFR	1	1.57	1.75	2.04	2.71	3.04	3.71	4.07
In [12]	EF (Hz)	122	202	205	301	308	367		
	PFR	1	1.65	1.67	2.46	2.51	3		
This paper	EF (Hz)	110	218	327	353	435	461	546	749
	PFR	1	1.98	2.97	3.21	3.95	4.19	4.96	6.8

5 Conclusion and Future Work

A modified EEMD including best IMF selection algorithm was described and its applications to the induction motor fault signal and the musical signal were shown. The IMF selection algorithm based on the PEC defined as the ratio between energy of the IMF and the total energy, was reasonable and efficient to extract the best IMFs including features from the target signals. Extracted IMFs contained physical significance such as ball pass frequency representing the bearing fault and harmonic structure of the musical instrument sound. However, we assumed the specific signal containing significant feature had high energy and the regularity of the partial distribution. In other words, we may not assure the effectiveness and validity of the proposed algorithm when it is employed to the noise-like signal in which harmonic structure is hardly observed or the amplitude of spectral peaks for certain oscillation is not conspicuous. Therefore, it is necessary to provide additional analysis about non-harmonic (or noise-like) signal and future work in making the algorithm more stable is required.

Acknowledgment This work was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (NRF-2012R1A1B6002600).

References

1. N.E. Huang, Z. Shen, S.R. Long, M. Wu, H. Shih, N. Zheng, C. Yen, C.C. Tung, H.H. Liu, The empirical mode decomposition and the Hilbert spectrum for non-linear and non-stationary time series analysis, in *Proceedings of the Royal Society London,* vol. A454, pp. 903–995
2. Y. Yu, Y. Dejie, C. Junsheng, A roller bearing fault diagnosis method based on EMD energy entropy and ANN. J. Sound Vib. **294**(1), 269–277
3. C. Junsheng, Y. Dejie, T. Jiashi, Y. Yu, Application of SVM and SVD technique based on EMD to the fault diagnosis of the rotating machinery. Shock Vibr. **16**(1), 89–98
4. Z.K. Peng, P.W. Tse, F.L. Chu, An improved Hilbert–Huang transform and its application in vibration signal analysis. J. Sound Vibr. **286**(1), 187–205

5. S. Cho, Analysis of the Jing sound using empirical mode decomposition, in *Proceedings of the Acoustical Society of Korea Spring Conference*, 9–10 May, 2013, Inchon, Korea, pp. 66–69
6. S. Cho, Y. Seo, Extraction of significant features using empirical mode decomposition and its application, Lecture Notes in Engineering and Computer Science, in *Proceedings of The World Congress on Engineering and Computer Science 2013, WCECS 2013*, 23–25 Oct, 2013, San Francisco, USA, pp. 527–530
7. S. Cho, Intrinsic frequency analysis for the percussion instrument based on the recorded sound, in *Proceedings of the Korea Institute of Signal Processing and Systems Summer Conference*, 12–13 July, 2013, Ulsan, Korea, pp. 98–100
8. Z. Wu, N. E. Huang, Ensemble empirical mode decomposition: a noise-assisted data analysis method. Adv. Adapt. Data Anal. **1**(1), 1–41
9. N.E. Huang, M.C. Wu, S.R. Long, S.S. P. Shen, W. Qu, P. Gloersen, K.L. Fan, A confidence limit for empirical mode decomposition and Hilbert spectral analysis, in *Proceedings of the Royal Society London*, vol. A459, pp. 2317–2345
10. B.S. Yang, T. Han, Z.T. Yin, Fault diagnosis system of induction motor using feature extraction, feature selection and classification algorithm, JSME Int. J. Ser. C **49**(3), 734–741
11. H.S. Kwon, Y.H. Kim, M. Rim, Acoustical characteristics of the Jing: an experimental observation using planar acoustic holography. J. Acoust. Soc. Korea **16**(2E), 3–13
12. H.J. Lee, H. Park, D.N. Lee, Manufacturing processes and acoustics of Jing and Kkwaenggwari. Bull. Korean Inst. Metals Mater. **12**(1), 10–18

Chapter 30
High Precision Numerical Implementation of Bandlimited Signal Extrapolation Using Prolate Spheroidal Wave Functions

Amal Devasia and Michael Cada

Abstract An efficient and reliable yet simple method to extrapolate bandlimited signals up to an arbitrarily high range of frequencies is proposed. The orthogonal properties of linear prolate spheroidal wave functions (PSWFs) are exploited to form an orthogonal basis set needed for synthesis. A significant step in the process is the higher order piecewise polynomial approximation of the overlap integral required for obtaining the expansion coefficients accurately with very high precision. A PSWFs set having a fixed Slepian frequency is utilized for performing extrapolation. Numerical results of extrapolation of some standard test signals using our algorithm are presented, compared, discussed, and some interesting inferences are made.

Keywords Bandlimited signals · High-precision numerical integration · Linear prolate spheroidal wave functions · Overlap integral · Signal extrapolation · Slepian series

1 Introduction

A remarkable discovery was made about half a century ago by David Slepian, an American mathematician, and his colleagues on a special set of functions called prolate spheroidal wave functions (PSWFs). These functions, also known as Slepian prolate functions, were bandlimited and exhibited interesting orthogonality relations. They are normalized versions of the solutions to Helmholtz wave equation [1] in spheroidal coordinates. In his paper [2–5], Slepian proposed the

A. Devasia (✉) · M. Cada
Department of Electrical and Computer Engineering, Dalhousie University, Halifax,
NS B3H 4R2, Canada
e-mail: amal.devasia@dal.ca

M. Cada
e-mail: michael.cada@dal.ca

H. K. Kim et al. (eds.), *Transactions on Engineering Technologies*,
DOI: 10.1007/978-94-017-9115-1_30,

idea of bandlimited signal extrapolation using PSWFs. Generating this set of functions practically seemed difficult because of the complexity involved and limited computational capabilities existed. Hence, there hasn't been any significant interest in this field up until very recently. However, with advanced numerical techniques and superior computational power, there has been noticeable activity in this field within the past decade [6–9]. In [10], Senay et al. proposed sampling and reconstruction of bandlimited as well as non-bandlimited signals using Slepian functions. They discussed the idea of modifying the Whittaker–Shannon sampling theory by replacing the sin*c* basis by Slepian functions for reconstruction of signals. Further to this, in [11, 12], they showed signal reconstruction using non-uniform sampling and level-crossing sampling with Slepian functions.

Considering signal extrapolation to be the subject of this paper and not just reconstruction or interpolation, we will shift our focus to the recent advancements in this context. While we consider the signals to be bandlimited in the Fourier transform domain, much attention has recently been on extrapolation of signals bandlimited in linear canonical transform (LCT) domain, this being a four-parameter family of linear integral transform [13, 14] that generalizes Fourier transform as one of its special cases. For extrapolation of LCT bandlimited signals, several iterative and non-iterative algorithms have been proposed [15–18]. Most of the iterative algorithms are centered on modifying the Gerchberg–Papoulis (GP) algorithm [19, 20] that relies on successive reduction of error energy. Although theoretical convergence of the result has been shown, there is still some uncertainty associated with the swiftness with which this is achieved. With respect to the non-iterative algorithms proposed in [16], the authors themselves admit that the extrapolation could become unstable with an increase in the number of observations. A comparison of the extrapolation of an LCT bandlimited signal, using an iterative GP algorithm and another algorithm based on signal expansion into a series of generalized PSWFs [18] is presented in [17]. The comparison showed better results for the iterative method (proposed by [17]) over the one described in [18], in terms of the normalized mean square error (NMSE). Gosse, in [21], performed Fourier bandlimited signal extrapolation by handling lower and higher frequencies of the signal separately. He used PSWFs for extrapolating lower frequency components while the higher frequencies were dealt with compressive sampling [22, 23] algorithms. The efficiency of the proposed method was highly dependent on the correlation between low and high frequencies in the signal (it should be weak for better results), the existence of a sparse representation of higher frequencies in the Fourier basis, and on a reasonable choice of extrapolation domain.

In this paper, we propose a non-iterative and simple method for bandlimited signal extrapolation valid up to an arbitrarily high range of frequencies using Slepian functions. Although we concentrate mainly on Fourier bandlimited signals, it however might also be applied for LCT bandlimited cases as is shown in one of our results below. Several comparisons are made with the results obtained in earlier related publications. They show that, within the prescribed bandwidth, the proposed method is far superior over several other methods referenced in this paper. PSWFs for analysis purposes need to be computed accurately and with rather high precision. Here, we rely on a proprietary algorithm developed theoretically and implemented numerically by Cada [24], for accurately generating the linear prolate functions (one-dimensional PSWFs, henceforth abbreviated as LPFs) set with desired high precision. Once the LPFs set is obtained with the corresponding eigenvalues (discussed later in this paper), they are employed in our algorithm for extrapolation. Here, we do not consider the storage of LPFs set as an issue (as put forward by Shi et al. in [17]) to be addressed, as it is not the subject of this paper.

Cada's algorithm exploits robust properties of certain formulae derived that are efficient, accurate and suitable for fast numerical evaluations of linear prolate functions and their eigenvalues. Previous methods [Slepian, Flammer [25], etc.] required lengthy cumbersome calculations with slowly converging series and necessary approximations that led to insurmountable numerical problems and/or failing when higher orders were concerned. Prolate functions and the eigenvalues change their properties drastically at certain parameter values (see below), which has caused described problems. Even standard professional high-quality commercial packages such as Mathematica or Matlab fail to compute these functions and eigenvalues correctly or at all for such a combination of parameters that is critical for extrapolation applications. Our algorithm enables to break through this numerical barrier and makes it possible to calculate linear prolate functions and their eigenvalues correctly for basically any parameters.

Signal extrapolation is an extension of a signal, $f(t)$, beyond the interval in which it is known to the observer. The region in which the signal is known is called the observation interval (here, $[-t_0, t_0]$). Bandlimited signals are bound in the frequency domain; their Fourier transform, $F(\omega)$, vanishes beyond a particular finite frequency interval. Thus if,

$$F(\omega) = 0, \quad |\omega| > \Omega, \tag{1}$$

then $f(t)$ is said to be Ω-bandlimited.

The following sections of the paper are organized as follows. In Sect. 2, LPFs and their relevant properties are discussed. The various steps involved in our proposed extrapolation algorithm are described throughout in Sect. 3. Section 4 is devoted to presenting the actual extrapolated results of various test functions, their comparison and error analysis. Finally, in Sect. 5, we conclude with our inferences and possible future prospects.

2 Linear Prolate Functions and Its Properties

Bandlimited signal extrapolation using PSWFs was first discussed by Slepian and his colleagues in [2]. They explained the use of PSWFs, or more precisely, linear prolate functions (LPFs) as an orthogonal basis set for decomposition and reconstruction of the signal using analysis and synthesis equations. Linear prolate functions are one-dimensional PSWFs denoted by $\Psi_n(c,t)$, where n is the order of LPF (non-negative integer), t is the time parameter and c is the bandwidth parameter also known as Slepian frequency. LPFs can be evaluated using:

$$\Psi_n(c,t) = \sqrt{\frac{\lambda_n(c)/t_0}{\int_{-1}^{1}(S_{0n}(c,t))^2 dt}} S_{0n}\left(c, \frac{t}{t_0}\right), \tag{2}$$

where $\lambda_n(c)$ is the eigenvalue of sin$c$ kernel with $\Psi_n(c,t)$ as eigenfunction (a measure of concentration of signal in the observation interval $[-t_0,\ t_0]$), t_0 is the observation boundary of the interval in which the function is known, and $S_{m,n}(c,\ \eta)$ are the angular solutions of the first kind to Helmholtz wave equation [1]. The eigenvalue $\lambda_n(c)$ is given by,

$$\lambda_n(c) = \frac{2c}{\pi}[R_{0,n}(c,1)^2], \tag{3}$$

where $R_{m,n}(c,\ \varepsilon)$ are the radial solutions of the first kind to Helmholtz wave equation.

Numerical evaluation of the LPFs set along with their corresponding eigenvalues practically seemed very difficult as it involved finding precise numerical values of the angular ($S_{m,n}(c,\ \eta)$) and radial ($R_{m,n}(c,\ \varepsilon)$) solutions. For obtaining this, a typical power series expansion was used which was predetermined by the association of Legendre and spherical Bessel functions to the angular and radial solutions respectively. Interested readers are referred to [8, 9, 26] for more details on LPFs derivation. These LPFs along with their corresponding eigenvalues have been used in our extrapolation algorithm. It should be noted that since extrapolation relies heavily on values of $\Psi_n's$ and $\lambda_n's$ when $n > 2c/\pi$, it is of paramount importance that one computes them accurately and with high precision. Ours is the first algorithm that offers such a capability.

2.1 Properties of LPFs

LPFs have many interesting properties of which some of the relevant ones related to this study are discussed below.

(i) *Bandlimited*
Bandlimiting property of LPFs is denoted by a free bandwidth parameter (Slepian frequency) c given by:

$$c = \Omega_0 t_0, \tag{4}$$

where Ω_0 is the finite bandwidth of $\Psi_n(c,t)$ for a given order n.

(ii) *Symmetry*

LPFs exhibit even and odd symmetries based on their integer order n. If n is even, $\Psi_n(c,t)$ is even symmetric. If n is odd, then it is odd symmetric.

(iii) *Orthogonality*

LPFs are linearly independent and orthogonal over finite (5) as well as infinite (6) intervals, unlike, for example, trigonometric functions that are orthogonal only over a finite domain.

$$\int_{-t_0}^{t_0} \Psi_n(c,t)\Psi_m(c,t)dt = \begin{cases} \lambda_n(c) & for\, n=m, \\ 0 & otherwise. \end{cases} \tag{5}$$

$$\int_{-\infty}^{\infty} \Psi_n(c,t)\Psi_m(c,t)dt = \begin{cases} 1 & for\, n=m, \\ 0 & otherwise. \end{cases} \tag{6}$$

where n, m are non-negative integers.

(iv) *Invariance to Fourier transform*

Fourier transforms of LPFs over both finite (7) and infinite (8) intervals are simply scaled versions of themselves.

$$\int_{-t_0}^{t_0} \Psi_n(c,t)\mathrm{e}^{j\omega t}dt = j^n\left(\frac{2\pi\lambda_n(c)t_0}{\Omega_0}\right)^{1/2}\Psi_n\left(c,\frac{\omega t_0}{\Omega_0}\right) \tag{7}$$

$$\int_{-\infty}^{\infty} \Psi_n(c,t)\mathrm{e}^{j\omega t}dt = j^n\left(\frac{2\pi t_0}{\Omega_0}\right)^{1/2}\Psi_n\left(c,\frac{\omega t_0}{\Omega_0}\right) \tag{8}$$

Expressions (7) and (8) show LPFs' invariance to Fourier transforms and are a further proof of their bandlimiting property.

3 Extrapolation Algorithm

This section explains the various steps with which we implemented our extrapolation algorithm.

3.1 Analysis and Synthesis

In general, any bandlimited signal can be decomposed into a linear combination of weighted orthogonal basis functions using the relation:

$$f(t) = \sum_{n=-\infty}^{\infty} \gamma_n \Phi_n(t) \quad \text{(synthesis)}, \tag{9}$$

where *f*(*t*) is the signal, γ_n is a set of scalar coefficients, and $\Phi_n(t)$ is the orthogonal basis set.The set of scalar coefficients is found from:

$$\gamma_n = \frac{\int_{-\infty}^{\infty} f(t)\Phi_n(t)dt}{\int_{-\infty}^{\infty} \Phi_n(t)\Phi_n(t)dt} \quad \text{(analysis)}. \tag{10}$$

Employing LPFs as the orthogonal basis set for a fixed Slepian frequency *c* yields:

$$\gamma_n(c) = \lambda_n^{-1}(c) \int_{-t_0}^{t_0} f(t)\Psi_n(c,t)dt \quad \text{(analysis)}, \tag{11}$$

where $\int_{-t_0}^{t_0} f(t)\Psi_n(c,t)dt$ is known as the overlap integral, and $\gamma_n(c)$ are the scalar expansion coefficient for a given order *n* of LPF.

The synthesis equation, used for signal extrapolation, is given by:

$$f(t) \cong \sum_{n=0}^{N} \gamma_n(c)\Psi_n(c,t) \quad \text{(synthesis)}, \tag{12}$$

where *N* is the truncation value for the order *n*.

The analysis and synthesis equations described in (11) and (12) respectively become the basis of our signal extrapolation algorithm.

3.2 LPFs Set

A LPFs set with a fixed Slepian frequency *c* was used as a potential orthogonal basis set along with its corresponding eigenvalues for the proposed algorithm. These functions were discretized in time for numerical implementation. Each of these sampled data has very high numerical precision of about 200 digits. The specifics are as follows:

$$c = 20\pi,\ n : 0 \to 101,\ t \to [-1.900, 1.900] \quad (\Psi_n(c,t)Set)$$

The time parameter *t* is specified with three digits of precision after the decimal point as the sampling period in the time axis is 0.001 s. Generally, for any LPFs set with a fixed *c*, as the order *n* increases, the concentration of the LPFs within $[-t_0,\ t_0]$ decreases. For $n = 2c/\pi$, the signal's maximum concentration reaches the boundary of the observation interval (see Fig. 1).

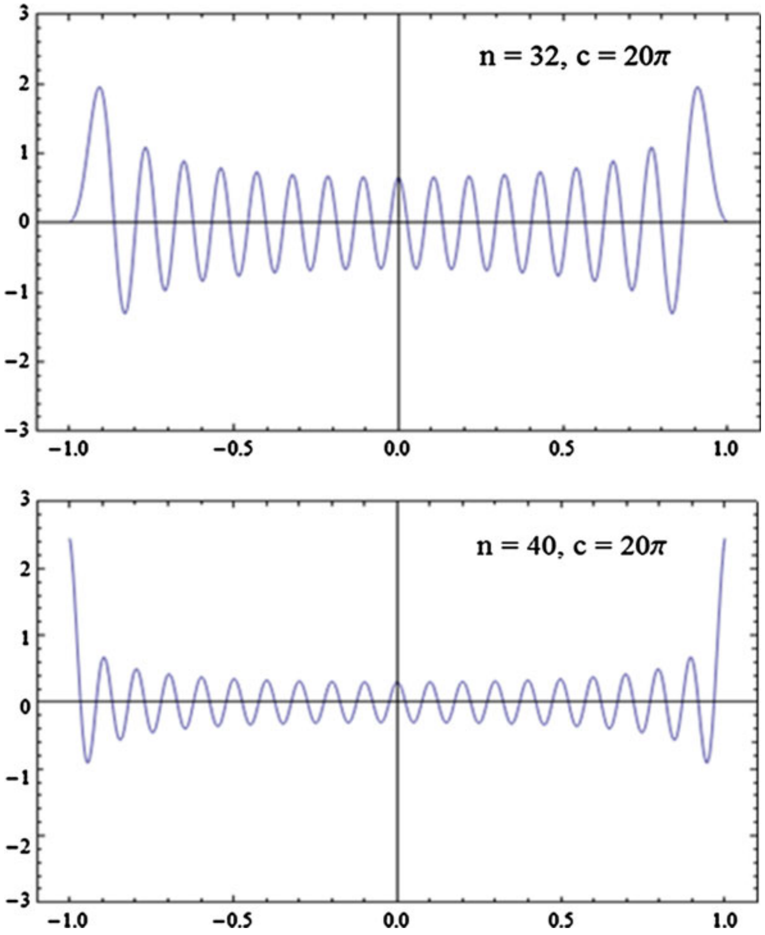


Fig. 1 Dependency of signal concentration on n for LPFs set with $c = 20\pi$

3.3 Overlap Integral

One can notice from the analysis (11) and synthesis (12) equations using LPFs that $f(t)$, the function to be extrapolated, is well defined in $[-t_0, t_0]$, i.e. $[-1, 1]$ (t_0 chosen to be 1). Numerical values of $\lambda_n(c)$ and $\Psi_n(c,t)$ as a set for a given c are also known (see Sect. 2). The only unknown factor is an efficient method to calculate the overlap integral given by $\int_{-1}^{1} f(t)\Psi_n(c,t)dt$. Efficient estimation of overlap integral is of paramount importance in obtaining accurate results for extrapolation. As LPFs, $\Psi_n(c,t)$, and eigenvalues, $\lambda_n(c)$, are required to be of high precision for high orders of n, the eigenvalues tend to become extremely small, close to zero, which makes this essentially a problem of high precision numerical integration. If and when computed inaccurately, the coefficients of expansion, $\gamma_n(c)$, of the synthesis Eq. (12) assume extremely large values for such $n's$ that are crucial and irreplaceable for extrapolation purposes. This, in turn, causes enormous numerical errors that thus render extrapolated signals completely incorrect and useless.

3.4 Calculation of Overlap Integral

A simple and efficient way to compute overlap integral is proposed and implemented to obtain satisfactory results. To make it simple and generic, polynomial approximation of the discrete samples of the scalar product $f(t)\Psi_n(c,t)$ was chosen. Our primary focus was to obtain the right polynomial approximation for the underlying common function, $\Psi_n(c,t)$ (for a given c), in any scalar product, irrespective of the bandlimited function $f(t)$. Another major task was to choose the right truncation value N for synthesis (12), which we determined by examining the behavior of the scalar coefficients $\gamma_n(c)$ obtained in analysis (11). After a series of thorough investigations using different kinds of polynomial interpolation and numerical integration techniques, it was found that piecewise polynomial approximation is best suited for the particular LPF set that was used.

Direct method of piecewise polynomial interpolation [27] is used. Given i discrete data points, (x_0, y_0) to (x_{i-1}, y_{i-1}), it can be approximated to a polynomial of order $i - 1$ as:

$$y = \beta_0 + \beta_1 x + \cdots + \beta_{i-1} x^{i-1}, \tag{13}$$

where the coefficients (β) can be found by solving this linear system of equations:

$$\begin{pmatrix} y_0 \\ y_1 \\ \vdots \\ y_{i-1} \end{pmatrix} = \begin{pmatrix} 1 & x_0 & x_0^2 & \cdots & x_0^{i-1} \\ 1 & x_1 & x_1^2 & \cdots & x_1^{i-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{i-1} & x_{i-1}^2 & \cdots & x_{i-1}^{i-1} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{i-1} \end{pmatrix}. \tag{14}$$

Applying this technique to our problem, there are 1,001 samples in the time interval [0, 1] and 2,001 samples in [−1, 1]. The whole interval of [−1, 1] is divided into eight equal segments, thus there are 251 samples in each segment of the samples of the scalar product $f(t)\Psi_n(c,t)$. Applying piecewise polynomial approximation to each segment containing 251 samples, one obtains a polynomial of order 250 for each segment. The desired overlap integral is finally obtained by directly integrating each segment.

4 Results and Discussion

For implementing our algorithm we used Mathematica, a software tool that is excellent for high precision computing. Extrapolation was carried out for some selected known test functions using the aforementioned LPFs set. Some of these results can also be found in [28, 29] and is used here for making comparisons and drawing conclusions. The only restriction imposed on the selected signals was that its maximum frequency should be less than or equal to that of the corresponding LPFs set used for their extrapolation. In the results shown below, the same extrapolation formula was employed for both reconstructing the signal within as well as extrapolating it beyond the interval $[-1, 1]$. To show the error estimates with respect to the original signal, common logarithm of the absolute error between the extrapolated and original data is also plotted. Extrapolation errors of the order of up to 10^{-2} or 10^{-3} are considered acceptable.

The range of extrapolation is limited mainly due to the series truncation, N, the value of which should be at least greater than $2c/\pi$ for performing extrapolation. We also verified the effective extrapolation range that can be achieved analytically by using simple sinusoidal signals, for which the analytical expressions are known [8]. We found that for the particular LPFs set used, the truncation value N was more or less equal to 100 (out of the total order of $n = 101$ being considered), which is much greater than $2c/\pi$ (with $c = 20\pi$, this equals $40 \ll 100$), thus allowing signal extrapolation.

The signals for which the extrapolation is shown in Figs. 2 and 3, although not exactly bandlimited because of the Gaussian functions involved, also gave good results (as seen in Fig. 3), which are comparable to what was achieved using strictly bandlimited case (Fig. 4). Figure 3 also demonstrates the effective extrapolation of a signal that was analyzed in a different observation interval, $[-13.5, -11.5]$, than that of the LPFs set used, $[-1, 1]$, for obtaining the scalar expansion coefficients needed for its synthesis.

To find the effectiveness of our algorithm, the results obtained were also compared with those obtained in earlier publications. In [21], Gosse used the same test signal given here as signal 1 (see below) for extrapolation. He considered it as a composite signal composed of low (first term) and high (second term) frequencies, and extrapolated the lower frequency part using a truncated prolate series expansion, while the higher frequency part was handled by compressive sampling techniques. The effective extrapolation range in our case (for signal 1) is significantly improved when compared with their results (see [21], p. 1277; and Fig. 2 of this paper). In the same context, the error analysis also shows better results as our method has absolute error magnitude around 10^{-38} (while it is 10^{-2} in [21]) within the reconstruction interval $[-1, 1]$. We also obtained better ratios of error magnitudes (varying smoothly from the order of 10^{-36}–10^{-3} using our algorithm, while oscillating between 10^{-2} and 10^{-1} in [21]) in the effective extrapolation region, i.e. outside $[-1, 1]$ (see absolute error plots of Fig. 2).

(1)

$$f(t) = \mathrm{e}^{-2t^2}\mathrm{e}^{3t}\sin(\pi t)\cos(3\pi t) + 0.5[\sin(5\pi t) - \cos(7\pi t)]$$

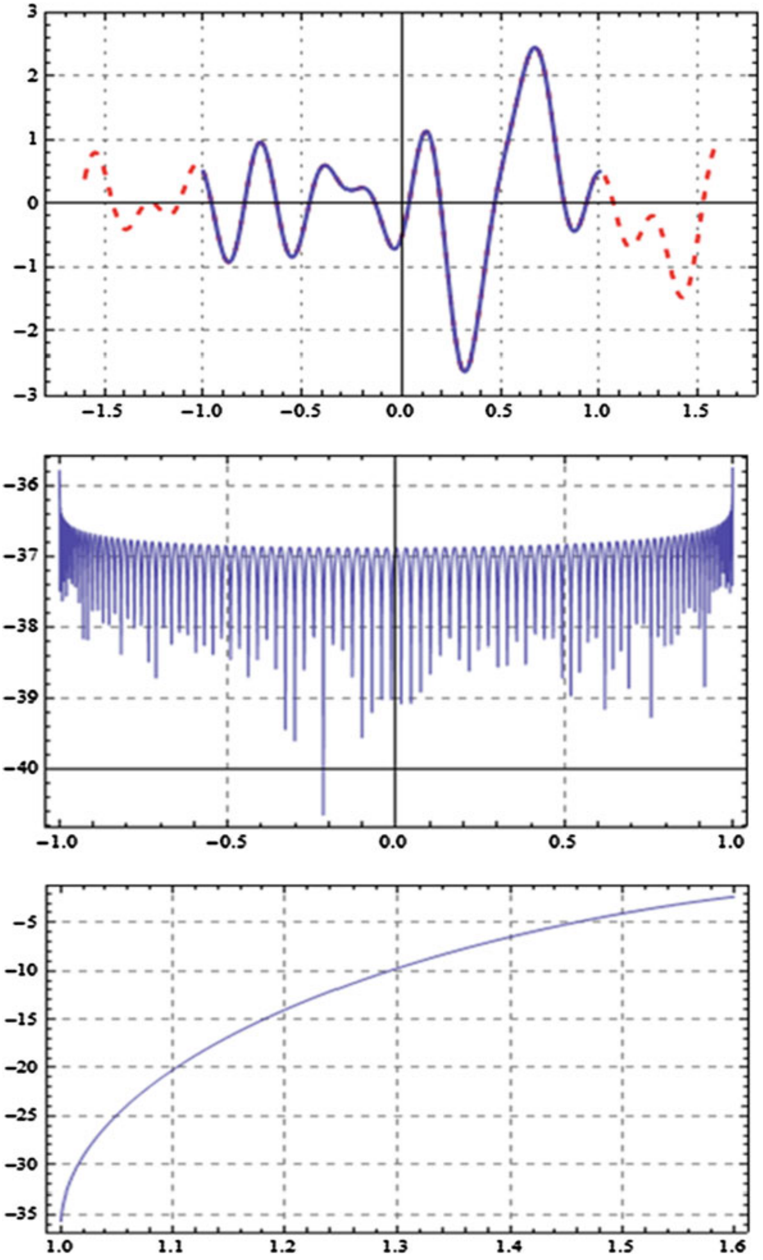


Fig. 2 $f(t)$ original (*solid*) and extrapolated (*dashed*) (*top*), logarithm of absolute error (*bottom two*) versus time; LPFs set with $c = 20\pi$ used

(2)

$$f(t) = \frac{1}{2}\mathrm{e}^{-2t^2}\cos(5\pi t)$$

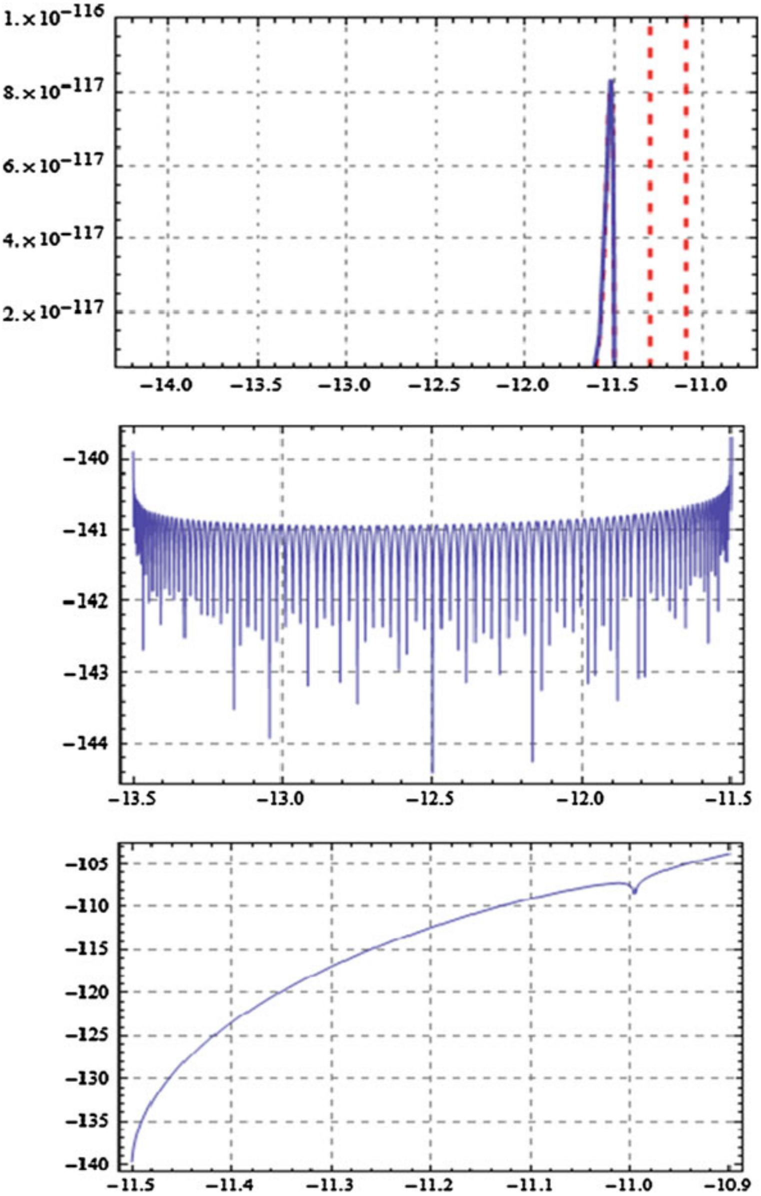


Fig. 3 $f(t)$ original (*solid*) and extrapolated (*dashed*) (*top*), logarithm of absolute error (*bottom two*) versus time; LPFs set with $c = 20\pi$ used

(3)

$$f(t) = \cos\left(20\pi t - \left(\frac{\pi}{11}\right)\right) + \cos\left(\frac{20\pi t}{7}\right) - \cos\left(\frac{15\pi t}{2}\right)$$

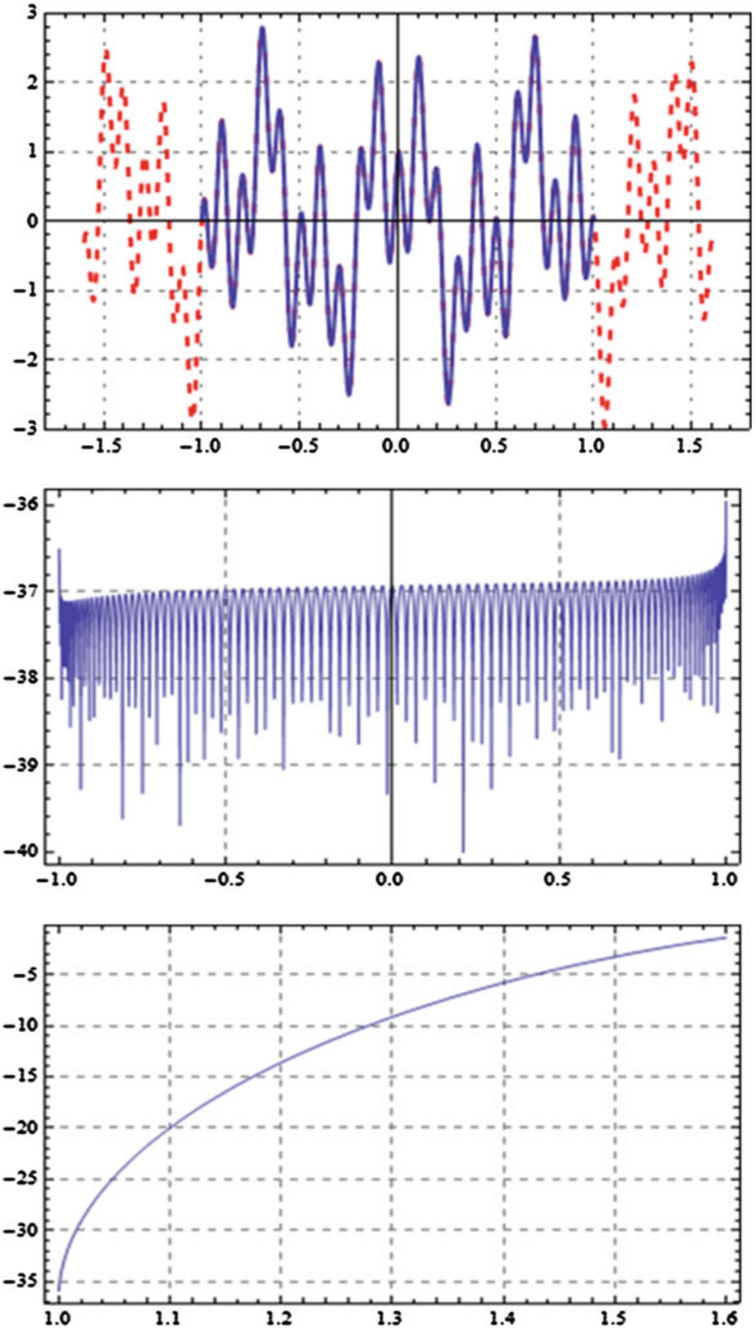


Fig. 4 $f(t)$ original (*solid*) and extrapolated (*dashed*) (*top*), logarithm of absolute error (*bottom two*) versus time; LPFs set with $c = 20\pi$ used

(4)

$$f(t) = \frac{20}{\pi} e^{j2t^2} \operatorname{sinc}\left(\frac{20t}{\pi}\right)$$

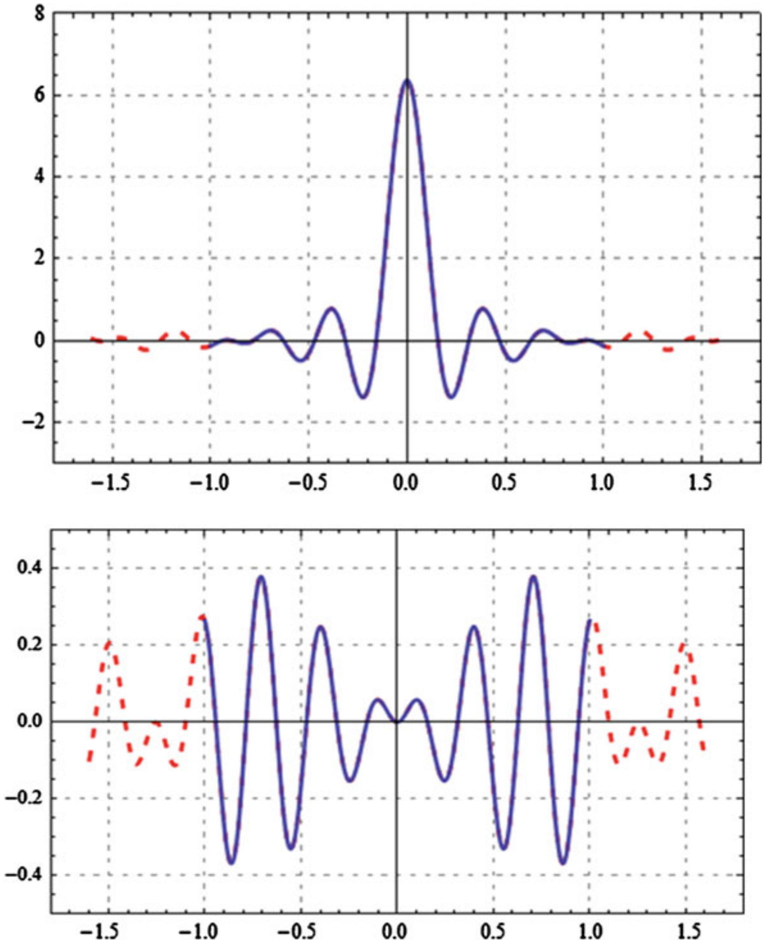


Fig. 5 Real part of $f(t)$ original (*solid*) and extrapolated (*dashed*) (*top*), imaginary part of $f(t)$ original (*solid*) and extrapolated (*dashed*) (*bottom*) versus time; LPFs set with $c = 20\pi$ used

As mentioned earlier, we also performed extrapolation on an LCT bandlimited signal to compare our method with existing ones. We chose the same signal used by Shi et al. in [17]. It is given as signal 4 (the sin*c* function used is a normalized sin*c* function) in the results. The extrapolation of the real and imaginary parts of the signal is shown in Fig. 5. The performance was measured using an error norm known as the normalized mean-square error (NMSE); following their [17] notation it is defined as:

$$NMSE = \frac{\|f_e - f\|^2}{\|f\|^2}, \tag{15}$$

where f_e is the extrapolated signal and f is the original signal. The NMSE of the extrapolated signal using our algorithm is 8.430×10^{-7}. The corresponding NMSE using the iterative algorithm proposed in [17] is 1.037×10^{-4}, and NMSE using the generalized PSWFs expansion method proposed in [18] is only 0.685. Thus our algorithm is shown performs superiorly even for LCT bandlimited signals, albeit it is not the fundamental goal of this paper.

5 Conclusion

We have presented an implemented robust and efficient algorithm for bandlimited signal extrapolation valid up to basically an arbitrarily high range of frequencies. Even though the algorithm is complex in the sense that it involves time consuming calculations and tedious computations with big matrices of very high precision, the overall idea is simple and easy to execute, thanks to the current computational speeds and available system memory. We believe that the accuracy with which the LPFs were computed with very high precision allowed this method to work efficiently thus making it suitable for extrapolating signals within the prescribed bandwidth. Characterizing the Slepian functions (LPFs set) finely and precisely into an appropriate polynomial expression is the key with which, this method could be extended to other LPF sets. During the characterization process, emphasis should be on incorporating more higher order (n) terms in the synthesis Eq. (12) thus making N sufficiently large for extrapolation. This is a promising development in the field of signal processing [15–17, 30, 31] and will be helpful in the characterization of both known and random bandlimited observations. It should be stressed, however, that the key to the successful better-than-others results of our extrapolation algorithm is, indeed, the accurate numerical evaluations of linear prolate functions and their eigenvalues employing our proprietary robust algorithm for computing them.

Acknowledgment The authors thank Jose Gonzalez-Cueto and William Phillips for their helpful suggestions on the material. Partial support by Applied Science in Photonics and Innovative Research in Engineering (ASPIRE), a program under Collaborative Research and Training Experience (CREATE) program funded by Natural Sciences and Engineering Research Council (NSERC) of Canada, is also acknowledged.

References

1. M. Abramowitz, I. Stegen, *Handbook of Mathematical Functions with Formulas, Graphs and Mathematical Tables* (Addison-Wesley, New York, 1972)
2. D. Slepian, H.O. Pollack, Prolate spheroidal wave functions, Fourier analysis, and uncertainty-I. Bell Syst. Tech. J. **40**, 43–63 (1961)
3. D. Slepian, Prolate spheroidal wave functions, Fourier analysis and uncertainty—IV: extensions to many dimensions; generalized prolate spheroidal functions. Bell Syst. Techn. J. **43**, 3009–3057 (1962)
4. D. Slepian, Some asymptotic expansions for prolate spheroidal wave functions. J. Math. Phys. **44**, 99–140 (1965)
5. D. Slepian, Some comments on Fourier analysis, uncertainty and modeling, SIAM Rev. **25**, 379–393 (1983)
6. K. Khare, N. George, Sampling theory approach to prolate spheroidal wave functions, J. Phys. A: Math. Gen. **36**, 10011–10021 (2003)
7. P. Kirby, Calculation of spheroidal wave functions, Comput. Phys. Comm. **175**(7), 465–472 (2006)
8. I.C. Moore, M. Cada, Prolate spheroidal wave functions, an introduction to the Slepian series and its properties. Appl. Comput. Harmon. Anal. **16**, 208–230 (2004)
9. H. Xiao, V. Rokhlin, N. Yarvin, Prolate spheroidal wave functions, quadrature and interpolation, Inverse Prob. **17**, 805–838 (2001)
10. S. Senay, L.F. Chaparro, A. Akan, Sampling and reconstruction of non-bandlimited signals using Slepian functions, in *EUSIPCO 2008*, Lousanne, 25–29 Aug 2008
11. S. Senay, L.F. Chaparro, L. Durak, Reconstruction of non-uniformly sampled time-limited signals using prolate spheroidal wave functions. Sig. Process. **89**, 2585–2595 (2009)
12. S. Senay, J. Oh, L.F. Chaparro, Regularized signal reconstruction for level-crossing sampling using Slepian functions. Sig. Process. **92**, 1157–1165 (2012)
13. M. Moshinsky, C. Quesne, Linear canonical transformations and their unitary representations. J. Math. Phys. **12**, 1772–1783 (1971)
14. H.M. Ozaktas, Z. Zalevsky, M.A. Kutay, *The Fractional Fourier Transform with Applications in Optics and Signal Processing* (Wiley, New York, 2000)
15. H. Zhao, R.Y. Wang, D.P. Song, D.P. Wu, An extrapolation algorithm for M-bandlimited signals. IEEE Signal Process. Lett. **18**(12), 745–748 (2011)
16. H. Zhao et al., Extrapolation of discrete bandlimited signals in linear canonical transform domain. Signal Process. (2013). http://dx.doi.org/10.1016/j.sigpro.2013.06.001
17. J. Shi, X. Sha, Q. Zhang, N. Zhang, Extrapolation of bandlimited signals in linear canonical transform domain, IEEE Trans. Signal Process. **60**(3), 1502–1508 (2012)
18. H. Zhao, Q.W. Ran, J. Ma, L.Y. Tan, Generalized prolate spheroidal wave functions associated with linear canonical transform. IEEE Trans. Signal Process. **58**(6), 3032–3041 (2010)
19. R. Gerchberg, Super-resolution through error energy reduction. Opt. Acta. **12**(9), 709–720 (1974)
20. A. Papoulis, A new algorithm in spectral analysis and band-limited extrapolation. IEEE Trans. Circuit Syst. **CAS-22**(9), 735–742 (1975)
21. L. Gosse, Effective band-limited extrapolation relying on Slepian series and ℓ^1 regularization. Comput. Math. Appl. **60**, 1259–1279 (2010)
22. E.J. Candès, Compressive sampling, in *Proceedings of the International Congress of Mathematicians*, Madrid, 2006
23. E.J. Candès, The restricted isometry property and its implications for compressed sensing. C. R. Acad. Sci. Paris, Ser. I **346**, 589–592 (2008)
24. M. Cada, Private communication, Department of Electrical and Computer Engineering, Dalhousie University, Halifax, NS, B3H 4R2, Canada, (2012)
25. C. Flammer, *Spheroidal Wave Functions* (Stanford Univ. Press, Stanford, 1956)

26. V. Rokhlin, H. Xiao, Approximate formulae for certain prolate spheroidal wave functions valid for large values of both order and band-limit. Appl. Comput. Harmon. Anal. **22**, 105–123 (2007)
27. A. Kaw, E. Kalu Egwu, D. Nguyen, *Numerical Methods with Applications*, 2nd ed. (Lulu, Raleigh, 2011) (Abridged)
28. A. Devasia, M. Cada, Extrapolation of bandlimited signals using Slepian functions, in *Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering and Computer Science 2013, WCECS 2013*, 23–25 October 2013, San Francisco, pp. 492–497
29. A. Devasia, M. Cada, Bandlimited Signal Extrapolation Using Prolate Spheroidal Wave Functions. IAENG Int. J. Comput. Sci. **40**(4), 291–300 (2013)
30. I. Kauppinen, K. Roth, Audio signal extrapolation-theory and applications, in *Proceedings of (DAFx-02)*, Hamburg, 26–28 Sept 2002
31. A. Kaup, K. Meisinger, T. Aach, Frequency selective signal extrapolation with applications to error concealment in image communication. Int. J. Electron. Commun. **59**, 147–156 (2005)

Chapter 31
Path Loss Prediction in Relay Station Environment

Masoud Hamid and Ivica Kostanic

Abstract Relays play important role in deployment of Long Term Evolution (LTE) and LTE-Advanced systems. This chapter addresses prediction of the propagation path loss on the link between eNodeB and relay stations. The path loss models are derived on a basis of an extensive measurement campaign conducted in 1,900 MHz frequency band. An effect of the relay station antenna height is studied and included in the path loss modeling. An antenna height correction factor is derived and included in the modeling. Finally, a relationship between the intercept, slope of the model and the relay antenna height is derived.

Keywords Channel model · LTE-Advanced · Path loss measurements · Propagation modeling · Relay antenna height correction factor · Relay

1 Introduction

LTE-Advanced is the upcoming global cellular technology that offers a very high performance air interface. One of the enabling technologies that supports such a performance are radio relays. Within LTE and LTE-Advanced, radio relays are used to extend coverage, enhance capacity, increase throughput and provide overall increase in the network performance [1–3].

In addition to performance enhancements, the relays reduce cost of the network deployment and facilitate speed of network roll-outs [4]. In many cases, relaying technique is considered as a viable solution for replacement of base stations.

M. Hamid (✉) · I. Kostanic
Electrical and Computer Engineering Department, Florida Institute of Technology,
Melbourne, FL 32901, USA
e-mail: mhamid2010@my.fit.edu

I. Kostanic
e-mail: kostanic@fit.edu

H. K. Kim et al. (eds.), *Transactions on Engineering Technologies*,
DOI: 10.1007/978-94-017-9115-1_31,

Relays cost significantly less than base stations. When deployed, relays act like base stations but without the need of wired connection to the backhaul.

From the network planning perspective one needs to successfully model the impact of the relay deployment within an LTE network. The first step in this modeling is the prediction of the path loss on the link between an eNodeB and a relay station. A literature review shows that there is a general shortage of measured data collection to help empirical understanding of the propagation conditions in relay environment. Nevertheless, there have been several studies that discussed this topic.

Some of propagation models have been suggested by 3GPP (third Generation Partnership Project) [3], WINNER (Wireless World Initiative New Radio) [5] and IEEE 802.16j task group [6]. Nonetheless, one notes that a general limitation of these models is that they are developed from already existing macroscopic propagation models which were derived under completely different assumptions. Hence their applicability to relay scenarios needs to be tested. Another limitation is that they were derived for certain levels of relay antenna height and therefore their validation for different heights still needs to be performed. The effect of receive antenna height on the received signal level in a LTE-Advanced relaying scenario was investigated [7]. Even though general dependence of path loss on relay station antenna height was obtained, study would have been more complete if the authors had proposed an empirical path loss model which can be applied in similar scenarios. Similarly, a new propagation model for relay scenarios was proposed [8]; however, this model was suggested just for urban environments. Webb and coauthors discussed a related work [9]; however, the maximum height of relay station antenna was limited to 5 m which is too low for most relay scenarios [3, 5].

When deployed, a multi-hop relaying system traverses two radio links. The first link is the link between the eNodeB and the relay and it is referred to as the backhaul link. The second link is the link between the relay and the end user and it is referred to as the access link. The study presented here focuses on the path loss encountered on eNodeB-relay link only. The main objective of the research described in this chapter is to develop statistical path loss models for outdoor relaying systems in 1,900 MHz frequency band [10]. The path loss modeling takes into account the impact of the relay antenna height and therefore, an antenna height correction factor is included in the modeling. The models are based on field measurements conducted in suburban environment.

2 Experimental Setup

2.1 Equipment Description

The data collection system consists of a transmitter, transmit antenna, receiver, receive antenna, GPS (Global Positioning System) antenna and a laptop with installed measurement software called Wireless Measurement System (WMS) from

Grayson wireless. The software is used to measure the strength of the received signal and map its value on the spectrum tracker screen along with the corresponding frequency. The transmitter is on the top of a multi-story building. The receiver is placed in a "boom-lift" as shown in Fig. 1 and moved between locations. The measurements are conducted over a period of couple of weeks, with essentially no changes in weather pattern and vegetation. Before measurements were conducted, spectrum clearing of the area was performed to verify that the frequency used for the path loss measurements is free from any sources of radiation.

2.2 Environment Description

The measurements obtained in this study are collected in a typical US suburban environment of Melbourne, FL, USA. Most houses in the selected area are single to double stories and their heights are about 4–9 m. In general, most of the buildings are made of wooden structures with exception of few buildings that are made of combined materials; concrete for frame or body and timber or plastered bricks for walls, glass for windows and concrete for floors. Few buildings have flat roofs while most of one story houses and two stories residential apartments have pitched roofs. The terrain in general is flat with moderate tree densities. Trees height is up to 13 m.

2.3 Measurement Procedure

The measurements are collected in 1,900 MHz band which is one of the principal bands for the deployment of the LTE and LTE-Advanced. The parameters associated with the measurements are provided in Table 1. As seen, the study is conducted using four different relay heights ranging from 4 to 16 m, which are typical heights where one would find relay deployment in various scenarios. The transmitter antenna height (i.e. eNodeB antenna) is fixed to 25.5 m.

For each examined relay antenna height, 124 of path loss measurement locations are collected. For each measurement location several hundred readings are averaged in time domain so that the fast-fading component of the path loss is smoothed out. At each measurement point, GPS is used to determine the coordinates of the receiver. The location of the measurement points is presented in Fig. 2.

3 Empirical Model

In general, the measured path loss in [dB] at Relay Node (RN) that is at distance *d* from eNodeB, is calculated as:

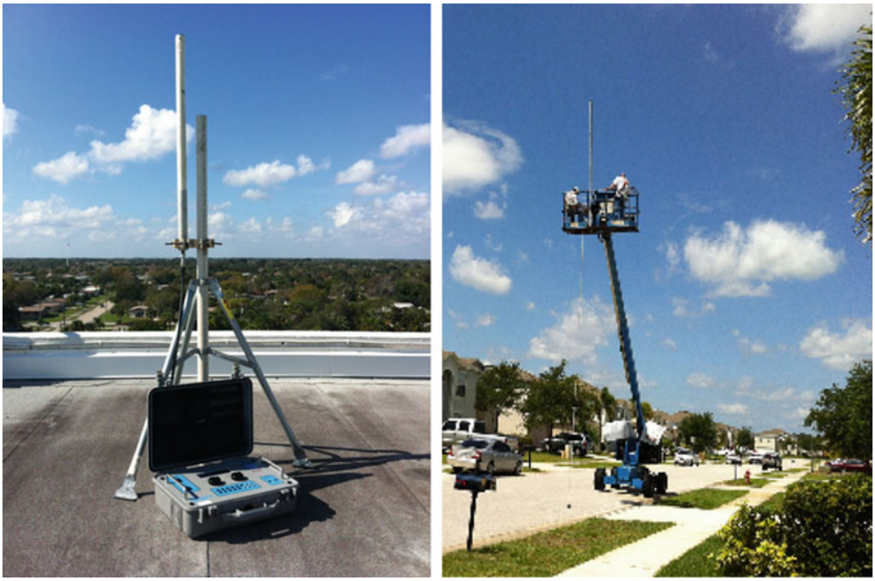


Fig. 1 Illustration of the transmitter (*left*) and the receiver (*right*)

Table 1 Parameters associated with the measurement campaign

Parameter	Value
Operating frequency	1,925 MHz
Transmit antenna height	25.5 m
Transmitting power	43 dBm
Transmit antenna gain	6 dBi
Cable and connector losses	0.7 dB
Receive antenna height	4, 8, 12, 16 m
Receive antenna gain including cable and connector losses	5 dBi
Noise figure	2 dB

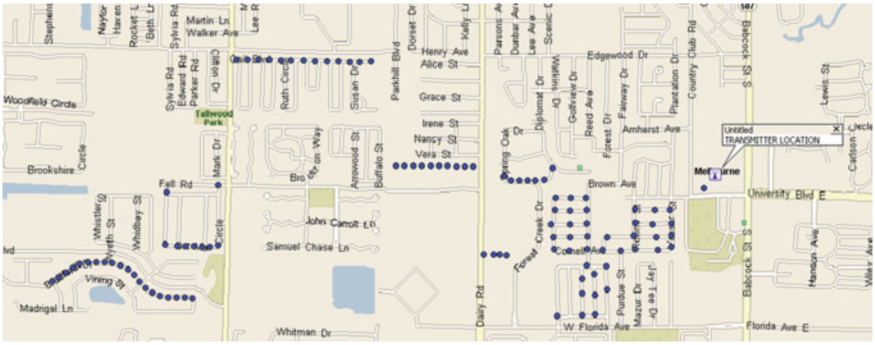


Fig. 2 Map of the area where the measurements were conducted. Transmitter location is Latitude: 28.064°N, Longitude: 80.624°W

$$PL_m(d) = ERP - RSL_m(d) \tag{1}$$

where *ERP* is the Effective Radiated Power in [dBm] and can be calculated as:

$$ERP = P_{TX} + G_{TX} - CL_{TX} \tag{2}$$

and RSL_m is the measured Received Signal Level in [dBm] and can be calculated as:

$$RSL_m(d) = P_r(d) + G_{RX} - CL_{RX} \tag{3}$$

where P_{TX} is the transmitted power in [dBm], G_{TX} is the transmit antenna gain in [dB], CL_{TX} is the cable losses of the transmitter in [dB], P_r is the received power in [dBm], G_{RX} is the receiving antenna gain in [dB] and CL_{RX} is the cable losses at reception side in [dB]. The value of all these parameters were given in Table 1 except P_r. The values for P_r are obtained from measurements.

3.1 Log-Distance Path Loss Model

In the first order approximation, the path loss in [dB] at any given distance *d* from the transmitter with respect to a reference distance d_0 may be estimated using the log-distance path loss model as:

$$PL_p = PL_0 + m\log(d/d_0) + X_\sigma \tag{4}$$

where PL_0 presents the intercept in [dB] and *m* is the slope of the model in [dB/decade]. X_σ is a log normally distributed random variable that describes the variability of the path loss due to shadowing effects. The parameters PL_0 and *m* are environmentally dependent parameters and are usually determined through statistical analysis of path loss measurements in a given environment.

3.2 Estimation of Model Parameters

The goal of the analysis is to develop an empirical propagation path loss model that explains the observed path loss data. MMSE (Minimum Mean Square Error) method was used to minimize the difference between prediction and measurements. Assuming that there are *N* measurements, the difference between measured and predicted path loss values for the *i*th measurement point is expressed as:

$$\delta_i = PL_{mi} - PL_{pi} \tag{5}$$

where δ_i is the prediction error for the ith point and $i = 1, 2, \ldots, N$. By substituting Eq. (4) into Eq. (5), one may write:

$$\delta_i = PL_{mi} - PL_0 - m\log(d_i/d_0) \tag{6}$$

Taking all measurement points (N) into account, Eq. (6) may be written in a matrix format as:

$$\begin{bmatrix} \delta_1 \\ \delta_2 \\ \vdots \\ \delta_N \end{bmatrix} = \begin{bmatrix} PL_{m1} \\ PL_{m2} \\ \vdots \\ PL_{mN} \end{bmatrix} - PL_0 \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} - m \begin{bmatrix} \log(d_1/d_0) \\ \log(d_2/d_0) \\ \vdots \\ \log(d_N/d_0) \end{bmatrix} \tag{7}$$

The objective here is to determine the optimum values of PL_0 and m that minimize the norm of the prediction error vector $\boldsymbol{\delta}$. In other words, the cost function which is given by:

$$J(\boldsymbol{\delta}) = \boldsymbol{\delta}^T\boldsymbol{\delta} = \sum_{i=1}^{N} \delta_i^2 \tag{8}$$

needs to be minimized.

Substituting Eq. (8) into Eq. (7), taking the partial derivatives with respect to PL_0 and m, and solving for the minimum yields:

$$\begin{bmatrix} C \\ D \end{bmatrix} = \begin{bmatrix} 1 & A \\ A & B \end{bmatrix} \begin{bmatrix} PL_0 \\ m \end{bmatrix} \tag{9}$$

where

$$A = \frac{1}{N}\sum_{i=1}^{N} \log(d_i/d_0), B = \frac{1}{N}\sum_{i=1}^{N} \log^2(d_i/d_0) \tag{10}$$

$$C = \frac{1}{N}\sum_{i=1}^{N} PL_{mi}, D = \frac{1}{N}\sum_{i=1}^{N} PL_{mi}\log(d_i/d_0) \tag{11}$$

Therefore, the optimum values of PL_0 and m can be given as:

$$PL_0 = \frac{CB - DA}{B - A^2}, \text{ and } m = \frac{D - AC}{B - A^2} \tag{12}$$

4 Analysis of Path Loss Measurements

The path loss measurements from which the antenna pattern effects have been taken out are presented in Fig. 3. Free space path loss is plotted as well and as it is seen, it represents a lower boundary for the measurements. One can easily observe that the measurements show some consistent trends. The increase of path loss is a linear function of the log of distance. The figure shows clearly the effect of the relay antenna height on the path loss value. The path loss decreases with the increase of the relay antenna height. Measurements show less dependency of path loss value on the relay antenna height when the receiver is close to the transmitter. This result may be explained by the fact that in such cases the receiver and the transmitter are in Line Of Sight (LOS) conditions in which relay antenna height does not have a significant impact on the received power. On the other hand, as the separation between the transmitter and the receiver becomes larger, this dependency is more pronounced especially for lower relay antenna heights. Table 2 summarizes obtained values of the slope m and intercept PL_0 for different relay antenna heights. As seen, the slope and intercept are functions of the relay antenna height. For the free space, $m = 20$. However, as the path between transmitter and receiver gets obstructed, m increases. In other words, m increases as the receive antenna height decreases. It is intuitive that the slope gets closer to the free space value as the relay antenna height is increased. In this work, the reference distance d_0 was defined as 100 m. Table 2 also shows the standard deviation (σ) of the error between the predicted and measured path loss values. In general, there are three parameters that describe the behavior and indicate the accuracy of prediction models. Firstly, the slope (m) of the model, which indicates how fast the path loss value increases as a function of distance. Secondly, the mean of prediction error (μ), which is the average difference between measured and predicted path loss value. Finally, the standard deviation (σ), which is a measure of the dispersion of the measured path loss from its local mean (μ).

Note that for the results in Table 2, the mean of prediction error (μ) is equal to zero. This is due to the fact that linear regression approach was used to develop those models and therefore the zero mean is obtained as a result of *curve fitting*. More meaningful statistical parameter that describes the accuracy of path loss models is the standard deviation (σ) of the prediction error.

Histograms in Fig. 4 present the distribution of the prediction errors about their means for different relay antenna heights. It is observed that errors are almost log normally distributed about a zero mean with a standard deviation that decreases with the increase of the relay antenna height. It is noteworthy to point out that the variation of the path loss around its mean is due to the shadowing effect. This variation becomes smaller as the path between the transmitter and the receiver gets clearer. When the relay antenna height increases, the received signal experiences less attenuation, reflection or diffraction that are caused by obstacles between transmitter and receiver. The standard deviation, σ, ranges from 6.34 dB for lowest relay height to 2.59 dB for the highest one. These values of σ are typical for suburban environment in which the measurements were conducted.

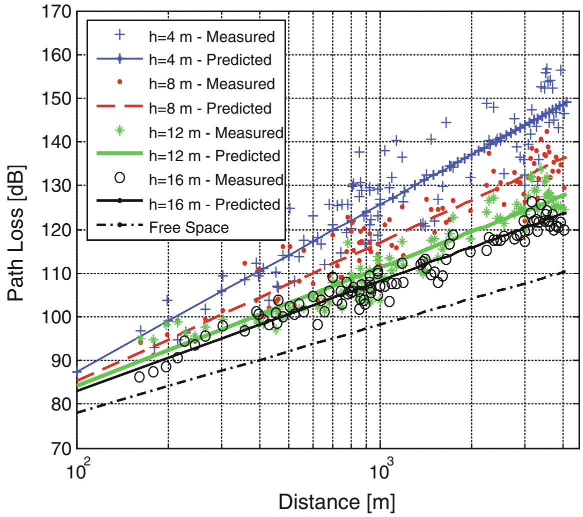


Fig. 3 Measured and predicted path loss for different relay heights

Table 2 Relay path loss propagation model parameters

Relay height (*h*) [m]	PL_0 [dB]	*m* [dB/decade]	σ [dB]
4	87.48	38.14	6.34
8	85.23	31.84	4.98
12	84.04	27.22	3.81
16	82.93	25.34	2.59
Free space	78.13	20	–

5 Path Loss Models for Relay Stations

According to the general path loss model given in Eq. (4) and propagation model parameters given in Table 2, one can write the propagation path loss model for any of the examined relay antenna heights. For example, in the case of 4 m relay, the model is given as:

$$PL(d) = 87.48 + 38.14 \log(d/d_0) \tag{13}$$

Similarly, other propagation models for the corresponding relay antenna heights 8, 12, and 16 m can be expressed as well. It is noteworthy to point out that these models are for the backhaul link only. Therefore, the considered models are valid for predicting path loss encountered in the relay link operates in 1,900 MHz band and distance ranges from 100 to 4,000 m.

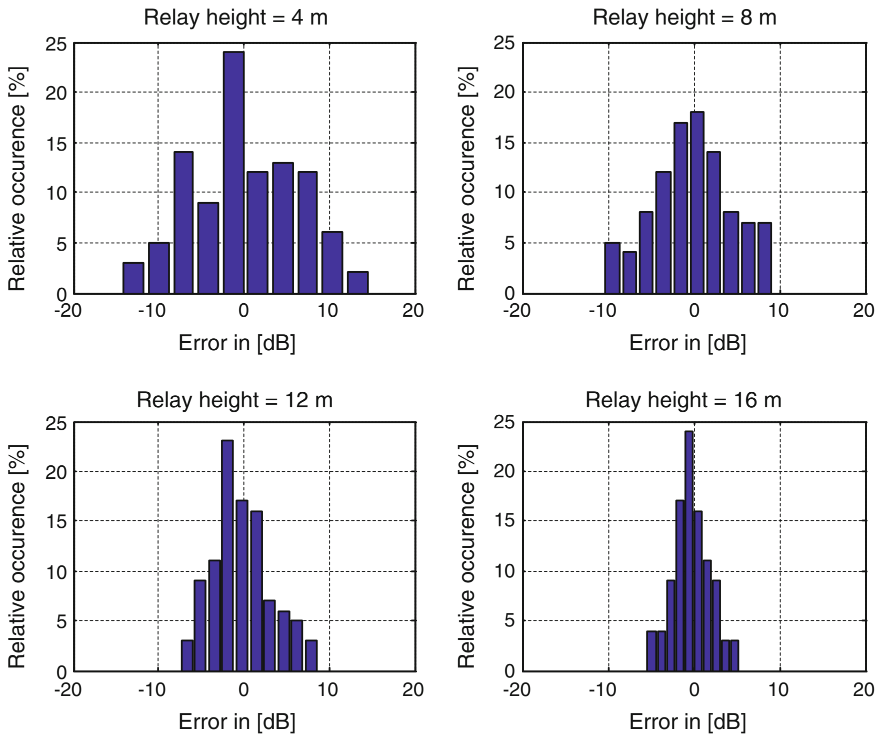


Fig. 4 Distribution of prediction error for the examined relay antenna heights

5.1 Path Loss Differences for the Examined Relay Heights

Table 3 shows the differences in path losses that were obtained for different relay antenna heights. For each pair of relay heights both mean and standard deviation of the difference is provided. The smallest average reduction of path loss of 3.21 dB is obtained when the relay antenna height is changed from 12 to 16 m. Similarly, an average of 18.46 dB path loss difference is observed when the relay height is raised from 4 to 16 m. The standard deviation of the path loss differences ranges from 3.59 to 8.27 dB.

5.2 Relay Antenna Height Correction Factor

The measured data suggest that there is a significant path loss difference between different relay heights. An empirical path loss propagation model that takes into account the relay antenna height in suburban environment can be given as:

Table 3 Average of path loss differences between relay heights

	Mean [dB]	Standard deviation [dB]
PL(h = 4)–PL(h = 8)	9.08	6.05
PL(h = 8)–PL(h = 12)	6.16	4.22
PL(h = 4)–PL(h = 12)	15.25	7.55
PL(h = 12)–PL(h = 16)	3.21	3.59
PL(h = 8)–PL(h = 16)	9.38	5.26
PL(h = 4)–PL(h = 16)	18.46	8.27

$$PL(d) = 87.48 + 38.14\log(d/d_0) - \Delta h \tag{14}$$

whereas the first two terms of Eq. (14) is the path loss model when $h = 4$ and Δh is the relay antenna height correction factor. In other words, Δh represents the reduction of the path loss as the result of the relay antenna height increases.

To the authors' knowledge, Δh for most of propagation models is formulated as a function either of only receiving antenna height h or of both operating frequency f and $h_.$ However, when f is fixed, as in the presented study, Δh becomes a function of h only. As a result, path loss difference between any two particular relay antenna heights remains constant for the entire distance range. However, the measurements in Fig. 3 show that Δh is not only a function of relay antenna height (h) but it is also a linear function of the log of distance (d). To understand and derive the relationship between Δh, h and d, following simple scenario is considered. According to the measurements (refer to Fig. 3), any two path loss models that have different relay antenna height h_1 and h_2, models might be given as:

$$PL_{1,2}(d) = PL_{01,02} + m_{1,2}\log(d/d_0) \tag{15}$$

where $h_1 > h_2$, $m_1 < m_2$, and $PL_{01} < PL_{02}$. The path loss difference ΔPL at any particular distance for two different antenna heights which is also equal to Δh can be given as:

$$\mathit{\Delta PL} = \mathit{\Delta h} = x\log(h_1/h_2) = \mathit{\Delta PL}_0 + \mathit{\Delta m}\log(d/d_0) \tag{16}$$

where $\Delta PL_0 = PL_{02} - PL_{01}$, $\Delta m = m_2 - m_{1.}$ and x is the slope of the antenna height dependence that needs to be determined. To generalize the expression given in Eq. (16), $h_2 = 4$ m and h_1 can be any value of the examined relay antenna heights (h). Therefore, the value of x may be calculated as:

$$x[\text{dB}] = \frac{\Delta m\log(d/d_0) + \Delta PL_0}{\log(h/4)} \tag{17}$$

Table 4 shows the values of ΔPL and Δm for different values of h. Based on Table 4 and Eq. (17), different expressions of x for corresponding h can be easily

Table 4 Slope and intercept differences

h [m]	PL_0 [dB]	ΔPL_0 [dB]	m [dB/dec]	Δm [dB/dec]
4	87.48	0	38.14	0
8	85.23	2.25	31.84	6.30
12	84.04	3.44	27.22	10.92
16	82.93	4.55	25.34	12.80

determined. To find the value of x that might be used for different relay heights, one may calculate the average across all values of x. Thus, the mean value of x is given by:

$$x[\mathrm{dB}] = 21.69\log(d/d_0) + 7.41 \tag{18}$$

By substituting Eq. (18) into Eq. (16), the antenna height correction factor is now can be given as:

$$\Delta h = [21.69\log(d/d_0) + 7.41]\log(h/4) \tag{19}$$

Δh is illustrated graphically in Fig. 5. This figure presents a family of curves of Δh as a function of distance (d) for different relay antenna heights (h). The figure clearly confirms that Δh is not only function of relay height (h) but also function of distance. Thus, Eq. (19) and Fig. 5 help explain how it is very important to take into account the distance dependency when considering the relay antenna height correction factor. This dependency has a great effect on the predicted path loss value and therefore on the model accuracy. Lack of this consideration might lead to very large errors in path loss predictions.

With the relay antenna correction factor, the goal of finding an empirical path loss model that takes into account the relay antenna height has been achieved. This model is given in Eq. (14) with its associated Δh given in Eq. (19). To verify the validity of the model after the implementation of Δh, a comparison between the model predictions and the measurements is provided in Table 5. The comparison is made in terms of the mean value of the prediction error (μ) and the standard deviation (σ). It is apparent from this table that implementation of the antenna height correction factor provides a very good agreement between the prediction and measured data. Since the path loss model for relay height equal to 4 m was taken as a reference for other models, there is no change in its μ and σ before and after implementation of antenna height correction factor.

Therefore, the model proposed in Eq. (14) along with the associated Δh given in Eq. (19) presents a general propagation model. This model might be used to predict the path loss value at any particular distance for any given relay station antenna height between 4 and 16 m. This is a suitable range of relay antenna height for environments that are similar to one surveyed in the measurement campaign.

Fig. 5 Relay antenna height correction factor (Δh) as a function of distance and relay height

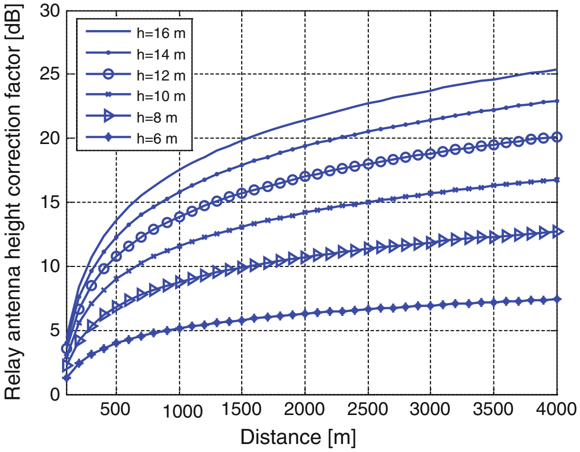


Table 5 Model comparison before and after implementation of Δh relative to the measurements

h [m]	Mean (μ) in [dB]		Standard deviation (σ) in [dB]	
	Before	After	Before	After
4	0	0	6.34	6.34
8	0	0.1	4.98	4.65
12	0	0.7	3.81	3.42
16	0	0.1	2.59	2.23

5.3 Dependence of Slope and Intercept on Antenna Height

After the determination of the relay antenna height correction factor, it is of a great interest to show how the slope (m) and the intercept (PL_0) of relay path loss propagation model change with the relay height. To derive the relationship between the model parameters (m and PL_0) and the relay height (h), Eq. (14) with the associated Δh given in Eq. (19) are to be reconsidered. By substituting Eq. (19) into Eq. (14) and rewriting the intercept and the slope as a function of h individually, one may obtain:

$$PL_0(h) = 87.48 - 7.41\log(h/4) \tag{20}$$

$$m(h) = (38.14 - 21.69\log(h/4))\log(d/d_0) \tag{21}$$

To express m as a function of h, log(d/d_0) in Eq. (21) needs to be equal to one. Therefore,

$$m(h) = 38.14 - 21.69\log(h/4) \tag{22}$$

Fig. 6 Intercept (*up*) and slope (*down*) as a function of relay antenna height

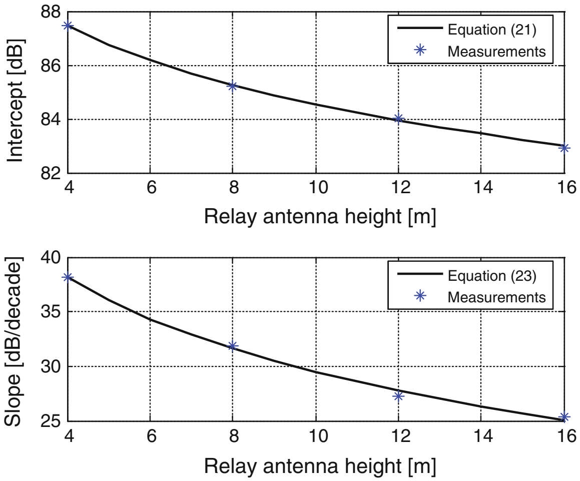


Figure 6 compares PL_0 and m obtained from the relationships (Eq. 20) and (Eq. 22) to the ones achieved by the measurements presented in Table 2. The figure shows that both intercept path loss and slope decrease with the increase of relay antenna height. Additionally, the figure indicates that for every doubling in relay antenna height there is a decrease of 2.23 dB in intercept and 6.53 dB/decade in slope. Formulas (Eq. 20) and (Eq. 22) represent mathematical expressions by which one can calculate the intercept (PL_0) and the slope (m) for any given relay antenna height ranging from 2 to 16 m. As a result, another approach of developing empirical propagation path loss models for relay environment has been achieved. This model might be given as:

$$PL(d) = PL_0(h) + m(h)\log(d/d_0) \tag{23}$$

where $PL_0(h)$ and $m(h)$ can be calculated by Eq. (20) and Eq. (22), respectively.

References

1. R. Pabst, B.H. Walke, D.C. Schultz, P. Herhold, H. Yanikomeroglu, S. Mukherjee, H. Viswanathan, M. Lott, W. Zirwas, M. Dohler, H. Aghvami, D.D. Falconer, G.P. Fettweis, Relay-based deployment concepts for wireless and mobile broadband radio. IEEE Commun. Mag. **42**(9), 80–89 (2004)
2. D. Soldani, S. Dixit, Wireless relays for broadband access [radio communications series]. IEEE Commun. Mag. **46**, 58–66 (2008)
3. 3GPP Technical Report 36.814 V9.0.0: Further advancements for E-UTRA physical layer aspects (Release 9), (2010), http://www.qtc.jp/3GPP/Specs/36814-900.pdf
4. F. Akyildiz, D.M. Gutierrez-Estavez, E.C. Reyes, The evolution to 4G cellular systems: LTE-Advanced. Physical Communications **3**(4), 217–244 (2010)

5. P. Kyösti, J. Meinilä, L. Hentilä, X. Zhao, T. Jämsä, M. Narandzić, M. Milojević, C. Schneider, A. Hong, J. Ylitalo, V. Holappa, M. Alatossava, R. Bultitude, Y. Jong, T. Rautiainen, IST-4-027756 WINNER II, D1.1.2 V1.0, WINNER II channel models, part II. Radio channel measurement and analysis results, Sept 2007
6. G. Senarath, W. Tong, P. Zhu, H. Zhang, D. Steer, D. Yu, M. Naden, D.K. Nortel, Multi-hop relay system evaluation methodology (channel model and performance metric), IEEE 802.16j-06/013r3, Feb 2007
7. C.Q. Hien, J.-M. Conrat, J.-C. Cousin, in *European Wireless Technology Conference (EuWIT).* On the impact of receive antenna height in a LTE-Advanced relaying scenario, pp. 129–132 (2010)
8. C.Q. Hien, J.-M. Conrat, J.-C. Cousin, Propagation path loss models for LTE-advanced urban relaying systems, IEEE International Symposium on Antennas and Propagation (APSURSI), pp. 2797–2800 (2011)
9. M. Webb, G. Watkins, C. Williams, T. Harrold, R. Feng, M. Beach, *Mobile Multihop: Measurements vs Models* (European Cooperation in the Field of Scientific and Technical Research, Europe, 2007)
10. M. Hamid, I. Kostanic, in *Proceedings of The World Congress on Engineering and Computer Science 2013, WCECS 2013*. Path loss measurements for relay stations in 1900 MHz band, Lecture Notes in Engineering and Computer Science, 23–25 October. (San Francisco, USA, 2013), pp. 736–741

Chapter 32
System Integration of a Home Network System with a Narrowband Power Line Communication Method and a Voice Recognition Command Controller

Lee Kyung Mog

Abstract After a home network system, consisting of some electric devices, had been constructed by a Narrowband power line communication method, the system was integrated with an external voice recognizing control board. Experimental data was measured for the communication characteristics between a central PC and a central-side PLC modem in Local loopback test, and then between the PC and a receive-side modem in Remote loopback test. The voice control board was connected to the PC via Bluetooth communication, and used to manipulate the home electric devices' setup statuses. A communication protocol was designed for the system, and by changing its length, the system's response time was measured in the local loopback test and the remote test. The response time of the system in the remote loopback test was about 345 ms for the protocol of the length of 13 Bytes. So, the proper protocol length for this system was decided to be 13 Bytes long. The command-recognizing time of the controller, measured by the Polling Process Technique, showed that it depended on the number of commands. For eight commands, the recognizing time of the voice board was about 100 ms. The Bluetooth and the power line communication data rates were 9,600 and 2,400 bps. Home electric devices' statuses were shown graphically on the computer screen according to the device's setup conditions, which were controlled by some voice commands.

Keywords Auto-home network · Intelligent home · Narrowband PLC communication · Remote voice controller · Bluetooth communication · Voice command smart home

L. K. Mog (✉)
E-Commerce Department, Semyung University, Jecheon, Chungbuk, South Korea
e-mail: ahkml@semyung.ac.kr

H. K. Kim et al. (eds.), *Transactions on Engineering Technologies*,
DOI: 10.1007/978-94-017-9115-1_32,

1 Introduction

Recently, speedy communication infrastructure has been built and Giga bit internet has been widely spread into our homes. Also, auto home network technology, such as the Smart Home, has been widely researched [1–3].

Home automation techniques have two kinds of communication methods, wireless or wire. In wire communications, there are ethernet, IEEE1397, USB, and Power Line Communications (PLC) [4, 5], etc. In wireless communication, there are wireless LAN in IEEE802.11x [6], Bluetooth [7], and Zigbee [8, 9], etc.

For the home networking, the existing power line can be used without any additional effort, just simply plug the communication modem into it. So, the PLC method preferred because an extra wire is not required. The PLC method is easy to use, safe from electric wave interference, and still works well.

In the power line communication methods, there are two types of PLC, a Narrowband and a Broadband. The Broadband PLC is used in the speedy multimedia communication, the Narrowband one for the home electric device control [10]. As long as both of the PLC technologies follow the standards of the first layer, the so called physical layer, of the OSI-7, the communication method is no longer a concern.

And, a research of an external voice recognition control board, which was used as a user interface to send user's commands to the PC and as a controller of a space war shooting computer game, was published [11, 12].

A home network system with an external voice recognition control board in manipulating remote home electric devices was published [13].

In this paper, the home network system, including a central control computer (PC) and integrated with an external voice recognition control board, was constructed with the Narrowband PLC method to manipulate remote home electric devices. A communication protocol was designed for this system and the system response time was measured in the remote loopback test. The PLC's communication data rate was 2,400 bps and the Polling control technology [14] was used in reading and setting up the remote electric devices' status. To get a voice commands from the external board, the Polling Method was used. Some home electric devices, such as a humidifier, a fan, a lamp, and a door lock, were controlled by voice commands.

2 Body of Paper

Figure 1 shows a home network system with the Narrowband PLC that was constructed in this paper. Here, four home electric devices, which were a humidifier, a fan, a light and a door lock, were plugged into the home power line. The central PC was controlling and monitoring of all the system.

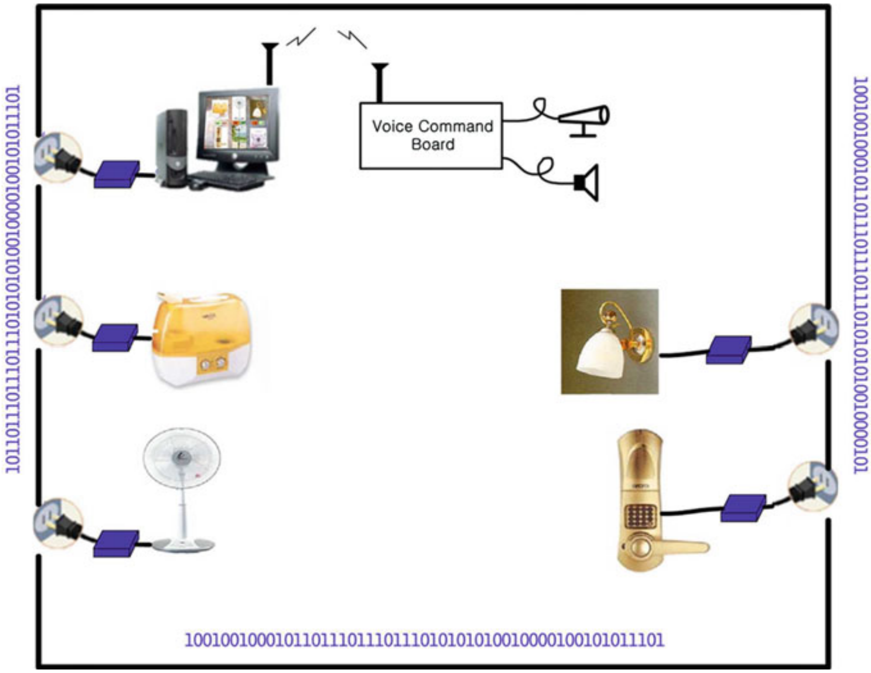


Fig. 1 Constructed Home network system with the Narrowband PLC

Figure 2 shows the detailed system construction of the central control PC and an external voice recognition control board. Figure 2a shows an external voice recognition control board. Figure 2b shows the central control PC connected to the power line through a central side PLC modem.

When a voice command was recognized on the external voice recognition control board, the code corresponding to a recognized command was transmitted to the central PC via Bluetooth module. Then, the PC sent the home electric device control signal to the central-side PLC modem through a RS-232C serial cable. Then the modem broadcasted an information protocol into the home power line.

The central-side PLC modem was able to communicate to the receive-side modems of the home electric devices through the power line. The receive-side modem sent the received information to the microprocessor of the AT89C2051 in serial. The distance between the central-side and the receive-side modem was changed. The communication environment was clear and quiet without noise generating devices such as motors.

The receive-side microprocessor of the AT89C2051 was configured for four of the eight ports that functioned as outputs of the device control and the other four as inputs of the device's ON/OFF status.

All the PLC modems were using the ST7537HS1 chips of SGS-THOMSON's company. The modulation rate was in half duplex asynchronous 2400 FSK.

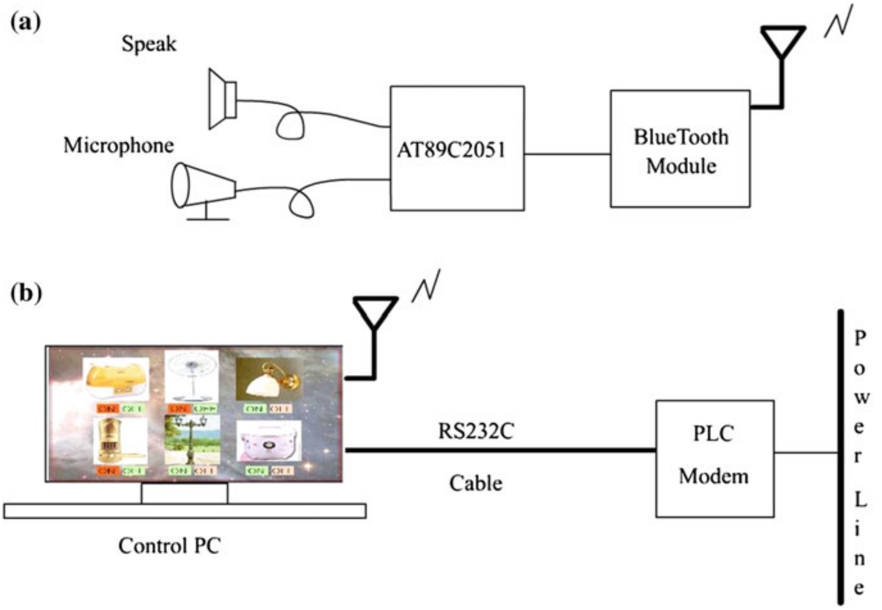


Fig. 2 Detailed system construction of a central control PC and an external voice recognition control board. **a** An external voice recognition control board. **b** A central control PC connected to the power line

The central-side PLC modem was connected to the COM1 port of the PC with RS-232C serial cable. The receive-side PLC modem was connected to the AT89C2051 microprocessor manufactured by the ATMEL company.

Figure 3 shows the diagram of the program blocks of the constructed system. Here (1) is the point of the Local Loopback test, and (2) the spot of the Remote Loopback test. When the program of the central PC started, it checked the ON/OFF status of the home electric devices and showed each device's status graphically on the screen. Then it checked the voice command from the external voice recognition control board. When there was a voice command, it sent a device control signal to the PLC modem of the PC. Then the modem sent the protocol containing the signal to all the receive-side PLC modems. The receive-side PLC modem retrieved the signal from the protocol after performing the error check and sent the signal to the AT89C2051's communication program. Then the microprocessor output the setup signal to the chosen device and vice versa. The microprocessor responded with an "ACK" protocol, containing the statuses of the devices to the central PC.

The PLC modems were communicating in the Polling process, so there were two kinds of stations, a primary one and a secondary one. The central-side PLC modem acted as the primary station and the receive-side one as the secondary. Also, the external voice control board was treated as another secondary station.

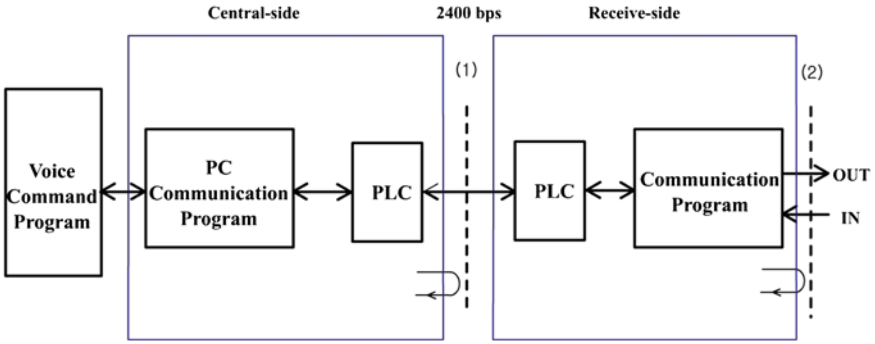


Fig. 3 Diagram of the program blocks of the constructed system

In the Stop and Wait process, the primary modem broadcasted the “Ready to Receive” protocol and waited for the “ACK” (Acknowledgement) protocol from the secondary modems, which could contain responding data.

The ACK protocol might contain data of the voice command code or of the statuses of the home electric devices.

The PC program checked the ACK protocol from external voice control board for a voice command. When there was a voice command, it sent the “Ready to Receive” protocol containing the device control signal corresponding to the command. Then the program checked the ACK protocol from the receive-side modem for the device statuses and changed the colors of the device pictures on the screen corresponding to the statuses.

Figure 4 shows the structure of the protocol for the PLC communication. The protocol consisted of characters, each one consists of 8bits. It begins with “STF” characters which mean the start of the protocol, following is “RTR:” characters as “Ready to Receive” signal, ADDR characters as the receive-side PLC address, “DATA” as the output of the Port 0 of the ATMEL microprocessor, and “Check Sum” for error check operation. In this experiment, total length of the protocol was designed to be 13 bytes long. The “Ready to Receive” protocol was as follows:

“Ready to Receive” protocol:
“STF” +“:” + “RTR” + “:” + “ADDR” + “:” + “DATA” + “:” + “CS”

Here, “STF”: 3Bytes as a frame start, “RTR”: 3 Bytes as the Ready-to-receive signal, “:”: 1 Byte as a data separator, “ADDR”: 1 Byte as the receiver PLC address, “DATA”: 1 Byte as the output data of the port 0, “CS”: 1 Byte for the checksum. The responding ACK protocol had the same structure as the “Ready to receive”, except using “ACK” instead of “RTR”.

And, the “ACK” protocol was as follows:
“STF” +“:” + “ACK” + “:” + “ADDR” + “:” + “DATA” + “:” + “CS”

The “DATA” part consists of two hex characters. The first one is for the output control signal and the second one for the input device. The microprocessor’s lower

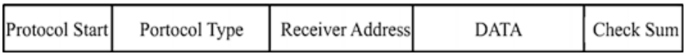


Fig. 4 The structure of a PLC communication protocol

Table 1 The response time from the receive-side device and the writing time of data on the PC's COM1 port

Number of bytes	1 Byte	2 Byte	6 Byte	11 Byte	13 Byte	15 Byte
Response time (ms)	156	172	234	312	345	375
Writing time (ms)	16	32	94	172	200	235

four ports (P0.0, P0.1, P0.2, P0.3) were used as inputs of the devices' statuses and the higher four ports (P0.4, P0.5, P0.6, P0.7) as outputs of the electric device control.

Table 1 shows the response time from the receive-side device in the Remote Loopback test and the writing time of data on the PC's COM1 port in the Local Loopback test. The writing time was about 16 ms for the data of 1 Byte. The response time of this polling process between the PC and the receive-side modem was measured to be about 156 ms for the data of 1 Byte. The above measured date showed that the longer the length of the protocol, the more delayed the writing time. But, the delay time, which was the response time minus the writing one, in the remote loopback test was always the same as 140 ms regardless of the length of the protocol. For this system integration, the best length of the protocol was chosen to be 13 bytes long for which the response time was 345 ms.

Figure 5 shows the polling process of the system built. To get a voice command, the controller's program sent a poll signal to the voice control board. When a voice command was detected, it sent the corresponding electric device control signal to the receive-side microprocessor to setup the device. To get the electric devices' statuses, it awaited the responding protocol from the receive-side microprocessor. Then it changed the color of the device picture on the screen according to each electric device's status. The connections of the receive-side AT89C2051 ports were as follows:

Receive AT89C2051	Home electric devices
P0.4	Humidifier
P0.5	Fan
P0.6	Lamp
P0.7	Door lock

Figure 6 shows the flow charter of the control program of the central command. When a voice command was detected, it generated an RTR protocol to send the device control signal to the receive-side microprocessor. After the microprocessor sent the devices' statuses to the program with the responding "ACK" protocol, the program changed the color of the devices' pictures on the screen. This process was repeated.

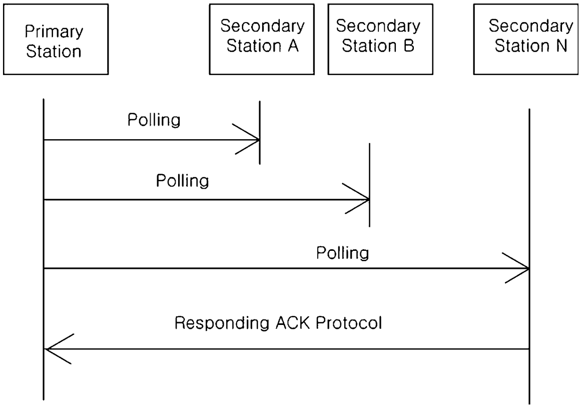


Fig. 5 A The polling process of the system built

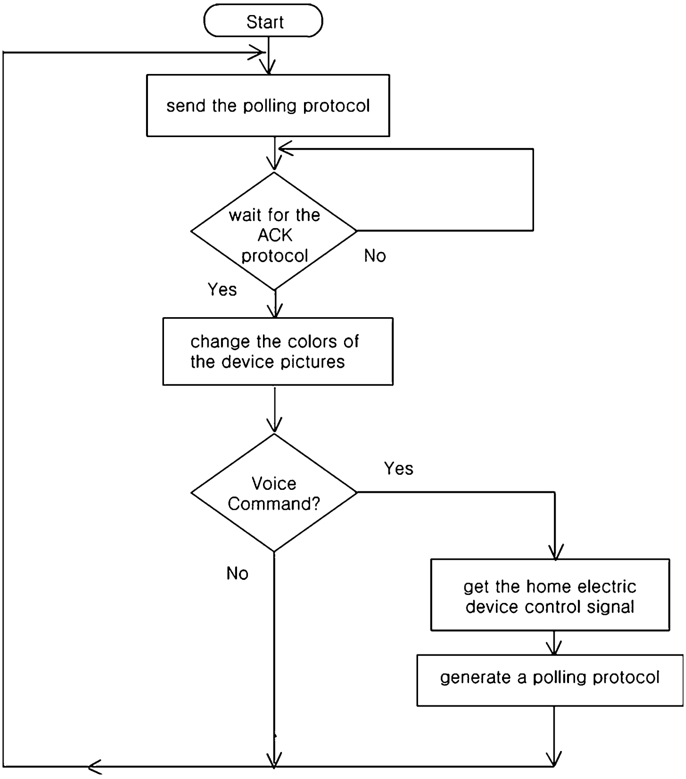
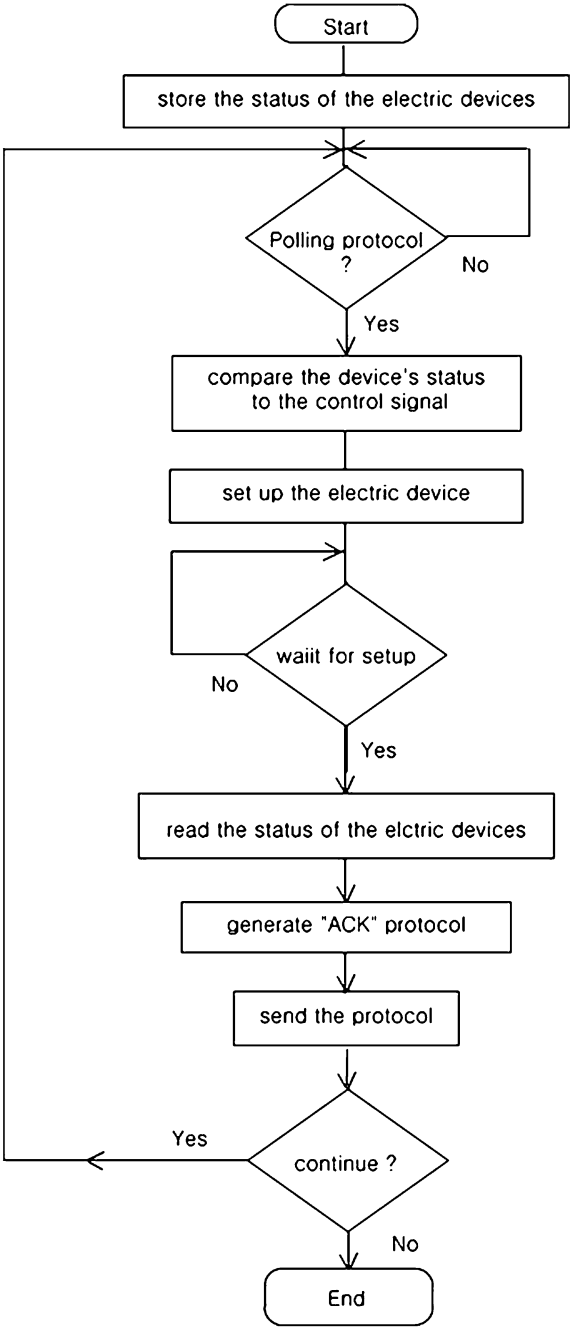

Start
send the polling protocol
wait for the ACK protocol
No
Yes
change the colors of the device pictures
Voice Command?
Yes
No
get the home electric device control signal
generate a polling protocol

Fig. 6 Flow charter of the control program of the central PC

Fig. 7 Flow charter of the receive-side microprocessor program


Start
store the status of the electric devices
Polling protocol ?
No
Yes
compare the device's status to the control signal
set up the electric device
waiit for setup
No
Yes
read the status of the elctric devices
generate "ACK" protocol
send the protocol
Yes
continue ?
No
End

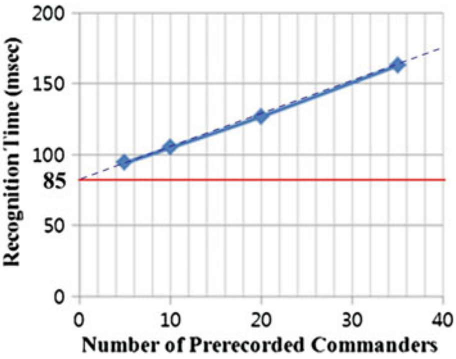


Fig. 8 Measured recognition time verse the number of pre-recorded voice commands

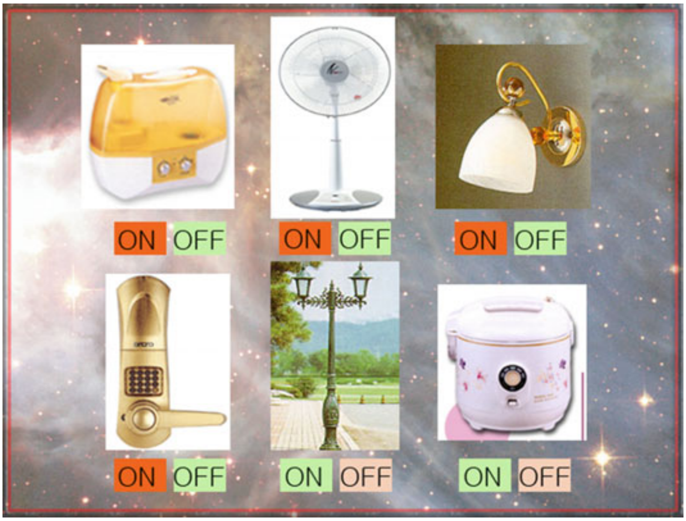


Fig. 9 The device setup status shown graphically on the screen

Figure 7 shows the flow charter of the receive-side microprocessor program. When started, the program checked and stored all the electric devices' statuses. When a polling signal was detected, it extracted the device control signal from the polling protocol. Then it compared them with the pre-stored devices' statuses. When there was a difference, the program set up the electric device according to the control signal. After checking the devices status again, the program sent the status to the control program in a responding ACK protocol. This process was repeated.

Figure 8 shows the measured recognition time verse the number of pre-recorded voice commands. For eight commands, the recognizing and processing time was about 100 ms. The data showed that the recognition time was increased about 2.5 ms for every newly added command.

Figure 9 shows the device setup statuses graphically on the screen according the command of the voice controller.

3 Conclusion

In this experiment, a home network system was integrated with a voice recognizing control board. The home system was constructed with the Narrowband Power Line Communication technology. The board was used in controlling the home electric devices by the user's voice command.

After a protocol for the communication was designed, the response time was measured by changing the length of the protocol in a remote loopback test. Then the best protocol for this system was decided to be 13 bytes long. With the polling technique, the PC's control program checked the voice control board for a voice command and set up the electric devices according to the command. The communication data rate between the central-side modem and the receive-side one was in 2,400 bps. The voice control board was connected to the PC via Bluetooth communication in 9,600 bps data rate.

The voice recognizing time of the voice controller was measured by changing the number of commands because it depended on the number. For eight commands, the recognizing time of the voice board was about 100 ms.

Finally, the devices' setup conditions were shown graphically on the PC screen according to the home electric devices' statuses.

Although this experiment was performed in a quiet environment condition, further research is required for the communication method in a noisy environment.

Acknowledgment This paper was supported by the Semyung University Research Grant of 2012.

References

1. R. Natarajan, A.P. Mathur, Development of an infrastructure for the management of smart homes, in *Computer Science Technical Reports. 2002,* Paper 1540
2. C. Jin, T. Kunz, Smart home networking: lessons from combining wireless and powerline networking. Smart Grid Renewable Energy **2**, 136–151 (2011)
3. L. Zoref, D. Bregman, D. Dori, Networking mobile devices and computers in an intelligent home, Int. J. Smart Home **3**(4) 15–22 (2009)
4. V. Oksman, J. Zhang, G.HNEM: the new ITU-T standard on narrowband PLC technology. IEEE Commun. Mag. **49**(12), 36–44 (2011)
5. K.M. Lee, The construction of a remote game control system by the power line communication. Korea Game Soc. **7**(1), 53–58 (2007)

6. IEEE Standards Associacion, IEEE Std. 802. 11-2012—Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications, *IEEE* (2012)
7. D.A. Gratton, *Bluetooth Profiles* (Prentice Hall, Uppper Saddle River, 2003)
8. S.-J. Woo1, B.-D. Shin, Efficient cluster organization method of zigbee nodes, Int. J. Smart Home **7**(3), 45–56 (2013)
9. N. Dou, Y. Mei, Z. Yanjuan, Z. Yan, The networking technology within smart home system-ZigBee technology. Comput. Sci.-Technol. Appl. **2**, 29–33 (2009)
10. H.C. Ferreira, L. Lampe, J. Newbury, T.G. Swart, *Power line communications: theory and applications for narrowband and broadband communications over power lines* (Wiley, New York, 2010)
11. K.M. Lee, Implementation of a computer game voice command board with a speaker-dependent recognition chip. J. Convergence Inf. Technol. **8**(14), 238–244 (2013)
12. K.M. Lee, Voice-game controller via bluetooth communication with a speaker-dependent recognition chip, in *Research Notes in Information Science(RNIS): Advance Institute of Convergence Information Technology*, vol. 14, pp. 61–66 (2013)
13. K.M. Lee, Construction of a home network system using a power line communication method and a voice recognition command control, in *Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering 2013*, WCE 2013, pp. 770–773, San Francisco, 23–25 Oct 2013
14. B.A. Forouzan, *Introduction to Data Communications and Networking* (International Editions, McGrawHill, 1998)

Chapter 33
A New Non-stationary Channel Model Based on Drifted Brownian Random Paths

Alireza Borhani and Matthias Pätzold

Abstract This paper utilizes Brownian motion (BM) processes with drift to model mobile radio channels under non-stationary conditions. It is assumed that the mobile station (MS) starts moving in a semi-random way, but subject to follow a given direction. This moving scenario is modelled by a BM process with drift (BMD). The starting point of the movement is a fixed point in the two-dimensional (2D) propagation area, while its destination is a random point along a predetermined drift. To model the propagation area, we propose a non-centred one-ring scattering model in which the local scatterers are uniformly distributed on a ring that is *not* necessarily centred on the MS. The semi-random movement of the MS results in local angles-of-arrival (AOAs) and local angles-of-motion (AOMs), which are stochastic processes instead of random variables. We present the first-order density of the AOA and AOM processes in closed form. Subsequently, the local power spectral density (PSD) and autocorrelation function (ACF) of the complex channel gain are provided. The analytical results are simulated, illustrated, and physically explained. It turns out that the targeted Brownian path model results in a statistically non-stationary channel model. The interdisciplinary idea of the paper opens a new perspective on the modelling of non-stationary channels under realistic propagation conditions.

Keywords Brownian motion processes · Channel modelling · Local autocorrelation function · Local power spectral density · Non-centred one-ring scattering model · Non-stationary channels · Targeted motions

A. Borhani (✉) · M. Pätzold
Faculty of Engineering and Science, University of Agder, 4898 Grimstad, Norway
e-mail: alireza.borhani@uia.no

M. Pätzold
e-mail: matthias.paetzold@uia.no

H. K. Kim et al. (eds.), *Transactions on Engineering Technologies*,
DOI: 10.1007/978-94-017-9115-1_33,

1 Introduction

To develop mobile communication systems, geometric channel models are recognized as one of the most effective candidates, which allow a fairly accurate system performance analysis. As an example, the one-ring scattering model [1–3], in which the local scatterers are uniformly distributed on a ring centered on the MS, is an appropriate model capturing the propagation effects in rural and sub-urban propagation areas. The unified disk scattering model (UDSM) [4] is also one of the most general geometric channel models, which covers numerous circularly-symmetric scattering models as special cases, including the one-ring model. In this regard, an overview of the most important geometric channel models can be found in [5].

Geometric channel models often profit from a common simplification, namely the stationarity assumption of the stochastic channel in time. Considering a very short observation time instant justifies a time-invariant AOA at the MS, which then results in a statistically stationary channel model. Many empirical and analytical investigations, e.g., [6–8], however, show that this property is only valid for *very* short travelling distances [9]. This calls for the need to develop and analyze stochastic channel models under non-stationary conditions.

Despite the drastic number of investigations on stationary geometric channel models, the literature lacks studies on non-stationary geometric channel models. Only a small number of analytical studies, e.g., [10–13], cope with the statistical properties of non-stationary channels. To the best knowledge of the authors, except the non-stationary one-ring scattering model studied in [14], none of the established geometric scattering models listed in [5] has been analyzed under non-stationary conditions. In [14], a non-stationary one-ring channel model has been derived by assuming that the MS moves from the center of the ring to the ring's border on a straight line. In this paper, we further expand the idea of [14] by allowing the MS to randomly fluctuate around a straight line, where its starting point is not necessarily the ring's center. It can be any point inside the ring of scatterers. To this end, we let the MS move in a semi-random way, but subject to follow a given preferred direction. By establishing an analogy between such a motion and the chaotic movement of particles suspended in fluids discovered by Robert Brown (see [15]), we model the travelling path of the MS by a BMD. We coin the term *targeted Brownian path model* to address the proposed path model of the MS. For a given BM process, the randomness of the path can be controlled by a single parameter. By eliminating the randomness of the path, the MS arrives at a fixed destination point via a straight path. Accordingly, the path model of [14] can be obtained as a special case of the proposed targeted Brownian path model.

Moving along a targeted Brownian path results in local AOAs and local AOMs, which are modelled by stochastic processes rather than random variables. We present the first-order density of the AOA and AOM processes in closed form. Expressions for the local PSD of the Doppler frequencies and ACF of the complex channel gain are also provided. These expressions are different from those presented in [16]. Numerical computations at 2.1 GHz illustrate the analytical results and verify the non-stationarity of the channel model. It is shown that non-stationarity in time contradicts the common isotropic propagation assumption on the channel. It is also proved that the one-ring scattering model can be obtained as a special case of the proposed channel model of this paper.

It is worth mentioning that 3D BM processes have been used to model fully random motions of mobile users [17]. However, 1D BM processes with drift have never been used to model semi-random motions of mobile users. Several other mobility models have also been employed in mobile ad hoc networks [18], but not in the area of channel modelling. In a nutshell, the novelty of this paper arises from the pioneering utilization of the BMD process as a path model for the modelling of non-stationary mobile fading channels.

The remainder of this paper is organized as follows. Section 2 gives a brief introduction to BM processes as a physical phenomenon, while Sect. 3 utilizes the BM process for developing the targeted Brownian path model. Section 4 describes the propagation scenario by means of the non-centred one-ring scattering model. The complex channel gain of the proposed channel model is then described in Sect. 5. Section 6 investigates the statistical properties of the channel model. Numerical results are provided in Sect. 7. Finally, Sect. 8 summarizes our main findings and draws the conclusions.

2 Principles of Brownian Motion Processes

BM was originally discovered in 1827 by the famous botanist, Robert Brown. It describes the chaotic movement of particles suspended in a fluid or gas [15]. In the 1860s, there were experimentalists who clearly recognized that the motion is due to the impact of suspending molecules. Finally, in 1906, Albert Einstein [19] offered an exact physical explanation of such a motion based on the bombardment of the suspended particles by the atoms of the suspending fluid. In 1908, a mathematical explanation of the BM was provided by Langevin [20].[1] BM processes have a wide range of applications, such as modelling of stock market fluctuations, medical imaging, and fractal theory [22]. In mobile ad hoc networks, 2D BM processes (random walk) are also employed to model irregular motions of mobile nodes [18]. The model is then used for network layer analysis.

A stochastic process $\{B(t) : t \in [0, T]\}$ is said to be a standard BM process if:

1. $B(0) = 0$.
2. $\forall 0 \leq s < t \leq T$, the random variable given by the increment $B(t) - B(s)$ follows the Gaussian distribution with zero mean and variance $t - s$, i.e., $B(t) - B(s) \sim N(0, t - s)$.
3. $\forall 0 \leq s < t < u < v \leq T$, the increments $B(t) - B(s)$ and $B(v) - B(u)$ are statistically independent.

[1] A translation of [20] into English has been provided in [21].

From the conditions above, it can be concluded that $B(t)$ is a Wiener process with normally and independently distributed increments.

3 Path Modelling

In what follows, we first provide an equivalent spatial representation of the temporal BM process. Subsequently, the proposed local BM process is used to model the targeted motion of the MS along a predetermined drift.

3.1 Spatial Representation of BM Processes

To establish an analogy between the BM process and the MS movement, let us first assume that the MS starts from a given point with Cartesian coordinates (x_s, y_s) in the 2D plane. The aim is to model the random path starting from (x_s, y_s) via the BM process described in Sect. 2. For this purpose, we establish a mapping from the temporal representation of the BM process $B(t)$ to the spatial representation of the BM process $B(x)$ by replacing the temporal variable t by the spatial variable x. Accordingly, the first condition of the BM process, i.e., $B(0) = 0$, changes to $B(x_s) = 0$. By assuming (x_d, y_d) as the terminal point of the movement, we introduce the scalar standard BM process over the range $[x_s, x_d]$ by means of the spatial stochastic process $B(x)$, which satisfies the following three conditions:

1. $B(x_s) = 0$.
2. $\forall x_s \leq x_p < x \leq x_d$, the random variable given by the increment $B(x) - B(x_p)$ follows the Gaussian distribution with zero mean and variance $x - x_p$, i.e., $B(x) - B(x_p) \sim N(0, x - x_p)$.
3. $\forall x_s \leq x_p < x < x_q < x_m \leq x_d$, the increments $B(x) - B(x_p)$ and $B(x_m) - B(x_q)$ are statistically independent.

For computational reasons, it is useful to consider the BM process at discrete values of x. To this end, we define $\Delta x = (x_d - x_s)/L$ for some positive integer L. Hence, $B_l = B(x_l)$ denotes the BM process at $x_l = x_s + l\Delta x$ $(l = 0, 1, \ldots, L)$. Now, with reference to Conditions 2 and 3, it can be concluded that $B_l = B_{l-1} + \Delta B_l$, where each ΔB_l is an independent normal distributed random variable of the form $N(0, \Delta x)$.

3.2 The Targeted Brownian Path Model

To model the targeted motion of the MS in the 2D plane, we propose a path with a controllable drift in a preferred direction, while the fluctuations of the path are modeled by the spatial BM process B_l. Accordingly, the path $\mathscr{P}$ of the MS is modelled as follows

$$\mathscr{P}: \left\{ (x_l, y_l) \middle| \begin{array}{c} x_l = x_s + l\Delta x, \\ y_l = ax_l + b + \sigma_y B_l, \end{array} \right\} \tag{1}$$

where $l = 0, 1, \ldots, L$ is the position index, the variable a denotes the slope of the drift, b is a constant shift along the y-axis, and σ_y allows to control the randomness of the path. Considering the fact that the randomness of the path $\mathscr{P}$ originates inherently from the randomness of the BM process B_l, the parameter σ_y provides an additional degree of freedom to control the randomness. For instance, by setting σ_y to 0, any point on the line represented by $y_l = ax_l + b$ can be reached. Whereas, increasing the value of σ_y reduces the chance of arriving at that point. However, the mean direction of the path remains unchanged. It is also noteworthy that the path model in (1) reduces to that in [14] if $\sigma_y = 0$. The model also enables to incorporate random fluctuations only along a specific line. For instance, by increasing a towards infinity, the fluctuations occur only along the y-axis. The same behaviour can be attained along any other line (axis) if we simply rotate the coordinate system.

In mobile communications, the proposed targeted Brownian path can be a very useful model to describe typical dynamics of users in motion, such as persons walking along a street, but not necessarily along a very smooth path. In vehicular communications, the model can also be used to explain the jittery motion of the vehicle antenna, while the vehicle is moving along a given direction.

4 The Propagation Scenario

To cope with the scattering effect caused by the propagation area, we propose a non-centred one-ring scattering model, in which the local scatterers are uniformly distributed on a ring that is not necessarily centred on the MS. The displacement of the MS from the ring's center results in a non-isotropic channel model. This model is an appropriate geometric scattering model to explain environments, in which the base station (BS) antenna is highly elevated to scattering-free levels, while the MS antenna is surrounded by a large number of local scatterers. This situation occurs mostly in rural and sub-urban areas.

Figure 1 shows the proposed non-centred one-ring scattering model with the uniform distribution of the local scatterers S_n $(n = 1, 2, \ldots, N)$ on a ring of radius R centered on the origin. In this regard, α_n^S denotes the angle-of-scatterer (AOS) associated with the nth scatterer. At a reference point in time t_0, the MS starts its movement from (x_0, y_0) and tracks the path $\mathscr{P}$ to reach (x_L, y_L) at time t_L. The position of the MS at time $t_l \in [t_0, t_L]$ is described by Cartesian coordinates (x_l, y_l). It is also assumed that the MS is moving with a constant velocity v_R in the direction indicated by the AOM $\alpha_v(l)$. Owing to high path loss, we assume that at time t_l, a wave emitted from the BS reaches the MS at the AOA $\mathbf{a}_n^R(l)$ after a single bounce by the nth randomly distributed scatterer S_n located on the ring. A realization of the

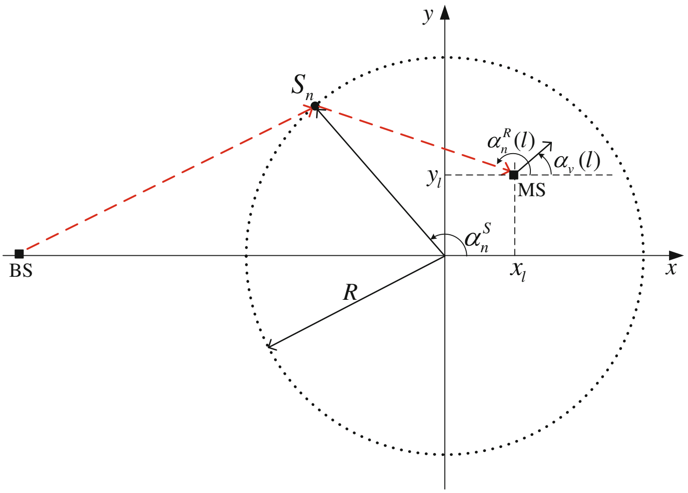
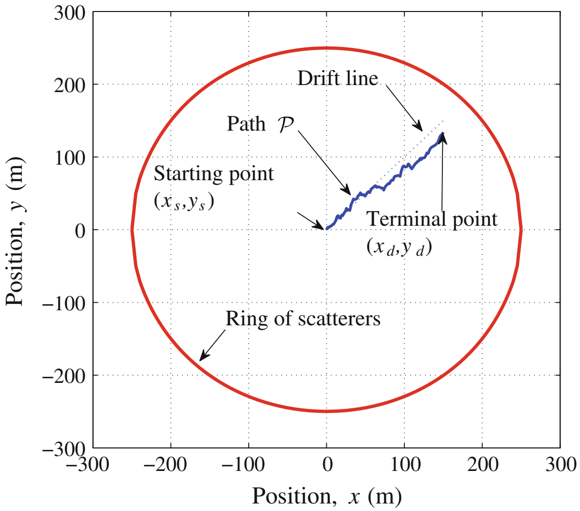


Fig. 1 The non-centred one-ring scattering model for a single-bounce scattering scenario

Fig. 2 Realization of a targeted Brownian path $\mathscr{P}$ in the ring of scatterers. The model parameters are $L = 100$, $a = 1$, $b = 0$, $\sigma_y = 2$, $x_s = 0$ m, $x_d = 150$ m, and $R = 250$ m



proposed Brownian path $\mathscr{P}$ in such a geometric scattering model is shown in Fig. 2, in which the starting point (x_0, y_0) of the path is set to the ring's center.[2]

The above mentioned propagation scenario is completely different from the one-ring scattering model [1–3], in which the MS is located at the center of the

[2] We have chosen the ring's center as the starting point of the movement to enable the verification of our numerical results (see Sect. 7) with the ones from the one-ring scattering model. However, the analytical results provided in the paper are not limited to such a special case.

ring, while its AOM is a deterministic variable. Therein, considering a very short observation time results in a stationary and isotropic channel model, while herein, the proposed path $\mathscr{P}$ justifies a non-stationary non-isotropic channel model. The proposed jittery path model $\mathscr{P}$ is also different from the smooth path model of [14]. Indeed, the random behavior of the AOM $\alpha_v(l)$ (see Fig. 2) allows a much more flexible non-stationary channel model than the one proposed in [14]. In what follows, after providing an expression for the complex channel gain, we study the statistical properties of the proposed non-stationary channel model.

5 The Complex Channel Gain

The propagation scenario presented in Sect. 4 is a non-stationary version of the typical fixed-to-mobile (F2M) scenario studied in [23, pp. 56–60]. Therein, the complex channel gain $\mu(t_l)$ of frequency-nonselective F2M channels was modeled by means of a complex stochastic process representing the sum of all scattered components as follows

$$\mu(t_l) = \lim_{N\to\infty} \sum_{n=1}^{N} c_n e^{j(2\pi f_n t_l + \theta_n)}. \tag{2}$$

In the equation above, c_n denotes the attenuation factor caused by the physical interaction of the emitted wave with the nth scatterer S_n, and f_n stands for the Doppler frequency[3] caused by the movement of the MS. In addition, the random variable θ_n represents the phase shift of the nth path, which is often assumed to be uniformly distributed between 0 and 2π [23, p. 59].

The complex channel gain in (2) suits the proposed non-stationary one-ring model, if we replace the Doppler frequency f_n by $f_n(t_l)$. This apparently minor change adds a great deal of mathematical computations to the statistical characterization of the channel.

6 Statistical Properties of the Channel

To investigate the statistical properties of the complex channel gain described in (2), let us start from the local AOA, which plays a key role in other statistical quantities. Notice that we defer the illustration and physical explanation of the analytical results to Sect. 7.

[3] The frequency shift caused by the Doppler effect is given by $f = f_{\max}\cos(\alpha)$, where $f_{\max} = f_0 v / c_0$ is the maximum Doppler frequency, f_0 denotes the carrier frequency, c_0 stands for the speed of light, and α equals the difference between the AOA and the AOM [24].

6.1 The Local Angles-of-Arrival

Referring to the geometric scattering model in Fig. 1, the AOA $\alpha_n^R(l)$ at the point (x_l, y_l) is given by

$$\alpha_n^R(l) = \arctan\left(\frac{R\sin(\alpha_n^S) - y_l}{R\cos(\alpha_n^S) - x_l}\right). \tag{3}$$

For a given position l, the only random variable in the right side of (3) is the AOS α_n^S. Since the number N of local scatterers tends to infinity in the reference model, it is mathematically convenient to assume that the discrete AOS α_n^S is a continues random variable α^S, which is assumed to be uniformly distributed between $-\pi$ and π (see Sect. 4). By applying the concept of transformation of random variables [25, p. 130] and performing some mathematical manipulations, it can be shown that the first-order density $p_{\mathbf{a}^R}(\alpha^R; l)$ of the stochastic process $\alpha^R(l)$ in (3) becomes

$$p_{\alpha^R}(\alpha^R; l) = \frac{1}{2\pi}\left(1 - \frac{x_l\cos(\alpha^R) + y_l\sin(\alpha^R)}{\sqrt{R^2 - (x_l\sin(\alpha^R) - y_l\cos(\alpha^R))^2}}\right) \tag{4}$$

in which $-\pi \leq \alpha^R < \pi$. It is worth mentioning that $p_{\alpha^R}(\alpha^R; l)$ in (4) depends strongly on the position (x_l, y_l) of the MS. This means that the AOA $\alpha^R(l)$ is *not* first-order stationary. As a special case, if the path P crosses the ring's center $(0, 0)$, then $p_{\alpha^R}(\alpha^R; l)$ in (4) reduces to $1/(2\pi)$, which is the AOA probability density function (PDF) of the one-ring model [1–3].

6.2 The Local Angles-of-Motion

By performing the linear interpolation scheme, the path $\mathscr{P}$ becomes continues and piecewise differentiable. This allows us to present the AOM $\alpha_v(l)$ at the location point (x_l, y_l) by the following expression

$$\begin{aligned}\alpha_v(l) &= \arctan\left(\frac{y_{l+1} - y_l}{x_{l+1} - x_l}\right)\\ &= \arctan\left(a + \sigma_y\frac{B_{l+1} - B_l}{x_{l+1} - x_l}\right).\end{aligned} \tag{5}$$

In the right side of (5), $B_{l+1} - B_l$ is the only random variable, which follows the Gaussian distribution of the form $N(0, \Delta x)$ (see Sect. 3.1). Again, by applying the concept of transformation of random variables, the PDF $p_{\mathbf{a}_v}(\alpha_v)$ of the AOM $\alpha_v(l)$ in (5) is given by

$$p_{\alpha_v}(\alpha_v) = \frac{1}{\sqrt{2\pi}\sigma\cos^2(\alpha_v)} e^{\frac{-(\tan(\alpha_v)-a)^2}{2\sigma^2}} \tag{6}$$

where $-\pi/2 \le \alpha_v \le \pi/2$ and $\sigma = \sigma_y/\sqrt{\Delta x}$. Notice that $p_{\mathbf{a}_v}(\alpha_v)$ in (6) is independent of the position (x_l, y_l) of the MS, meaning that the AOM α_v is first-order stationary. It can be shown that the mean α_v equals $\arctan(a)$, in which a is the slope of the drift of the path $\mathscr{P}$.

6.3 The Local Power Spectral Density

The local Doppler frequency $f(l)$ is obtained through a non-linear transformation of the local AOA $\alpha^R(l)$ and the local AOM $\alpha_v(l)$ of the MS. It follows

$$f(l) = f_{\max}\cos\left(\alpha^R(l) - \alpha_v(l)\right) \tag{7}$$

where $\alpha^R(l)$ is described by the first-order density $p_{\alpha^R}(\alpha^R; l)$ in (4). In addition, the angle $\alpha_v(l)$ denotes the AOM at the point (x_l, y_l) after realizing the path $\mathscr{P}$ (see 5). Accordingly, for a given position index l, the only random variable on the right-hand side of (7) is the AOA $\alpha^R(l)$. This is different from the approach used in [16], where the AOM is also assumed to be a random variable. To compute the first-order density $p_f(f; l)$ of the Doppler frequencies $f(l)$, we fix the position index l, and then we apply the concept of transformation of random variables. It follows

$$\begin{aligned} p_f(f; l) = {} & \frac{1}{2\pi f_{\max}\sqrt{1 - (f/f_{\max})^2}} \\ & \times \left(2 - \frac{x_l\cos(A_+(f;l)) + y_l\sin(A_+(f;l))}{\sqrt{R^2 - (x_l\sin(A_+(f;l)) - y_l\cos(A_+(f;l)))^2}}\right. \\ & \left. - \frac{x_l\cos(A_-(f;l)) + y_l\sin(A_-(f;l))}{\sqrt{R^2 - (x_l\sin(A_-(f;l)) - y_l\cos(A_-(f;l)))^2}}\right) \end{aligned} \tag{8}$$

where

$$A_\pm(f; l) = \arctan\left(\frac{y_{l+1} - y_l}{x_{l+1} - x_l}\right) \pm \arccos\left(\frac{f}{f_{\max}}\right) \tag{9}$$

for $-f_{\max} \le f \le f_{\max}$.

Referring to [23, p. 85], the first-order density $p_f(f;l)$ of the Doppler frequencies is proportional to the local PSD $S_{\mu\mu}(f;l)$ of the complex channel gain $\mu(t_l)$. This allows us to present $S_{\mu\mu}(f;l)$ by the following expression

$$S_{\mu\mu}(f;l) = 2\sigma_0^2 p_f(f;l). \tag{10}$$

In the equation above, $2\sigma_0^2$ is the mean power of $\mu(t_l)$, and $p_f(f;l)$ is given by (8). For the special case that the path $\mathscr{P}$ crosses the ring's center $(0,0)$, the first-order density $p_f(f;l)$ in (8) reduces to

$$p_{\mathbf{f}}(f;l) = \frac{1}{\pi f_{\max}\sqrt{1-(f/f_{\max})^2}} \tag{11}$$

which, after its multiplication by the mean power $2\sigma_0^2$, results in the Jakes PSD [24]. Notice that the Jakes PSD is not only associated with the stationary one-ring scattering model [1–3], but also with any other scattering model that is circularly symmetric with respect to the MS [4].

6.4 The Local Autocorrelation Function

With reference to the generalized Wigner-Ville spectrum [26, pp. 282–285], the local ACF $r_{\mu\mu}(\tau;l)$ of the non-stationary complex channel gain $\mu(t_l)$ can be attained by taking the inverse Fourier transform of the local PSD $S_{\mu\mu}(f;l)$ in (10). Accordingly, one can write

$$r_{\mu\mu}(\tau;l) = \int_{-f_{\max}}^{f_{\max}} S_{\mu\mu}(f;l)e^{j2\pi f\tau}df. \tag{12}$$

As a special case, if the path $\mathscr{P}$ goes across the ring's center, a scaled version of $p_f(f;l)$ in (11) can be employed to compute the inverse Fourier transform in (12). In this case, the local ACF $r_{\mu\mu}(\tau;l)$ in (12) reduces to $2\sigma_0^2 J_0(2\pi f_{\max}\tau)$, where $J_0(\cdot)$ stands for the zeroth-order Bessel function of the first kind [27, Eq. (8.411.1)].

7 Numerical Results

Channel modelling at the 2 GHz band is of great importance in mobile communications. With reference to the operating frequency of the universal mobile telecommunications system (UMTS), the carrier frequency $f_0 = 2.1$ GHz has been

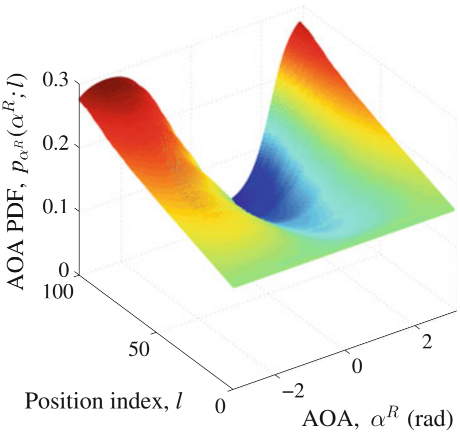


Fig. 3 The behavior of the first-order density $p_{\alpha^R}(\alpha^R;l)$ in (4) for the propagation scenario illustrated in Fig. 2

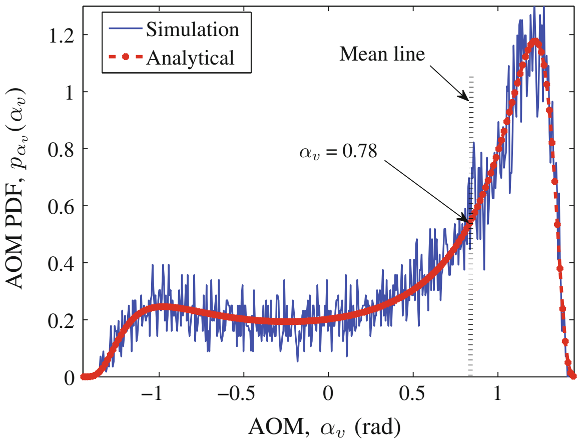
chosen in our numerical computations. In addition, we consider the path $\mathscr{P}$ shown in Fig. 2 as the travelling path of the MS. This allows us to have the positions (x_l, y_l) for $l = 0, 1, \ldots, L$. It is also assumed that the MS is moving with a velocity v_R of 80 km/h, which results in a maximum Doppler frequency $f_{\max}$ of 155.5 Hz. The mean power $2\sigma_0^2$ has been set to unity.

Figure 3 illustrates the first-order density $p_{\alpha^R}(\alpha^R;l)$ of the AOA process $\alpha^R(l)$ provided in (4). With reference to the path $\mathscr{P}$ shown in Fig. 2, the MS starts its movement from the center of the ring. This circularly symmetric starting point explains the uniform distribution of the AOA at $l = 0$. By moving along the path $\mathscr{P}$, the probability of receiving signals from the scatterers ahead reduces, whereas the probability of receiving from the scatterers behind increases. This behavior continuous up to $l = 100$, where $p_{\alpha^R}(\alpha^R, 100)$ takes its minimum value at $\alpha^R = \arctan(1) = 0.78$ radian.

Figure 4 displays the PDF $p_{\alpha_v}(\alpha_v)$ of the AOM $\alpha_v(l)$ in (6). The simulated AOM is also shown in this figure. An excellent match between the simulation and analytical results can be observed. The mean α_v equals $\arctan(1) = 0.78$ radian as shown in the figure. The plot shows explicitly the tendency of the MS to follow the predetermined drift of the path $\mathscr{P}$. This tendency depends solely on the slope a of the drift, which has been set to 1 herein. It is noteworthy that if the randomness σ_y of the path tends to zero, the AOM PDF approaches the delta function at $\alpha_v = 0.78$ radian.

Figure 5 depicts the local PSD $S_{\mu\mu}(f;l)$ presented in (10). The classical Jakes PSD with a U-shape can be observed in the stationary case ($l = 0$), where the MS is located at the ring's center. At this position, $S_{\mu\mu}(f, 0)$ is a symmetric function with respect to f, indicating that the channel is instantaneously isotropic. However, this feature does not hold if the MS continuous its motion along the path $\mathscr{P}$. In this regard, by increasing l, an asymmetric behavior of the local PSD $S_{\mu\mu}(f;l)$ can be observed. Notice that moving along the path $\mathscr{P}$ results in confronting a lower number of scatterers ahead and a higher number of them behind the MS. This

Fig. 4 The behavior of the AOM PDF $p_{\alpha_v}(\alpha_v)$ in (6) for the propagation scenario illustrated in Fig. 2



allows a higher and a lower probability of negative and positive Doppler shifts as shown in Fig. 5.

Figure 4 displays the PDF $p_{\alpha_v}(\alpha_v)$ of the AOM $\alpha_v(l)$ in (6). The simulated AOM is also shown in this figure. An excellent match between the simulation and analytical results can be observed. The mean α_v equals $\arctan(1) = 0.78$ radian as shown in the figure. The plot shows explicitly the tendency of the MS to follow the predetermined drift of the path $\mathscr{P}$. This tendency depends solely on the slope a of the drift, which has been set to 1 herein. It is noteworthy that if the randomness σ_y of the path tends to zero, the AOM PDF approaches the delta function at $\alpha_v = 0.78$ radian.

Figure 5 depicts the local PSD $S_{\mu\mu}(f;l)$ presented in (10). The classical Jakes PSD with a U-shape can be observed in the stationary case ($l = 0$), where the MS is located at the ring's center. At this position, $S_{\mu\mu}(f,0)$ is a symmetric function with respect to f, indicating that the channel is instantaneously isotropic. However, this feature does not hold if the MS continuous its motion along the path $\mathscr{P}$. In this regard, by increasing l, an asymmetric behavior of the local PSD $S_{\mu\mu}(f;l)$ can be observed. Notice that moving along the path $\mathscr{P}$ results in confronting a lower number of scatterers ahead and a higher number of them behind the MS. This allows a higher and a lower probability of negative and positive Doppler shifts as shown in Fig. 5.

Figure 6 shows the absolute value of the local ACF $r_{\mu\mu}(\tau;l)$ given in (12). Notice that due to the asymmetric behavior of the PSD $S_{\mu\mu}(f;l)$ (see Fig. 5), the ACF $r_{\mu\mu}(\tau;l)$ is in general complex. A quite time-varying behavior of the ACF can be observed. It is also noteworthy that for a given time difference $\tau \neq 0$, the correlation increases through some fluctuations if l grows.

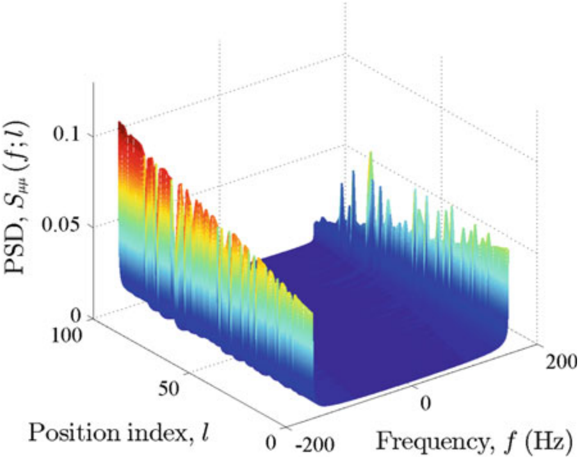


Fig. 5 The behavior of the local PSD $S_{\mu\mu}(f;l)$ in (10) for the propagation scenario illustrated in Fig. 2

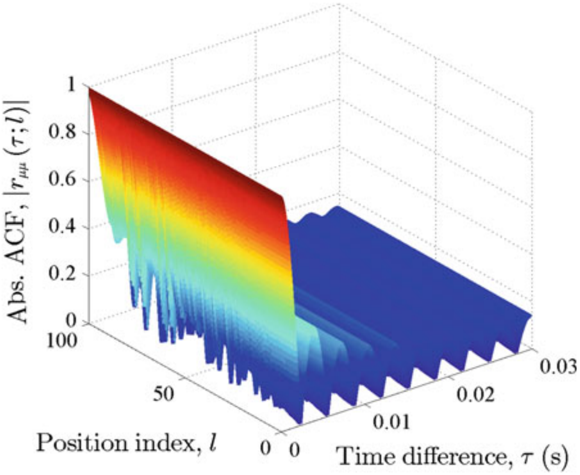


Fig. 6 The behavior of the absolute value of the local ACF $|r_{\mu\mu}(\tau;l)|$ (see (12)) for the propagation scenario illustrated in Fig. 2

8 Conclusion

In this paper, we have proposed a targeted Brownian path model to explain the travelling path of the MS. The proposed path model has a tendency to follow a preferred direction. To describe the propagation area, we have proposed a non-centred one-ring scattering model, in which the MS is not necessarily located at the ring's center. We have assumed that the MS is moving along the proposed targeted Brownian path model in such a geometric scattering model. It has been turned out that the proposed path model results in a non-stationary non-isotropic channel model. As a special case, the stationary isotropic one-ring scattering model can also be obtained from the proposed non-stationary channel model. The statistical properties of the proposed channel model have been derived, illustrated, and discussed extensively. It has been shown that the AOA process is first-order non-stationary, while the PDF of the AOM is stationary. The corresponding PSD of the Doppler frequencies and ACF of the complex channel gain have also been provided, showing that these characteristics are heavily time-dependent. Validating the analytical results by means of empirical data needs to be addressed in future works.

References

1. A. Abdi, M. Kaveh, A space-time correlation model for multielement antenna systems in mobile fading channels. IEEE J. Sel. Areas Commun. **20**(3), 550–560 (2002)
2. M. Pätzold, B.O. Hogstad, A space-time channel simulator for MIMO channels based on the geometrical one-ring scattering model, in *Proceedings of the 60th IEEE Semiannual Vehicular Technology Conference, VTC 2004-Fall*, vol. 1 (Los Angeles, 2004), pp. 144–149
3. D.S. Shiu, G.J. Foschini, M.J. Gans, J.M. Kahn, Fading correlation and its effect on the capacity of multielement antenna systems. IEEE Trans. Commun. **48**(3), 502–513 (2000)
4. A. Borhani, M. Pätzold, A unified disk scattering model and its angle-of-departure and time-of-arrival statistics. IEEE Trans. Veh. Technol. **62**(2), 473–485 (2013)
5. K.T. Wong, Y.I. Wu, M. Abdulla, Landmobile radiowave multipaths' DOA-distribution: assessing geometric models by the open literature's empirical datasets. IEEE Trans. Antennas Propag. **58**(2), 946–958 (2010)
6. A. Gehring, M. Steinbauer, I. Gaspard, M. Grigat, Empirical channel stationarity in urban environments, in *Proceedings of the 4th European Personal Mobile Communications Conference*, Vienna, 2001
7. A. Ispas, G. Ascheid, C. Schneider, R. Thom, Analysis of local quasi-stationarity regions in an urban macrocell scenario, in *Proceedings of the 71th IEEE Vehicular Technology Conference, VTC 2010-Spring*. Taipei, 2010
8. D. Umansky, M. Pätzold, Stationarity test for wireless communication channels, in *Proceedings of the IEEE Global Communications Conference, IEEE GLOBECOM 2009*. Honolulu
9. A. Paier, J. Karedal, N. Czink, H. Hofstetter, C. Dumard, T. Zemen, F. Tufvesson, A.F. Molisch, C.F. Mecklenbräucker, Characterization of vehicle-to-vehicle radio channels from measurement at 5.2 GHz. Wirel. Pers. Commun. **50**(1), 19–32 (2009)
10. A. Chelli, M. Pätzold, A non-stationary MIMO vehicle-to-vehicle channel model based on the geometrical T-junction model, in *Proceedings of the International Conference on Wireless Communications and Signal Processing, WCSP 2009*. Nanjing, 2009
11. A. Ghazal, C. Wang, H. Hass, R. Mesleh, D. Yuan, X. Ge, A non-stationary MIMO channel model for high-speed train communication systems, in *Proceedings of the 75th IEEE Vehicular Technology Conference, VTC 2012-Spring*. Yokohama, 2012
12. J. Karedal, F. Tufvesson, N. Czink, A. Paier, C. Dumard, T. Zemen, C.F. Mecklenbräuker, A.F. Molisch, A geometry-based stochastic MIMO model for vehicle-to-vehicle communications. IEEE Trans. Wirel. Commun. **8**(7), 3646–3657 (2009)
13. G. Matz, On non-WSSUS wireless fading channels. IEEE Trans. Wirel. Commun. **4**(5), 2465–2478 (2005)
14. A. Borhani, M. Pätzold, A non-stationary one-ring scattering model. In: *Proceedings of the IEEE Wireless Communications and Networking, Conference (WCNC'13)*. Shanghai, 2013
15. P. Pearle, B. Collett, K. Bart, D. Bilderback, D. Newman, S. Samuels, What Brown saw and you can too. Am. J. Phys. **78**(12), 1278–1289 (2010)

16. A. Borhani, M. Pätzold, Modelling of non-stationary mobile radio channels incorporating the Brownian mobility model with drift, in *Proceedings of the World Congress on Engineering and Computer Science, WCECS 2013*. Lecture Notes in Engineering and Computer Science, San Francisco, 23–25 Oct, pp. 695–700
17. B.H. Fleury, D. Dahlhaus, Investigations on the time variations of the wide-band radio channel for random receiver movements, in *Proceedings of the IEEE International Symposium on Spread Spectrum Techniques and Applications (ISSSTA '94)*, vol. 2. Oulu, pp. 631–636
18. T. Camp, J. Boleng, V. Davies, A survey of mobility models for ad hoc network research. Wirel. Commun. Mobile Comput. **2**(5), 483–502 (2002)
19. A. Einstein, Über die von der molekularkinetischen theorie der wärme geforderte bewegung von in ruhenden flüssigkeiten suspendierten teilchen. Ann. Phys. **17**, 549–560 (1905)
20. P. Langevin, Sur la théorie du mouvement brownien. C. R. Acad. Sci. Paris **146**, 530–533 (1908)
21. D.S. Lemons, A. Gythiel, On the theory of Brownian motion. Am. J. Phys. **65**(11), 530–533 (1979)
22. R.C. Earnshaw, E.M. Riley, *Brownian Motion: Theory, Modelling and Applications* (Nova Science Pub Inc, New York, 2011)
23. M. Pätzold, *Mobile Fading Channels*, 2nd edn. (Wiley, Chichester, 2011)
24. W.C. Jakes (ed.), *Microwave Mobile Communications* (IEEE Press, Piscataway, 1994)
25. A. Papoulis, *Probability, Random Variables, and Stochastic Processes*, 3rd edn. (McGraw-Hill, New York, 1991)
26. F. Hlawatsch, F. Auger, *Time-Frequency Analysis: Concepts and Methods* (Wiley, London, 2008)
27. I.S. Gradshteyn, I.M. Ryzhik, *Table of Integrals, Series, and Products*, 7th edn. (Elsevier Academic Press, Amsterdam, 2007)

Chapter 34
On the Performance of MU SIMO/MIMO UWB Communication Systems Applying Time-Reversal Technique

Vu Tran-Ha, Duc-Dung Tran, Dac-Binh Ha and Een-Kee Hong

Abstract Time Reversal (TR) is a technique to focus broadband signals tightly in time and space domains. Previously, this technique has been used in acoustics, medical and especially in underwater communication applications. TR technique is combined with Ultra Wideband (UWB) technology to offer a new efficient solution for the cost and complexity degradation of UWB receivers. It is capable of interference reduction such as: inter-symbol interference (ISI), inter-user interference (IUI)…, meanwhile the user equipments have no need of complex equalizers at the transmitter and receiver. In this book chapter, we focus on presenting the operational mechanism and giving some specific results which concern with applying TR technique for UWB communication systems. Specifically, the channel capacity is investigated in SIMO and MIMO UWB systems.

Keywords Channel capacity · ISI · IUI · MIMO · Time reversal · UWB

1 Introduction

Ultra Wideband (UWB) is an attractive research direction in recent years because of its capability of high-speed communication in a short distance [1–4]. UWB solved effectively the problems of bandwidth limit in wireless environments [5].

V. Tran-Ha (✉) · E.-K. Hong
School of Electronic and Information, Kyung Hee University, Yongin, South Korea
e-mail: havutran.dhkh@gmail.com

E.-K. Hong
e-mail: ekhong@khu.ac.kr

V. Tran-Ha · D.-D. Tran · D.-B. Ha
Institute Research and Development, Duy Tan University, Danang, Vietnam
e-mail: dung.td.1227@gmail.com

D.-B. Ha
e-mail: hadacbinh@duytan.edu.vn

H. K. Kim et al. (eds.), *Transactions on Engineering Technologies*,
DOI: 10.1007/978-94-017-9115-1_34,

However, it is realized that channels in reality are multi-path fading channels, so problems affecting quality of transmission in UWB systems serving multi-user (MU) are really complex. We can resolve these problems by combining UWB systems and TR technique to improve transmission rate and minimize the interference of channels which decrease the quality of UWB systems [3, 6, 7].

In UWB systems, the dense multi-path components can be useful for the purpose of both data communications and correct positioning, however, the multi-path effects also cause the great difficulty and complexity for the UWB synchronizer and equalizer. Besides, in order to harvest even half of the energy distributed in the entire impulse responses, Rake receivers with at least 20 taps, may be potentially much more, must be constructed [8]. TR technique has been extensively used in acoustic, medical applications and underwater communications [9, 10]. Its advantage is decreasing bad effects caused by environments, such as: Inter-Symbol Interference (ISI), Inter-User Interference (IUI), …without the need of using complex equalizers at transmitters and receivers. The combination between time-reversal (TR) technique and UWB systems offers a new possibility for decreasing the cost and complexity of the UWB receiver. And by convering the signal at one specific time instant and one specific location, it also provide a solution to multi-user access and secure communication schemes. In the TR systems, multi-access mechanism is based on the unique of channel impulse responses of the environments where a base station (BS) is forwarded to any users [11].

In [2], Vu Tran-Ha and the co-authors presented the results related to the channel capacity of MU MIMO UWB TR system in environment conditions when the correlation between antennas is considered. In this book chapter, we focus on introducing TR technique and simulating its operational mechanism. Besides, we collapse the problem in [2] when the correlation between antennas is not considered and show a number of specific results related to the comparison of capacity of SIMO/MIMO UWB systems in the case with/without TR applied. From the simulation results obtained, applying TR technique for UWB systems has highly effective in increasing the channel capacity significantly.

The rest of this chapter is organized as follow: part II is time reversal technique description, channel capacity of UWB systems are described in part III, part IV are numerical results and discussion, and part V is conclusion.

2 Time Reversal Technique

TR is a technique to focus broadband signals tightly in space and time where the multi-path channel with rich scattering is exploited by active modulating the signal at the transmitter side. The modulation scheme base on the state channel information, instead of being processed at the receiver by equalizers or Rake combiners as in the traditional communication systems [8]. This technique has been extensively used in acoustic, medical applications and underwater communications.

The main advantages of the TR technique are [8]

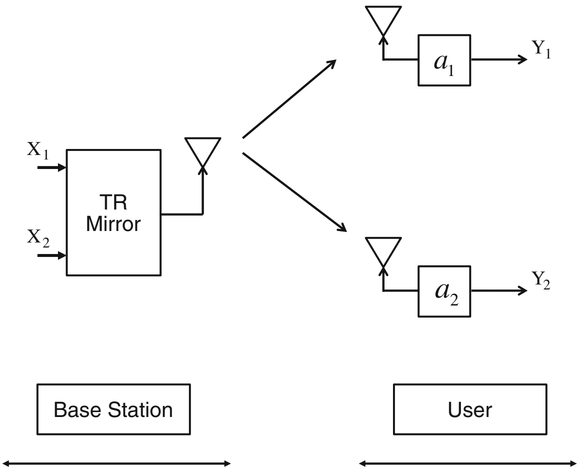


Fig. 1 UWB TR system

- Temporal focusing: The received signal is compressed in the time domain. Owing to this property, the inter-symbol interference (ISI) caused by the original multi-path channel is greatly reduced.
- Spatial focusing: The received signal is focused on the intended user at some specific position in space. This is very useful in realistic environments where the interference from co-channel users limits the capacity of each user. If the transmitter is able to focus precisely, an ideal space-division multiple access (SDMA) technique and the location-based security might be enabled.

Because of the simplicity in the principle and aforementioned advantages of TR technology, the idea of applying TR technique in wireless communication has gained much attention recently [2, 3, 11]. The principle of TR technique is to use state channel information to create waveform used to transmit signals. Suppose that we have a structure similar to the above example includes one transmit base station and two receivers. The system is simplified in only one antenna at each user, as Fig. 1.

In order to use TR technique, the state channel information needs to be known in advance. Therefore, first of all, receiver has to send to base station an impulse. When the impulse is transmitted through environment base station, received signal is the channel impulse response (CIR). The received signal at the base station (is called X1) look like the Fig. 2.

At this time, the base station store the received signal and then reverse it in time axis. The time-reversed signal will be used to transfer data between the base station and the receiver. The time-reversed signal (is called X2) look like the Fig. 3.

Thus, we have identified signal waveforms that will be used to transmit data based on state channel information. When X2 is used for communication, the received signal at the receiver (is called X3) is the waveform as Fig. 4.

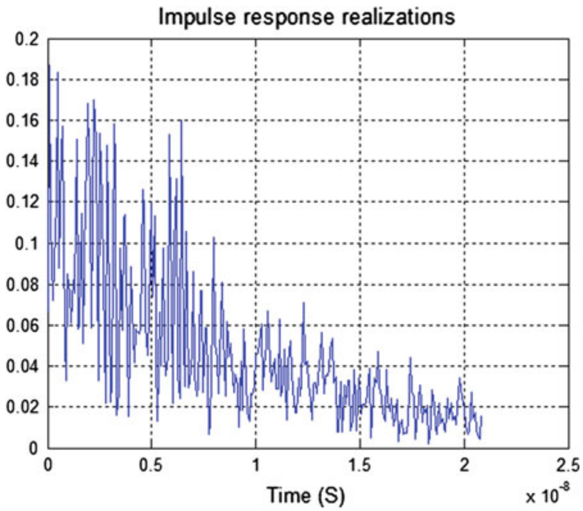


Fig. 2 X1 signal form

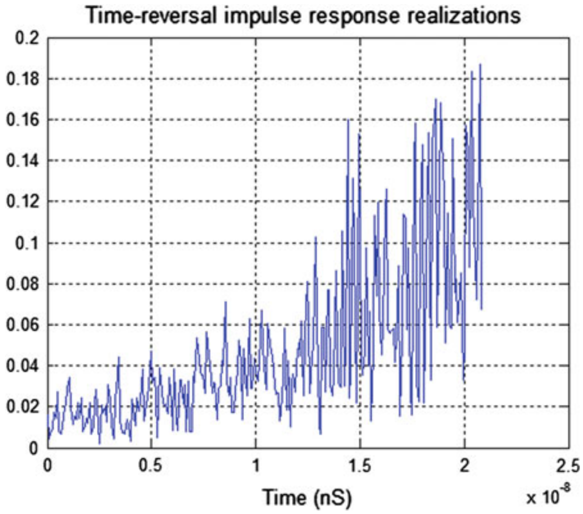


Fig. 3 X2 signal form

The energy of received signal is converged at a certain position in the time domain. Furthermore, if and only if X2 signal form transmitted through environment which its CIR is exactly X1 form, we will get X3 signal form at the receiver. It means that, only the intended receiver receive the signal form above. For the other receivers which are not intended, the received signals have the waveform (X4 form) as Fig. 5.

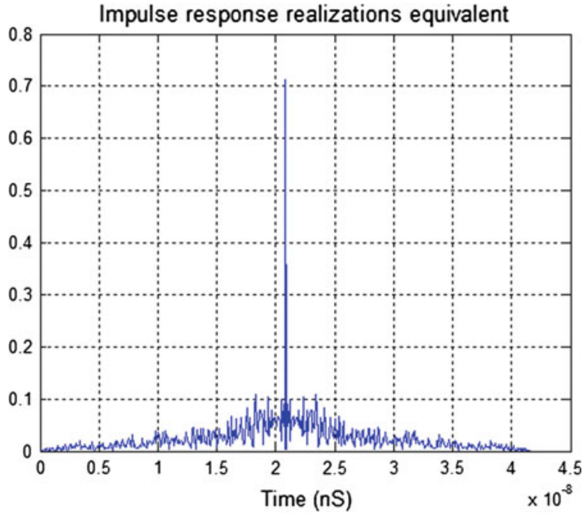


Fig. 4 X3 signal form

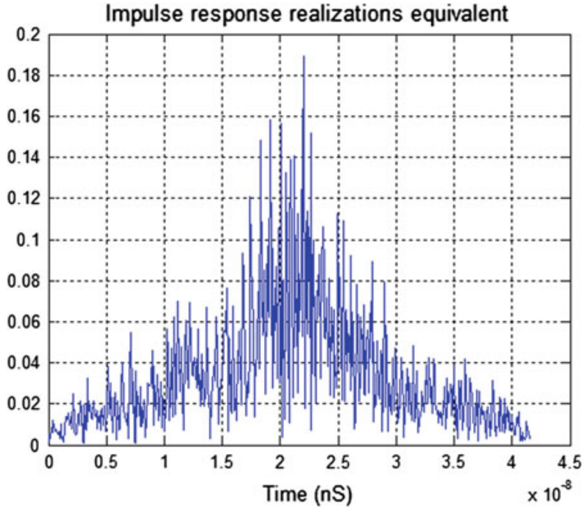


Fig. 5 X4 signal form

Therefore, the focused energy of X4 signal is much smaller than X3 signal. It is also considered as amount of interference between receive stations or inter-user interference (IUI). In summary, TR UWB systems takes advantages of the diversification and the unique of CIRs to focus signal energy as well as provide a simple and effective multi-access mechanism. This multi-access method is similar to (Code Division Multiple Access) CDMA, in which, each CIR will play a role as Pseudo Noise (PN) code [11].

3 Channel Capacity of SIMO/MIMO UWB Systems

In this part, we will describe SIMO/MIMO UWB system and show the formulas to calculate its channel capacity when it is applied and not applied TR technique (UWB and UWB TR).

The difference between UWB and UWB TR system is that, UWB TR system only operates when it anticipates the CIRs' information forwarded each user.

Therefore, in UWB TR system, first of all, intended users will send an impulse to base station, the received signal form at BS is CIRs form of that environment. When the base station received CIRs information from the users, the block Time Reversal Mirror will use this information of CIRs to create waveforms which are used for communication between the base station and intended users.

In this paper, we do not consider the correlation between transmit antennas and receive antennas. We suppose that the number of transmit antennas is M_T and the number of receive antennas is M_R.

3.1 SIMO UWB and SIMO UWB TR Systems

SIMO (Single Input Multi Output) system is a wireless system using single transmit antennas ($M_T = 1$) and multiple receive antennas ($M_R > 1$). The block diagram of SIMO UWB and SIMO UWB TR systems are shown in Figs. 6 and 7, respectively.

The received signal at the users has the following form

$$\mathbf{Y} = \mathbf{H_{SM}} * \mathbf{X} + \mathbf{n}, \tag{1}$$

where $\mathbf{Y}$ is the received signal vector at the users; $\mathbf{X}$ is the transmitted signal vector; $\mathbf{H_{SM}}$ is the CIRs matrix with dimension $M_R \times M_T$ from base station to the users and $\mathbf{n}$ is the white Gaussian noise vector at the users. $\mathbf{H_{SM}}$ can be written as

$$\mathbf{H_{SM}} = \begin{bmatrix} h_{11} \\ h_{21} \\ \vdots \\ h_{M_R 1} \end{bmatrix} \tag{2}$$

Where h_{i1} is the CIR between the transmit antenna and the i-th receive antenna of the user. It is represented as

$$h_{i1}(t) = \sum_{l=0}^{L-1} \alpha_l^{i1} \delta(t - \tau_l^{i1}), \quad i = 1, \ldots, M_R, \tag{3}$$

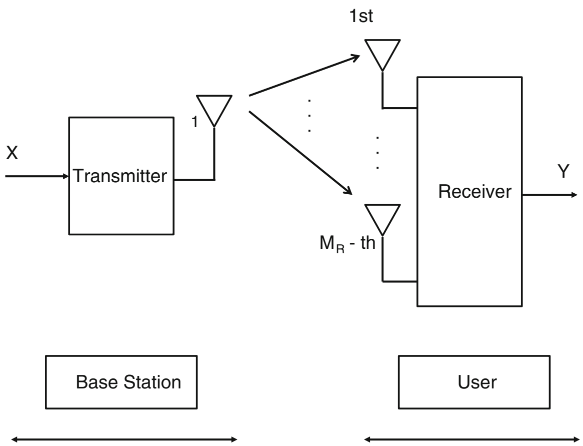


Fig. 6 Block diagram of SIMO UWB system

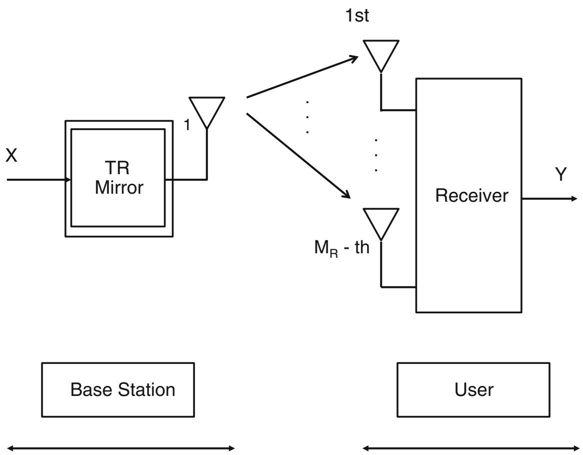


Fig. 7 Block diagram of SIMO UWB TR system

with α_l^{i1} and τ_l^{i1} are the amplitude and the delay of the l-th tap, respectively. The discrete time form of $h_{i1}(t)$ is expressed as

$$h_{i1} = [h_{i1}[0]\ h_{i1}[1]\ \ldots\ h_{i1}[L-1]], \tag{4}$$

where $h_{i1}[k]$, $k = 0, \ldots, L-1$ is the k-th tap of CIR with the length of L, $\delta[]$ is the Dirac pulse function. For each downlink, we assume that there are independent circular symmetric complex Gaussian (CSCG) random variables with zero mean and variance

$$\mathbf{E}\left[|h_{i1}[k]|^2\right] = e^{-\frac{kT_S}{\sigma_T}},\ 0 \le k \le L-1, \tag{5}$$

with T_S is the sampling time of the system; σ_T is the delay spread of the channel.

Channel capacity of SIMO UWB system will be calculated by the following formula:

$$C_{No\_TR}^{SM} = log_2\left[det\left(\mathbf{I}_{M_R} + SNR.\mathbf{H_{SM}}.\mathbf{H}_{\mathbf{SM}}^{H}\right)\right], \tag{6}$$

where $\mathbf{I}_{M_R}$ is the unit matrix with dimension $M_R \times M_R$, SNR is the signal to noise ratio and

$$\mathbf{H}_{\mathbf{SM}}^{H} = \begin{bmatrix} h_{11}^{*} & h_{21}^{*} & \ldots & h_{M_R1}^{*} \end{bmatrix}, \tag{7}$$

with h_{i1}^{*} is complex conjugate of h_{i1}, $i = 1,\ldots,M_R$.

With aforementioned UWB TR system, the TR Mirror block of base station records and stores received information which is used for processing transmitted signal. Let $\mathbf{G_{SM}}$ is TR Mirror's matrix, which is expressed as

$$\mathbf{G_{SM}} = \begin{bmatrix} g_{11} & g_{12} & \ldots & g_{1M_R} \end{bmatrix} \tag{8}$$

with the elements of $\mathbf{G_{SM}}$ are [11]

$$g_{1i}[k] = \frac{h_{i1}^{*}[L-1-k]}{\sqrt{E\left[\sum_{l=0}^{L-1} |h_{i1}[l]^2|\right]}}, \quad 0 \leq \mathrm{k} \leq \mathrm{L}-1, \tag{9}$$

which are the normalized complex conjugate of time-reversed $\{h_{i1}[k]\}$ $(1 \leq i \leq M_R)$, $E[.]$ is expectation operator.

Let $\hat{\mathbf{H}}_{\mathbf{SM}}$ is the equivalent CIRs matrix, which is represented as

$$\hat{\mathbf{H}}_{\mathbf{SM}} = \mathbf{H_{SM}} * \mathbf{G_{SM}} = \begin{bmatrix} \hat{h}_{11} & \hat{h}_{12} & \ldots & \hat{h}_{1M_R} \\ \hat{h}_{21} & \hat{h}_{22} & \ldots & \hat{h}_{2M_R} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{h}_{M_R1} & \hat{h}_{M_R2} & \ldots & \hat{h}_{M_RM_R} \end{bmatrix}, \tag{10}$$

where $\hat{h}_{ij} = h_{i1} * g_{1j},\ i,j = 1,\ldots,M_R$.

And then, Eq. (1) can rewrite as follows

$$\mathbf{Y} = \hat{\mathbf{H}}_{\mathbf{SM}} * \mathbf{X} + \mathbf{n}, \tag{11}$$

Channel capacity of MIMO UWB TR is

$$C_{TR}^{SM} = log_2\left[det\left(\mathbf{I}_{M_R} + SNR.\hat{\mathbf{H}}_{\mathbf{SM}}.\hat{\mathbf{H}}_{\mathbf{SM}}^{H}\right)\right], \tag{12}$$

where $\mathbf{I}_{M_R}$ is the unit matrix with dimension $M_R \times M_R$, SNR is the signal to noise ratio and

$$\hat{\mathbf{H}}_{\mathbf{SM}}^{H} = \begin{bmatrix} \hat{h}_{11}^{*} & \hat{h}_{21}^{*} & \cdots & \hat{h}_{M_R 1}^{*} \\ \hat{h}_{12}^{*} & \hat{h}_{22}^{*} & \cdots & \hat{h}_{M_R 2}^{*} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{h}_{1 M_R}^{*} & \hat{h}_{2 M_R}^{*} & \cdots & \hat{h}_{M_R M_R}^{*} \end{bmatrix}, \tag{13}$$

with $\hat{h}_{ij}^{*}$ is complex conjugate of $\hat{h}_{ij}$, $i = 1, \ldots, M_R$; $j = 1, \ldots, M_R$.

3.2 MIMO UWB and MIMO UWB TR Systems

MIMO (Multi Input Multi Output) system is a wireless system using multiple transmit antennas ($M_T > 1$) and multiple receive antennas ($M_R > 1$). The block diagram of MIMO UWB and MIMO UWB TR systems are shown in Figs. 8 and 9, respectively.

Similar to 3.1, we have the following results

In this case, the CIRs matrix (with dimension $M_R \times M_T$) is written as:

$$\mathbf{H}_{\mathbf{MM}} = \begin{bmatrix} h_{11} & h_{12} & \cdots & h_{1 M_T} \\ h_{21} & h_{22} & \cdots & h_{2 M_T} \\ \vdots & \vdots & \ddots & \vdots \\ h_{M_R 1} & h_{M_R 2} & \cdots & h_{M_R M_T} \end{bmatrix}; \quad \mathbf{H}_{\mathbf{MM}}^{H} = \begin{bmatrix} h_{11}^{*} & h_{21}^{*} & \cdots & h_{M_R 1}^{*} \\ h_{12}^{*} & h_{22}^{*} & \cdots & h_{M_R 2}^{*} \\ \vdots & \vdots & \ddots & \vdots \\ h_{1 M_T}^{*} & h_{2 M_T}^{*} & \cdots & h_{M_R M_T}^{*} \end{bmatrix}, \tag{14}$$

TR Mirror's matrix $\mathbf{G}_{\mathbf{MM}}(1 \times M_R)$ is

$$\mathbf{G}_{\mathbf{MM}} = \begin{bmatrix} g_{11} & g_{12} & \cdots & g_{1 M_R} \\ g_{21} & g_{22} & \cdots & g_{2 M_R} \\ \vdots & \vdots & \ddots & \vdots \\ g_{M_T 1} & g_{M_T 2} & \cdots & g_{M_T M_R} \end{bmatrix} \tag{15}$$

with the elements of $\mathbf{G}_{\mathbf{MM}}$ are [11]

$$g_{ji}[k] = \frac{h_{ij}^{*}[L-1-k]}{\sqrt{E\left[\sum_{l=0}^{L-1} |h_{ij}[l]^{2}|\right]}}, \quad 0 \le \mathrm{k} \le \mathrm{L}-1, \tag{16}$$

where $1 \le i \le M_R$, $1 \le j \le M_T$.

From there, the equivalent CIRs matrix $\hat{\mathbf{H}}_{\mathbf{MM}}$ is written as

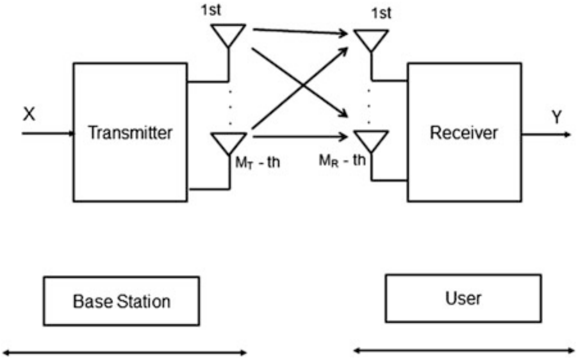


Fig. 8 Block diagram of MIMO UWB system

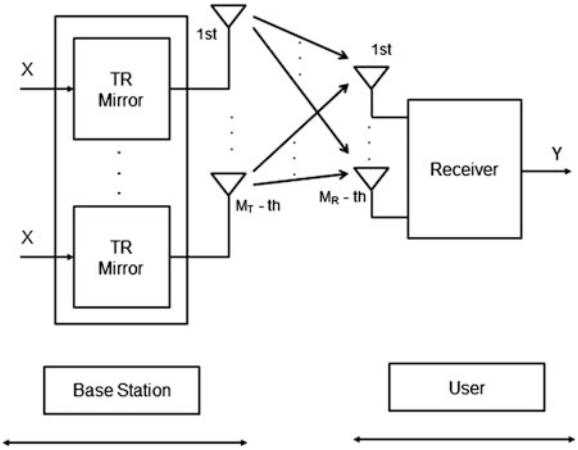


Fig. 9 Block diagram of MIMO UWB TR system

$$\hat{\mathbf{H}}_{\mathbf{MM}} = \mathbf{H}_{\mathbf{MM}} * \mathbf{G}_{\mathbf{MM}} = \begin{bmatrix} \hat{h}_{11} & \hat{h}_{12} & \dots & \hat{h}_{1M_R} \\ \hat{h}_{21} & \hat{h}_{22} & \dots & \hat{h}_{2M_R} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{h}_{M_R1} & \hat{h}_{M_R2} & \dots & \hat{h}_{M_RM_R} \end{bmatrix}, \tag{17}$$

where $\hat{h}_{ij} = \sum_{m=1}^{M_T} h_{im} * g_{mj},\ i,j = 1,\ldots,M_R$.

Channel capacity of MIMO UWB and MIMO UWB TR are respectively

$$C_{No\_TR}^{MM} = log_2\left[det\left(\mathbf{I}_{M_R} + SNR.\mathbf{H}_{\mathbf{MM}}.\mathbf{H}_{\mathbf{MM}}^H\right)\right], \tag{18}$$

$$C_{TR}^{MM} = log_2\left[det\left(\mathbf{I}_{M_R} + SNR.\hat{\mathbf{H}}_{\mathbf{MM}}.\hat{\mathbf{H}}_{\mathbf{MM}}^H\right)\right]. \tag{19}$$

Table 1 Simulation parameters

Parameters	System values
Number of transmit antennas (M_T)	[1, 3]
Number of receive antennas (M_R)	[3, 5, 7]
Environment	Rayleigh
Number of users (N)	3
Sampling time of the system (T_S)	$125T_S$
Delay spread of the channel (σ_T)	$\frac{1}{6}.10^{-9}$
Length of CIRs (L)	300
SNR	−10 to 20 dB

4 Numerical Results and Discussion

To carry out this comparison, we simulated the channel capacity of SIMO/MIMO UWB system with/without TR technique applied (SIMO/MIMO UWB and SIMO/MIMO UWB TR systems). The simulation parameters are shown in Table 1.

The simulated channel capacity of aforementioned UWB systems are shown as Figs. 10 and 11.

Figure 10 shows that, with $SNR = 5\text{dB}$, $L = 300$ the channel capacity of 1 × 3, 1 × 5, 1 × 7 SIMO UWB systems and 1 × 3, 1 × 5, 1 × 7 SIMO UWB TR systems are: $C_{No-TR}^{1\times3}$ approximates 6 bps/Hz while $C_{TR}^{1\times3}$ approximates 11 bps/Hz; $C_{No-TR}^{1\times5}$ approximates 10 bps/Hz while $C_{TR}^{1\times5}$ approximates 21 bps/Hz; $C_{No-TR}^{1\times7}$ approximates 14 bps/Hz while $C_{TR}^{1\times7}$ approximates 33 bps/Hz. Thus, the channel capacity of MIMO UWB TR systems are higher than the channel capacity of MIMO UWB systems significantly.

Figure 10 also indicates that, in considered SIMO UWB systems and SIMO UWB TR systems, 1 × 7 SIMO UWB systems and 1 × 7 SIMO UWB TR systems obtains the highest channel capacity. This means that, the more the number of antennas is used, the more increasing the channel capacity of the systems are.

Figure 11 shows how the channel capacity of MIMO UWB and MIMO UWB TR systems vary with SNR for variable the number of antennas and L = 300. Similar to 10, in this figure, we can also see that, the channel capacity of MIMO UWB (C_{No_TR}) is less than MIMO UWB TR (C_{TR}). And the more the number of antennas is used, the more increasing the channel capacity of the systems are.

The combination between UWB system and Time Reversal (TR) technique help to improve transmission rate without using complex equalizers at the transmitters and receivers. In other words, this combination has increased channel capacity of UWB system significantly meanwhile the cost and the complexity of UWB receiver is reduced.

The Gain (G) of the UWB system is defined as

Fig. 10 Channel capacity of SIMO UWB and SIMO UWB TR systems and different the number of antennas

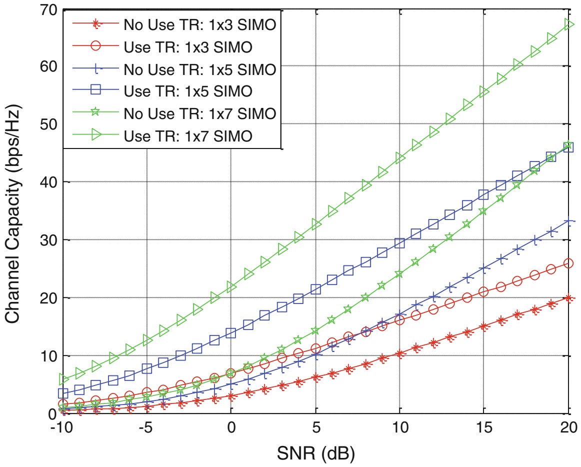
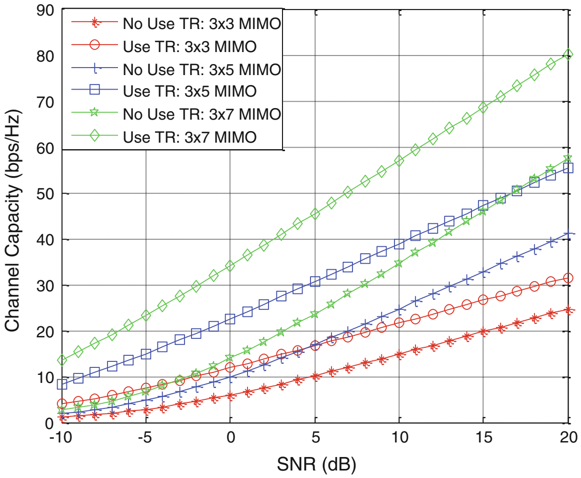


Fig. 11 Channel capacity of MIMO UWB and MIMO UWB TR systems and different the number of antennas



$$G = \frac{C_{TR}}{C_{No-TR}}, \tag{20}$$

where C_{TR} and C_{No-TR} are the channel capacity of UWB system with and without TR technique applied, respectively. The simulation results of gain G of SIMO UWB and MIMO UWB systems are shown as Figs. 12 and 13, respectively.

From Fig. 12, we obsever that, in the 1 × 3, 1 × 5, 1 × 7 SIMO systems, the 1 × 7 SIMO system has the highest gain and the 1 × 3 SIMO system has the smallest gain. At the same time, the more increasing the SNR (dB) value, the more decreasing the gain of the UWB system. In other words, the UWB systems combined with TR technique show superior performance in high noise environments (SNR is low). Similar conclusions is also obtained in Fig. 13.

Fig. 12 Gain of SIMO UWB systems

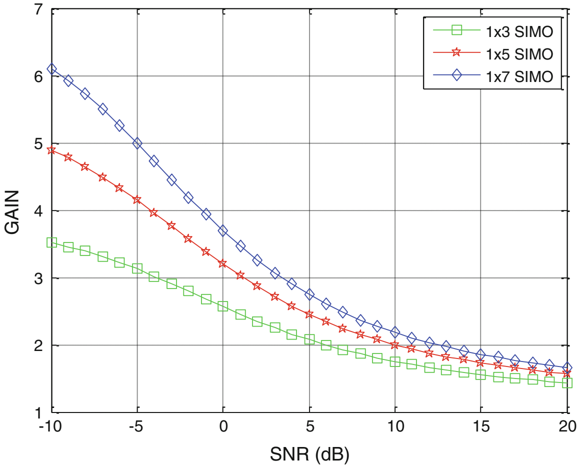
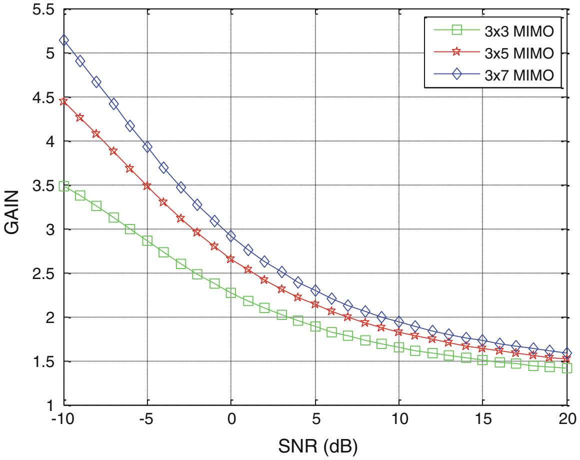


Fig. 13 Gain of MIMO UWB systems



5 Conclusion

UWB technology has been showing its preeminence and advantages in high-speed communication in short distance. And UWB systems operate more effectively (both quality and cost) when they are combined with TR technique. In this book chapter, we have focused on investigating and simulating the channel capacity of SIMO/MIMO UWB systems in two cases, when the TR technique is applied and not applied. Furthermore, we also have considered the impact of channel estimation error on the channel capacity of the systems. The simulation results indicated that the applying TR technique increased significantly channel capacity of UWB system. At the same time, these results also illustrated that the affectation of channel estimation error on the channel capacity of the UWB TR systems.

References

1. F. Han, Y.-H. Yang, B. Wang, Y. Wu, L. K.J.R., Time-reversal division multiple access in multi-path channels, in *Proceeding of the Global Telecommunications Conference*, 2011, pp. 1–5
2. T.H. Vu, N.T. Hieu, H.D.T. Linh, N.T. Dung, L.V. Tuan, Channel capacity of multi user TR-MIMO-UWB communications system, in *Proceedings of the International Conference on Computing, Management and Telecommunications (ComManTel)*, 2013, pp. 22–26
3. H. Nguyen, Z. Zhao, F. Zheng, T. Kaiser, Preequalizer design for spatial multiplexing SIMO-UWB TR systems. IEEE Trans. Veh. Commun. **59**(8), 3798–3805 (2010)
4. D.-D. Tran, V. Tran-Ha, D.-B. Ha, Applying time-reversal technique for MU MIMO UWB communication systems, in *Proceedings of the World Congress on Engineering and Computer Science, WCECS 2013*, San Francisco, 23–25 Oct, 2013. Lecture Notes in Engineering and Computer Science, pp. 724–729
5. L.Y. Georgios, B. Giannakis, Ultra-wideband communications: an idea whose time has come. IEEE Sign. Process. Mag. **21**(6), 26–54 (2004)
6. H.T. Nguyen, I.Z. Kovacs, P.C.F. Eggers, A time reversal transmission approach for multiuser uwb communications. IEEE Trans. Antennas Propag. **54**(11), 3216–3224 (2006)
7. R.C. Qiu, A theory of time-reversed impulse multiple-input multiple-output (mimo) for ultra-wideband (uwb) communications, in *Proceedings of the IEEE 2006 International Conference on Ultra-Wideband*, 2006, pp. 587–592
8. T. Kaiser, F. Zheng, Ultra-wideband systems with MIMO. 1em plus 0.5em minus 0.4em (Wiley, Chichester, 2010)
9. M. Fink, Time reversal of ultrasonic fields. i. basic principles. IEEE Trans. Ultrason. Feroelectr. Freq. Control **39**(5), 555–566 (1992)
10. P. Derode, A. Roux, M. Fink, Robust acoustic time reversal with high-order multiple scattering. Phys. Rev. Lett. **75**(23), 4206–4209 (1995)
11. F. Han, Y.-H. Yang, B. Wang, Y. Wu, K. Liu, Time-reversal division multiple access over multi-path channels. IEEE Trans. Commun. **60**(7), 1953–1965 (2012)

Chapter 35
Efficient Bandwidth Allocation Methods in Upstream EPON

S. K. Sadon, N. M. Din, N. A. Radzi, Mashkuri Bin Yaacob, Martin Maier and M. H. Al-Mansoori

Abstract The Internet has become the world's leading universal global communication infrastructure. Optical solutions are sought after at the access network to support the ever increasing demand in bandwidth. Passive Optical Networks (PONs) are seen to provide a cost effective solution for this. PON Dynamic Bandwidth Allocation (DBA) scheme provides the means for upstream traffic allocation. In this paper, the operation of several bandwidth allocation algorithms in upstream Ethernet PON (EPON) is presented. An Efficient Distributed DBA (EDDBA) that supports Quality of Service (QoS) for both inter and intra ONU allocation is proposed. The proposed scheme introduces an identical DBA algorithm running simultaneously in each ONU. The simulation performance for the proposed DBA was conducted using Prolog and shows flexibility, reliability in handling data, voice, and video traffic.

Keywords DBA · Decentralized allocation · EPON · PROLOG · Qos · SBA

S. K. Sadon (✉) · N. M. Din · N. A. Radzi · M. B. Yaacob
Center for Communications Service Convergence Technologies, College of Engineering,
Universiti Tenaga Nasional, Jalan IKRAM-UNITEN, 43000 Kajang, Selangor, Malaysia
e-mail: saja.almola@yahoo.com

N. M. Din
e-mail: norashidah@uniten.edu.my

N. A. Radzi
e-mail: asyikin@uniten.edu.my

M. B. Yaacob
e-mail: mashkuri@uniten.edu.my

M. Maier
Optical Zeitgeist Laboratory, INRS, 800, Gauchetière West suite 6900, Montréal, QC H5A 1K6, Canada
e-mail: maier@emt.inrs.ca

M. H. Al-Mansoori
Faculy of Engineering, Sohar University, PO Box 44, PCI 311 Sohar, Oman
e-mail: mansoori@ieee.org

H. K. Kim et al. (eds.), *Transactions on Engineering Technologies*,
DOI: 10.1007/978-94-017-9115-1_35,

1 Introduction

The notable success of the Internet has created new challenges. There has been a recorded exceptional increase in data traffic since the last decade [1]. Though the internet was originally designed for data transfer to cater for specific needs of daily life, many of the current applications such as Internet protocol (IP) telephony, video conferencing and video broadcasting need to be supported with additional requirements [2].

The first Passive Optical Network (PON) standard from the International Telecommunication Union (ITU) is based on the asynchronous transfer mode (ATM), i.e. ITU-T G.983 known as APON (ATM PON) [3]. It is used largely for business applications, and earmarked for PON devoted to deliver voice and data services. Further improvements to the original APON standard led to the emergence of broadband PON (BPON).

BPON is as an extended version of the APON network with better transmission capabilities and extra services. BPON is able to support wavelength-division multiplexing (WDM), dynamic and advanced up-stream bandwidth allocation, as well as larger sustainability than APON. BPON is defined under the ITU-T G.983.2 and G.983.3 standards [4, 5]. Although the structures of both these PONs were flexible and adaptable to different scenarios, they did not gain much popularity due to the complexity of the ATM protocol. The advent of technology saw the introduction of gigabit PON (GPON) which was introduced in 2003 by the Full Service Access Network (FSAN) group. GPON provides ATM interconnections. GPON is a PON version that is able of distributing data traffic with a capability of gigabit-per-second bit-range. It is defined in ITU-T G.984. It supports higher rates, enhances security, and has a choice of Layer 2 protocols. The GPON transportation mechanism is referred to as GPON Encapsulation Method (GEM) [6, 7].

Ethernet technology was introduced in the 70s as a local area network (LAN) technology for interconnecting desktop computers [8]. It catered for high bandwidth, low operation cost, and simplicity of installation and usage. The media access control (MAC) protocol used by Ethernet is the carriers sense multiple accesses with collision detection (CSMA/CD) for local area networking [9]. Ethernet PON was established by the Institute of Electrical and Electronics Engineers, Inc. (IEEE) 802.3ah task force [10].

GPON and EPON standards have already been established and deployments are seen worldwide. The existing and next generation PON should be flexible in supporting multiple existing and emerging services across multiple market segments, such as consumer, business, and mobile backhaul and therefore requires appropriate traffic management mechanisms. Dynamic Bandwidth Allocation (DBA) can play an effective part in improving the network performance and optimising capacity in the Fiber-to-the-x (FTTx) networks. A PON DBA scheme provides the means for upstream traffic control by the OLT of the ONUs.

In EPON, the available bandwidth is broadcasted to each ONU from the OLT through a splitter or network of splitters during downstream transmission. In the case of the upstream direction, the available bandwidth is shared amongst all ONUs at the optical splitter using the same fibre. To prevent congestion and collisions between frames from different ONUs, an OLT assigns a non-specific and non-repeated time-slot to each ONU using time division multiplexing (TDMA) technology. This facilitates transmission without overlapping difficulties. There is a need to effectively assign time slots for ONUs [11].

This paper focuses on several efficient bandwidth allocation algorithms for upstream EPON. Section 2 explains the static and dynamic bandwidth allocation schemes. Section 3 elaborates on DBA examples of efficient dynamic bandwidth allocation scheme including a new decentralised scheme. Section 4 provides the conclusion.

2 Static and Dynamic Bandwidth Allocation

The support of Quality of Service (QoS) within an EPON domain is currently one of the main areas of research [11] as they are less structured than in GPON. The assignment of bandwidth to each ONU could be either fixed or variable. Fixed timeslot allocation, also known as Static Bandwidth Allocation (SBA) [12], is straight forward to implement. In the case of SBA, once the bandwidth is allocated to a network user, it becomes unavailable to other users on that network [13]. Given the unpredictable nature of the network traffic, there will be situations where some timeslots overflow even under a very light load and causes delay for some packets. On the other hand, other timeslots may not be totally used even during heavy traffic, resulting in underutilized usage of the upstream bandwidth. For this reason, these shortfalls in SBA are addressed by allocating the bandwidth dynamically.

In order to optimize bandwidth utilization, it is desirable that the OLT allocates a varying timeslot in a dynamic manner to each ONU based on the actual demand of the ONUs. The DBA can be strategized to provide the necessary QoS as demanded by applications. Various DBA algorithms have been established to report this aspect [14–17]. Figure 1 depicts the DBA research areas. The DBA schemes are divided into two categories: centralized and decentralized where both may comprise of single level inter-ONU allocation or hierarchical with inter- and intra-ONU allocations. Though DBA algorithms have been reported in previous years [18–23], these algorithms have difficulty in estimating the proper allocation of the bandwidth to each ONU, especially for the discontinuous data traffic such as IP traffic [18]. In addition, the QoS and the satisfaction of subscribers are still open issues.

The first and largest category of bandwidth allocation scheme is the centralized bandwidth allocation approach. In this scheme the reports travel to the OLT from the ONUs to inform about their bandwidth needs for upstream bandwidth admission. The decision controller is located at the OLT. This means that a bandwidth decision controller located in the OLT would receive information from the ONUs and dynamically and individually schedules each queue located in the multiple ONUs.

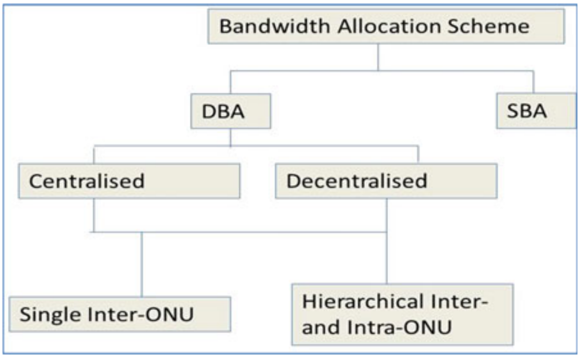

Bandwidth Allocation Scheme
DBA
SBA
Centralised
Decentralised
Single Inter-ONU
Hierarchical Inter- and Intra-ONU

Fig. 1 Bandwidth assignment research categorization

However, any software failure in the OLT will halt the ONU upstream transmission [23] adding additional delay if the bandwidth allocations failed in one cycle and the ONUs have to wait several more cycles to get grants. Centralized scheme could be either single-level or hierarchical. The classical and first example of centralized allocation with single-level algorithm is Interleaved Polling with Adaptive Cycle Time (IPACT) scheme by Kramer et al. [24] and it was the pioneering algorithm produced for DBA. In this algorithm, the OLT polls each of the ONUs individually in a round-robin fashion to dynamically allocate the bandwidth in accordance with the requested bandwidth of each ONU.

The hierarchical scheme involved algorithms at the OLT for inter and intra ONU bandwidth allocation. The hierarchical allocation allows expansion and efficient manage of bandwidth's resources besides resolving the need of separating GATE and REPORT messages, so they can be sent together to each queue over an EPON system [25]. Numerous hierarchical DBA algorithms are illustrated in the references [26, 27]. Assi et al. [28] upgraded the initial interleaved polling with adaptive cycle time (IPACT) algorithm by making more efficient usage of the under loaded unwanted excessive bandwidth. The excess bandwidth is effectively utilized by evenly assigning it amongst the highly loaded ONUs. However, this DBA may cause wasted bandwidth since the highly loaded ONUs can receive more bandwidth than requested. Hence, Bai et al. [29] improved the algorithm by adding a weight-based DBA scheme that is called weighted-DBA. The excess bandwidth from classes that need less than their maximums in lightly loaded ONUs is distributed among the classes that need more than their maximums in highly loaded ONUs.

Lately, a few researches dealt with the decentralized bandwidth allocation approaches for the EPON, as EPON's bandwidth allocation schemes were relying on centralized architecture. OLT is regarded as the only intelligent device that could arbitrate time-division access to the shared channel. However, any failure in the OLT bandwidth allocation would halt the allotment for ONUs. Nevertheless, most of the solutions for decentralised bandwidth allocation included additional expenses and complexity to the original architecture proposed by IEEE 802.3ah. The original EPON architecture had to be modified, such that each ONU's upstream transmission is echoed at the splitter to all ONUs. Each ONU is equipped with an additional receiver to obtain the copied transmission. As such, the entire ONUs is able to monitor the transmission of each ONU and determine the upstream channel access.in a dispersed manner [30].

Sherif et al. [31], proposed an Ethernet over a star coupler-based PON architecture that uses a distributed TDMA arbitration scheme. The paper anticipated several QoS-based DBA algorithms in which the OLT is excluded from the implementation of the time slot assignment. Wong [32], simulated a network frame in a Local Area Network (LAN) and all the ONUs were connected with one star coupler by two fibres. The ONUs was able to accept the frames transmitted by the other ONUs.

Delowar et al. [33] conducts a study of a distributed DBA scheme over a decentralized architecture and explored the control plane feasibility of such architecture. However, the proposed decentralized architecture saw increased complexity and cost of the ONUs. Also, Feng et al. [34] published an article analysing a distributed DBA by using simulation. The proposed algorithm does not alter the existing EPON architecture, but only transferred the requested bandwidth information transmitted to all ONUs by using the broadcasting capability of the OLT. It was completely in congruence with the Multi Point Control Protocol (MPCP), as regulated by the IEEE 802.3ah standard and only needed to add some new fields in the REPORT and GATE frames. However, this algorithm did not support traffic classes, e.g. Differentiated Services (Diffserv), and could lead to an inefficient QoS.

A new decentralised scheme for the emerging long reach (LR)-PON was proposed by Helmy et al. [35], where the PON network range is extended to over than 100 km where the propagation delays will increase and severely affect the performance of the algorithms as they are based on the bandwidth. However, this technique requires additional ONU transceivers.

3 DBA for Upstream EPON Examples

Several bandwidth allocation scheme proposed previously by researchers have been reviewed in the introduction. A discussion of efficient bandwidth allocation works by the authors is elaborated in this section.

3.1 Centralized Allocation: Single Level

The global priority DBA is a centralized DBA solution proposed by the authors in [36] as shown in Fig. 2. In global priority DBA, three queues are inside the ONU, OLT and the user's side, consistent with the needs of supporting QoS based on DiffServ. The three types of services are voice-DiffServ Expedited Forwarding (EF), video-DiffServ Assured Forwarding (AF), and data—DiffServ Best Effort (BE) respectively. The ultimate highest priority traffic is the voice bandwidth

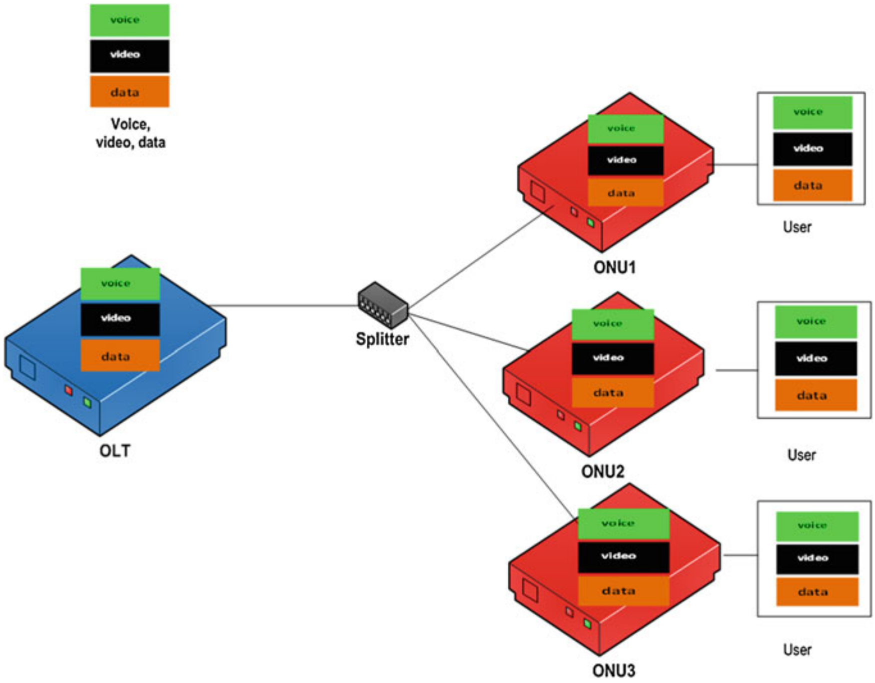


Fig. 2 DBA with global priority

because it requires strict delay and jitter guarantee. The medium priority traffic is the video bandwidth that requires bandwidth assurances. Finally, the lowest priority traffic is the data bandwidth that is more concerned with throughput.

The amount of excess bandwidth for lightly loaded queues is calculated and then allocated to the highly loaded queues according to their priority. Universally the needs of higher priority traffic would be met first, followed by the medium and low priority traffic. Thus, this method ensures the QoS support in the EPON system.

3.2 Hierarchical Allocation

3.2.1 Russian Doll Model (RDM) Algorithm

The DiffServ Aware Traffic Engineering (DS-TE) architecture provides the recommendation for the establishments of how the bandwidth could be allocated to different traffic classes in order for network suppliers to adopt the wide range of services, such as voice-over-IP (VoIP), IPTV, and video on demand (VoD) [37]. The bandwidth constraints (BC) model considers one of the most essential aspects

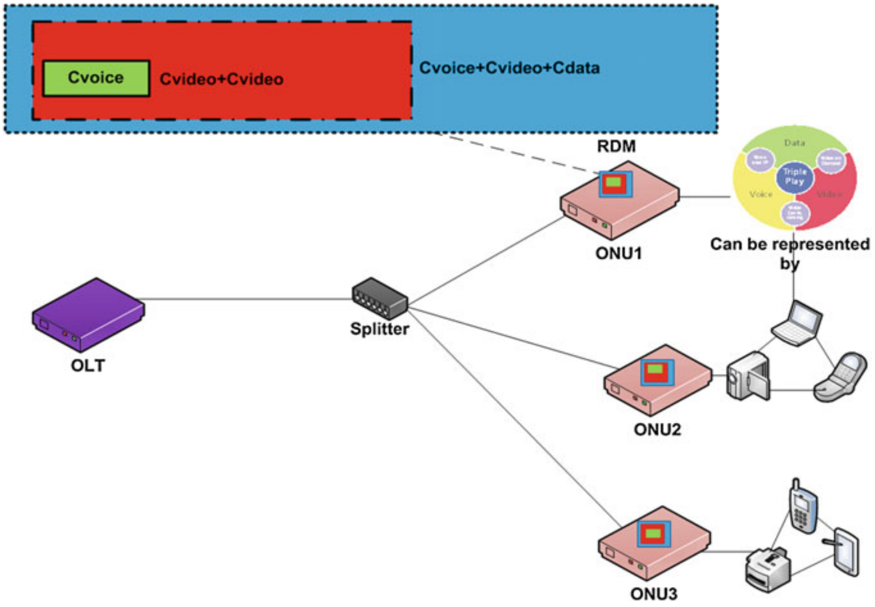


Fig. 3 DBA with RDM

of the DiffServ. The RDM [12, 13] is a technique that could be used to guarantee bandwidth efficiency and QoS of many types of services. It could also be used to simultaneously achieve isolation across the three types of traffic classes, so that each class type is guaranteed its share of bandwidth and bandwidth efficiency, with prevention of QoS degradation for all traffic types. The RDM improves bandwidth efficiency by allowing the triple play services to share the bandwidth, whereby the lower priority class are able to use the available excess bandwidth from a higher priority class of up to the summation of their bandwidth constraint values.

The hierarchical RDM architecture is shown in Fig. 3, C_{voice} is the traffic class with the highest strictest QoS requirements, C_{video} is the medium priority, while C_{data} is the best effort traffic class. The algorithm provides the allocation of bandwidth between the optical ONUs and within the ONUs based on RDM rules. The RDM bandwidth constraints are as expressed in Table 1.

The hierarchical dynamic bandwidth allocation algorithm based on RDM which is referred as the Russian Doll Dynamic Bandwidth Allocation (RDDBA) was able to distribute the bandwidth in an effective way for inter- and intra-ONU in an EPON. The allocation of bandwidth is strategized according to the ordering and prioritization of triple play services. The simulation results indicated that the proposed algorithm addresses the perfect differentiated services by balancing priority and fairness by supporting the triple-play services rather well, i.e. video, voice, and data, as well as makes effective adjustment in balancing bandwidth sharing between the ONUs.

Table 1 RDM rules

Symbol	Description	Amount	RDM rules
$C_{voice}(I)$	Aggregated bandwidth for voice	20 % of total bandwidth	$BC2 = C_{voice}(I)$
$C_{video}(I)$	Aggregated bandwidth for video	50 % of total bandwidth	$BC1 = C_{voice}(I) + C_{video}(I)$
$C_{data}(I)$	Aggregated bandwidth for data	30 %of total bandwidth	$BC0 = C_{voice}(I) + C_{video}(I) + C_{data}(I)$

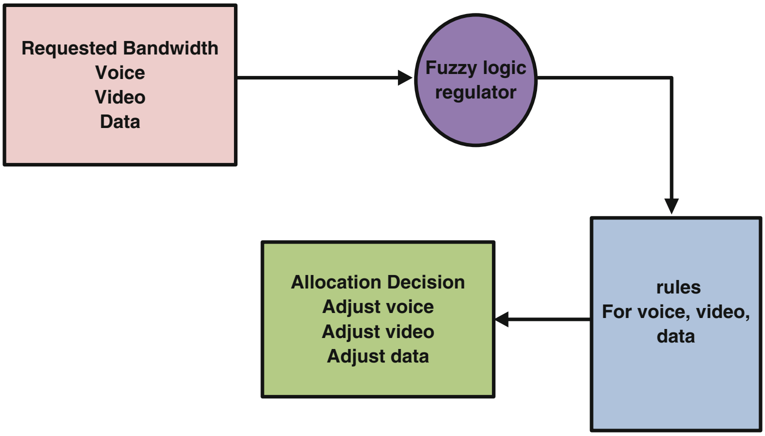


Fig. 4 DBA based fuzzy logic

3.2.2 Fuzzy Logic Algorithm

A hierarchical fuzzy logic was also incorporated in the intra-ONU allocation as shown in Fig. 4. The bandwidth allotment for each ONU in this algorithm is done accordingly by using fuzzy logic. The service discipline bandwidth that is assigned first is voice traffic. Then, the remaining bandwidth is allocated to video and finally to the data bandwidth.

The fuzzy logic regulator mechanism involves three input parameters implied by requested bandwidth for voice, video and data, and one output parameter. The output decision used to allocate bandwidth in each ONU is triggered when there is conflict for bandwidth between the classes [37, 38].

The hierarchical fuzzy logic scheme for EPON is called intelligent fuzzy logic-based dynamic bandwidth allocation algorithm (IFLDBA) [37]. The algorithm offers an intelligent decision making scheme for the allotment of bandwidth between the ONUs and within the ONUs. Fuzzy logic is used to improve the bandwidth allocation for inter and intra ONU scheduling. The results obtained showed that the IFLDBA has a better bandwidth utilization of up to 21 % and lower delay than the earlier introduced classical algorithm called the broadcast polling algorithm.

Table 2 The fuzzy logic rules

Rule	rvoice	rvideo	rdata	Decision
1	Low	Low	High	Adjust data
2	Low	High	Low	Adjust video
3	Low	High	High	Adjust video
4	High	Low	Low	Adjust voice
5	High	Low	High	Adjust EF
6	High	High	Low	Adjust EF
7	High	High	High	Adjust EF

The fuzzy decision rules for bandwidth allocation decision are shown in Table 2. There are seven rules that are related to the three inputs with the fuzzy output.

The fuzzy output range and the resultant allocation decision are based on the Sugeno inference method. It is noted that the delay for the IFLDBA increased linearly as the offered load increases to 1 Gbps with bandwidth limitation is set to avoid voice monopoly for the EF traffic class. The AF traffic class delay is seen be reduced as the offered load increases to 1 Gbps [37].

3.3 Decentralized Allocation

The flow chart of the proposed Efficient Distributed DBA (EDDBA) is depicted on Fig. 5. EDDBA is based on the existing EPON structure with no modification [39, 40] and only transfer the information window transmitted by one ONU to the other ONUs by broadcasting, similar to the OLT functionality. In the downstream direction, the OLT will insert a “broadcast” link ID which will be accepted by every ONU which complies with the MPCP requirements as outlined in the IEEE 802.3ah standard. The OLT will then add some fields in the REPORT frame and the GATE frame. To avoid frame duplication when an ONU receives its own frame, the ONU accepts a frame only if the frame’s link ID is different from the link ID assigned to that ONU. All information will be kept in a look- up table to be used later in bandwidth allocation decision making.

The decentralized algorithm will run simultaneously at each ONU, and the requested bandwidth for each ONU is then calculated.

If the requested bandwidth is less than the allowed bandwidth then the bandwidth given is the requested bandwidth based on defined criteria where the voice will get up to 20 % of the available bandwidth with the highest priority, the video’s portion is 40 % with medium priority and the data’s share is 40 % with lowest priority.

If the total demanded bandwidth is higher than the available bandwidth then excessive bandwidth from other ONUs will be considered as in equations below in the same manner as the algorithm in [28]. Then the extra bandwidth will be distributed between the ONUs.

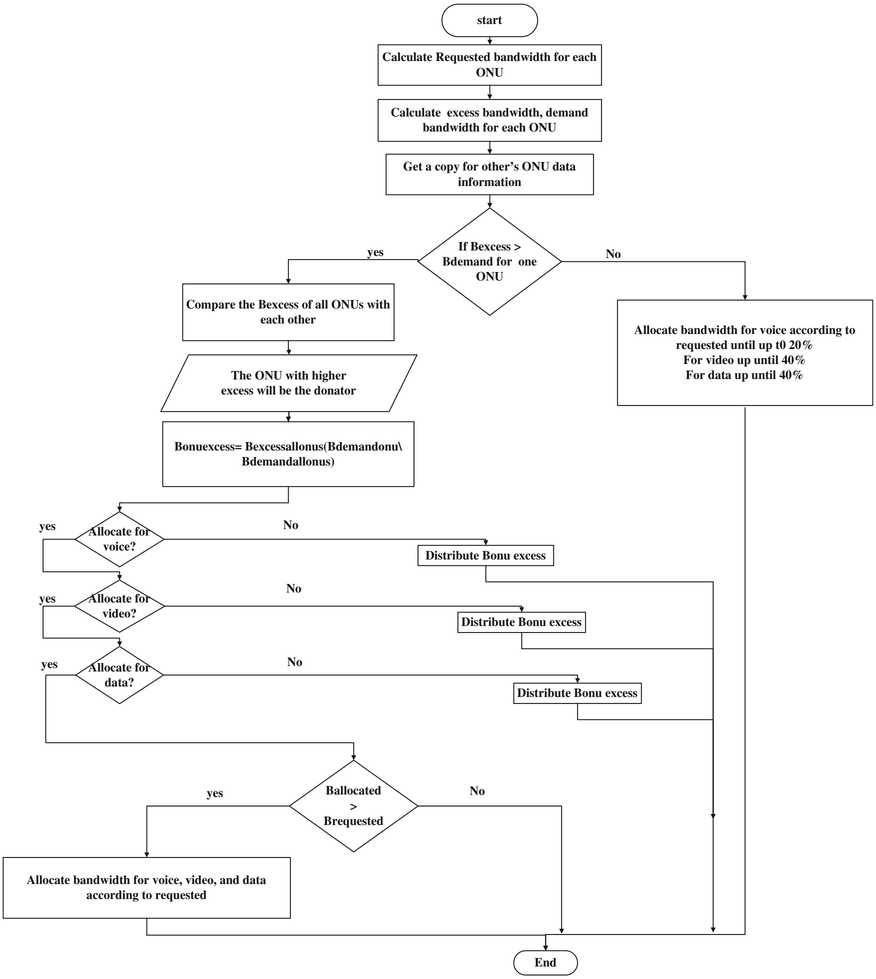

start
Calculate Requested bandwidth for each ONU
Calculate excess bandwidth, demand bandwidth for each ONU
Get a copy for other's ONU data information
If Bexcess > Bdemand for one ONU
yes
No
Compare the Bexcess of all ONUs with each other
The ONU with higher excess will be the donator
Bonuexcess= Bexcessallonus(Bdemandonu\ Bdemandallonus)
Allocate bandwidth for voice according to requested until up t0 20%
For video up until 40%
For data up until 40%
yes
Allocate for voice?
No
Distribute Bonu excess
yes
Allocate for video?
No
Distribute Bonu excess
yes
Allocate for data?
No
Distribute Bonu excess
yes
Ballocated > Brequested
No
Allocate bandwidth for voice, video, and data according to requested
End

Fig. 5 Flow chart of the efficient distributed DBA

To ensure the efficient utilization of excess bandwidth, the excessive bandwidth should be shared according to the requested bandwidth within the overloaded group. The performance model comprise of ONUs that maintain three separate priority queues which share the same buffering space to allow straight forward mapping of DiffServ's expedited forwarding (EF), assured forwarding (AF), and best effort (BE) classes [31].

A simulation study for EDDBA was performed as in [41] with the traffic scenarios of considering three priority classes P0, P1, and P2, with P0 being the highest priority constant-bit-rate (CBR) used for delivering voice, and P1 for variable-bit-rate or (VBR) traffic, which represents the video stream, and P2 the best-effort data. The simulation study was conducted using the PROLOG simulation [42].

The starting point for our comparative study is to look at how the offered load affects the bandwidth utilization of the IPACT [24] and EDDBA. An efficient decentralized DBA algorithm strives to achieve as high bandwidth utilization as possible. The improvement of the bandwidth utilization for EDDBA shows a useful performance via PROLOG simulation in comparison to the IPACT algorithm due to the efficient reuse of unclaimed access bandwidth.

4 Conclusion

This paper provides the background on DBA methods in upstream EPON systems. The operation of several bandwidth allocation algorithms in upstream EPON has been highlighted for centralised (single level and hierarchical) and decentralised DBA schemes. A new decentralised DBA for EPON, i.e. EDDBA has been proposed and deliberated in this paper.

Acknowledgment We would like to acknowledge and thank the Ministry of Science, Technology and Innovation Malaysia for funding this project under the Escience Fund 01-02-03-SF0247.

References

1. K.G. Coman, in *Handbook of Massive Data Sets*. Internet Growth: Is There a “Moore’s Law” for Data Traffic (Kluwer Academic, Dordrecht, 2001), pp. 47–93
2. N.M. Din, Fuzzy logic traffic control for differentiated service-aware generalized multi-protocol label switching network. Ph.D. thesis, Universiti Teknologi Malaysia, July 2007
3. ITU-T Recommendation G.983.1, *Broadband optical access systems based on Passive Optical Networks (PON)*. (1998), pp. 1–110
4. ITU-T Recommendation G.983.2, *ONT management and control interface specification for B-PON*. (2005), pp. 1–359
5. ITU-T Recommendation G.983.3, *A broadband optical access system with increased service capability by wavelength allocation*. (2001), pp. 1–51
6. ITU-T Recommendation G.983.4, *A broadband optical access system with increased service capability using dynamic bandwidth assignment*. (2001), pp. 1–82
7. I. Cale, A. Salihovic, M. Ivekovic, Gigabit passive optical network: GPON.2007, in *Proceedings of the 29th International Conference on Information Technology Interfaces, ITI* 2007, pp. 679–684
8. C.F. Lam, *Passive Optical Networks* (Academic Press, Burlington, 2007)
9. J. Majithia, W. Zhang, Performance results of CSMA/CD ethernet with various acknowledgement schemes, in *Proceedings of the International Conference of IEEE Region 10 Technology Enabling Tomorrow: Computers, Communications and Automation towards the 21st Century, TENCON ‘92*, Nov 1992, pp. 6–10
10. IEEE Standards, IEEE P802.3ah ethernet in the first mile task force (2004), http://www.ieee802.org/3/efm/public/comments/
11. C. Assi, Y. Ye, S. Dixit, M. Ali, Dynamic bandwidth allocation for quality-of-service over ethernet PONs. IEEE J. Sel. Areas Commun. **21**(9), 1467–1477 (2003)

12. F. Le Faucher, Russian dolls bandwidth constraints model for diffserv-aware MPLS traffic engineering. (The Internet Society, RFC 4127, 2005), http://community.roxen.com/developers/idocs/rfc/rfc4127.html
13. S.K. Sadon, N.M. Din, M.H. Al-Mansoori, N.A. Radzi, M. Mustafa, M.S.A. Majid, Dynamic hierarchical bandwidth allocation using Russian doll model in EPON. J. Comput. Electr. Eng. **38**, 1480–1489 (2012)
14. B. Skubic, J. Chen, J. Ahmed, L. Wosinska, and B. Mukherjee, A comparison of dynamic bandwidth allocation for EPON, GPON, and next-generation TDM PON. IEEE Communications Magazine (2009), pp. 540–548
15. G. Kramer, B. Mukherjee, G. Pesavento, Interleaved polling with adaptive cycle time (IPACT): a dynamic bandwidth distribution scheme in an optical access network. Photonic Netw. Commun. **4**(1), 89–107 (2002)
16. IEEE Standards, IEEE 802.1D local and metropolitan area networks: media access control (MAC) bridges (2004). http://www.dcs.gla.ac.uk/~lewis/teaching/802.1D-2004.pdf
17. M. McGarry, M. Maier, M. Reisslein, Ethernet PONs: a survey of dynamic bandwidth allocation (DBA) algorithms. IEEE Commun. Mag. **42**(8), 8–15 (2004)
18. H.J. Byun, J.M. Nho, J.T. Lim, Dynamic bandwidth allocation algorithm in ethernet passive optical networks. Electr. Lett. **39**(13), 1001–1002 (2003)
19. S.I. Choi, J.D. Huh; Dynamic bandwidth allocation algorithm for multimedia services over ethernet PONs. ETRI J. **24**(6), 465–468 (2002)
20. J. Xie, S. Jiang, Y. Jiang, A dynamic bandwidth allocation scheme for differentiated services in EPONs IEEE Commun.Mag. **42**(8), 32–39 (2004)
21. K.H. Ahn, K.E. Han, and Y.C. Kim, Hierarchical dynamic bandwidth allocation algorithm for multimedia services over ethernet PONs, ETRI J. **26**, 321–333 (2004)
22. Y. Zhu, M. Ma, T. Hiang Cheng, An efficient solution for mitigating light-load penalty in EPONs. Comput. Electr. Eng. **32**(6), 426–431, 2006
23. ITU-T Recommendation G.983.5, *A broadband optical access system with enhanced survivability* (2002), pp. 1–51
24. G. Kramer, B. Mukherjee, G. Pesavento, Interleaved polling with adaptive cycle time (IPACT): a dynamic bandwidth distribution scheme in an optical access network. Photonic Netw. Commun. **4**(1), 89–107 (2002)
25. IEEE Standards, IEEE 802.1D local and metropolitan area networks: media access control (MAC) bridges (2004). http://www.dcs.gla.ac.uk/~lewis/teaching/802.1D-2004.pdf
26. J. Zheng, H.T. Mouftah, A survey of dynamic bandwidth allocation algorithms for ethernet passive optical networks. J. Opt. Switching Netw. **6**, 151–162 (2009)
27. M. McGarry, M. Maier, M. Reisslein, Ethernet PONs: a survey of dynamic bandwidth allocation (DBA) algorithms. IEEE Commun. Mag. **42**(8), 8–15 (2004)
28. C. Assi, Y. Ye, S. Dixit, M. Ali, Dynamic bandwidth allocation for quality-of-service over Ethernet PONs. IEEE J. Sel. Areas Commun. **21**, 1467–1477 (2003)
29. X. Bai, A. Shami, C. Assi, On the fairness of dynamic bandwidth allocation schemes in ethernet passive optical networks. Comput. Commun. **29**(11), 2123–2135 (2006)
30. M. Maier, N. Ghazisaidi, FiWi Access Networks (Cambridge university press, Cambridge, 2012). www.cambridge.org
31. S.R. Sherif, A. Hadji Antonis, G. Ellinas, C. Assi, M.A. Ali, A novel decentralized ethernet-based PON access architecture for provisioning differentiated Q0S. J. Light wave Technol. **22**, 2483–2497 (2004)
32. E. Wong, C. Chang-Joon, Efficient dynamic bandwidth allocation based on upstream broadcast in ethernet passive optical networks. Opt. Fiber Commun. Conf. **6**, 3 (2005)
33. A.S.M. Delowar Hossain, H. Erkan, M.A. Ali, A distributed control plane architecture for EPON, in *Proceedings of the 2nd International Conference on Innovations in Information Technology*, 2005
34. F. Cao, D. Liu, M.Z. Kang Yang, L.D YinboQian, A distributed dynamic bandwidth allocation algorithm in EPON. Mod. Appl. Sci. **4**, 20–24 (2010)

35. A. H. Helmy, Habib Fathallah, and Adel Abdennour, Decentralized media access versus credit-based centralized bandwidth allocation for LR-PONs, in *Proceedings of the High Capacity Optical Networks and Enabling Technologies (HONET)*, Dec 2011, pp. 329–333
36. D. Nikolova, B.V. Houdt, C. Blondia, QoS issues in EPON, in *Proceedings of the Community Nets and FTTH/P/x Workshop*, 2003
37. N.A.M. Radzi, N.M. Din, M.H. Al-Mansoori, I.S. Mustafa, S.K. Sadon, Intelligent dynamic bandwidth allocation algorithm in upstream EPONs. J. Opt. Commun. Netw. **2**(3), 148–158
38. N.A.M. Radzi, N.M. Din, M.S.A. Majid, M.H. Al-Mansoori, Global priority DBA using fuzzy logic in ethernet PON, in *Proceedings of the 17th Asia Pacific Communications Conference (APCC)*, 2011, pp. 113–116
39. B. Lung, PON architecture 'future proofs' FTTH. Lightwave Mag. **16**(10), 104–107 (1999)
40. D. Sala, A. Gummalla, PON functional requirements: services and performance, in *Proceedings of the IEEE 802.3ah Meeting in Portland OR*, July 2001
41. S.K. Sadon, N.M. Din, N.A. Radzi, M.H. Al-Mansoori, Efficient decentralized dynamic bandwidth allocation in EPON, in *Proceedings of the World Congress on Engineering and Computer Science, WCECS 2013*. Lecture Notes in Engineering and Computer Science. San Francisco, 23–25 Oct 2013, pp. 746–749
42. Logic Programming Associate, Available: http://www.lpa.co.uk/win_det.htm

Chapter 36
Soft Tissue Characterisation Using a Force Feedback-Enabled Instrument for Robotic Assisted Minimally Invasive Surgery Systems

Mohsen Moradi Dalvand, Bijan Shirinzadeh, Saeid Nahavandi, Fatemeh Karimirad and Julian Smith

Abstract An automated laparoscopic instrument capable of non-invasive measurement of tip/tissue interaction forces for direct application in robotic assisted minimally invasive surgery systems is introduced in this chapter. It has the capability to measure normal grasping forces as well as lateral interaction forces without any sensor mounted on the tip jaws. Further to non-invasive actuation of the tip, the proposed instrument is also able to change the grasping direction during surgical operation. Modular design of the instrument allows conversion between surgical modalities (e.g., grasping, cutting, and dissecting). The main focus of this paper is on evaluation of the grasping force capability of the proposed instrument. The mathematical formulation of fenestrated insert is presented and its non-linear behaviour is studied. In order to measure the stiffness of soft tissues, a device was developed that is also described in this chapter. Tissue characterisation experiments were conducted and results are presented and analysed here. The experimental results verify the capability of the proposed instrument in accurately measuring grasping forces and in characterising artificial tissue samples of varying stiffness.

Keywords Actuation mechanism · Force measurement · Laparoscopic instrument · Modularity · Power transmission mechanism · Robotic assisted minimally invasive surgery (RAMIS) · Strain gages

M. M. Dalvand (✉) · S. Nahavandi · F. Karimirad
Centre for Intelligent Systems Research (CISR), Deakin University, Waurn Ponds Campus,
Melbourne, VIC 3216, Australia
e-mail: mohsen.m.dalvand@deakin.edu.au

B. Shirinzadeh
Robotics and Mechatronics Research Laboratory (RMRL), Department of Mechanical
and Aerospace Engineering, Monash University, Melbourne, Australia

J. Smith
Department of Surgery, Monash Medical Centre, Monash University, Melbourne, Australia

H. K. Kim et al. (eds.), *Transactions on Engineering Technologies*,
DOI: 10.1007/978-94-017-9115-1_36,

1 Introduction

Having improved the disadvantages of the traditional laparoscopic surgery, Robotic Assisted Minimally Invasive Surgery (RAMIS) has negatively affected the surgeon's ability in palpating and diagnosing soft tissues of varying stiffness during surgery [1, 2]. The lack of force feedback has motivated several researchers to explore possible methods of restoring this feature to RAMIS by making laparoscopic instruments capable of measuring tip/tissue interaction forces. Strain gauges were applied on the tip and handle of laparoscopic surgical forceps to characterise tissues [3–6]. Retrofitting of laparoscopic forceps with a commercial six-axis force/torque sensor encapsulated in the instrument shaft [7] and a force sensor on the handle of the tool [8] were studied. A two Degree Of Freedom (DOF) force sensing sleeve for 5 mm laparoscopic instruments was developed with advantages of compatibility and modularity among several types of surgical instruments [9].

A micro-machined piezoelectric tactile sensor embedded beneath silicon teeth of the grasper jaws was developed that has some disadvantages including ability to measure only dynamic forces, susceptibility to damage from shear forces, non-sterilizability, and high cost [10]. At more complex level, a distal force/torque sensor for laparoscopic instruments was developed that uses Stewart platform to locate six strain gauges for measuring forces and torques along all of its six measurement axes [11, 12]. Preliminary results on design and fabrication of a cutting tool with an integrated tri-axial force sensor to be applied in fetal surgery procedures were reported [13, 14]. A miniature uniaxial force sensor for use within a beating heart during mitral valve annuloplasty was presented [15].

Besides retrofitting conventional laparoscopic instruments, research efforts have also been conducted in developing automated laparoscopic tools with force measurement capability [16, 17]. Most of these tools incorporate the advantage of actuation mechanisms for utilising in robotic surgical systems [18, 19]. Further to the force sensing laparoscopic instruments, robotic surgical systems were proposed with force measurement capabilities including Black Falcon [20] and Blue-DRAGON [21]. The possibility of using a trocar sensor for measurement of the surgical forces was also investigated in a robotic surgical system [22]. Although these research efforts were steps toward introducing force-feedback in robotic assisted surgical systems, there still exist many problems within the designs of surgical tools and their use in robotic surgery including modularity, size and force measurement issues.

In this chapter, an automated modular laparoscopic instrument is introduced that provides tool/tissue interaction force measurement capability directly from the surgery site [2, 23]. The proposed instrument is presented in Fig. 1 and different components of it are highlighted in this figure. The proposed instrument is incorporated with a micropositioning parallel manipulator and the RCM control algorithms has already been developed and evaluated [24–26]. The rest of the chapter is organized into three sections. In the next section, the modelling and

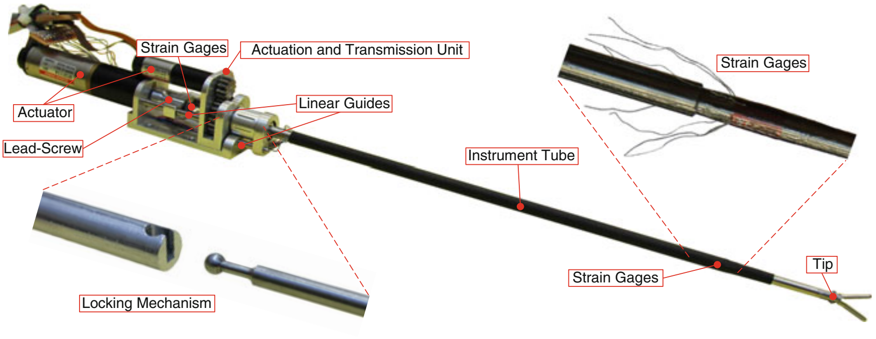


Fig. 1 Proposed force feedback-enabled minimally invasive surgery instrument

development of the proposed instrument are described that cover modularity feature, force sensing capability and modelling of tool tip. In Sect. 3, a device developed to measure stiffness of soft tissues is described. Experimental results are also presented in this section to verify the capabilities of the proposed instrument in probing and characterising soft tissues. Concluding remarks are made in Sect. 4.

2 Modelling and Development

2.1 Interchangeability Feature

Interchangeability feature of the proposed instrument enables the operator to easily and quickly change between variety of laparoscopic insert types e.g., grasping, cutting, and dissecting without loss of the force sensing capability. A stainless steel tube with 7 mm in diameter and 33 cm in length was designed in the way that it can be fitted concentrically, like a sleeve, over any 5 mm insert type manufactured by Matrix Surgical company. The nut (Fig. 2) couples the sleeve and the insert assembled to it to the base module. To minimise any potential error caused by unwanted clearance in the assembly, a part called nut-base was designed that is attached to the nut using a ball-bearing. Fastening the nut slides two supportive guides (Fig. 2) incorporated to the base inside two holes of the nut-base supporting the long tube of the instrument (Fig. 2). By help of this nut, insert, long tube and base module can be quickly disassembled and the insert type can be converted to the variety of insert types available.

Two actuators from Maxon Motor company were employed to actuate two DOFs related to the tip direction and operation (Fig. 2). Transmission and conversion of power required to change grasping direction and to operate tip jaws were achieved by spur gears and a lead-screw mechanism (Fig. 2). The size of the screw used in the lead-screw mechanism is $M6$ with 1 mm pitch. The mechanism to lock the insert to the lead-screw is also shown in Fig. 1.

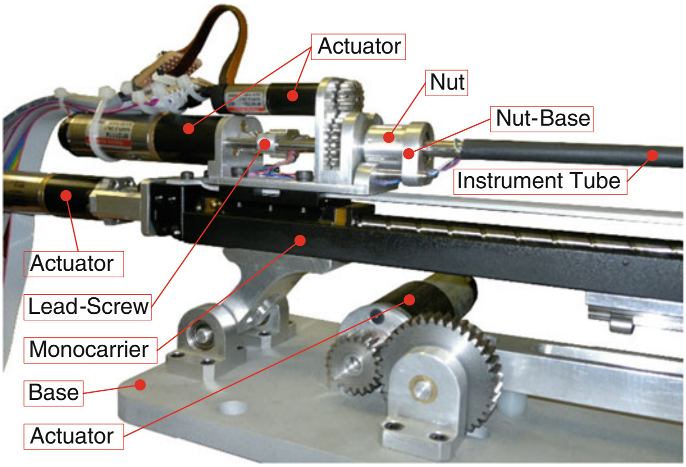


Fig. 2 The proposed instrument incorporated with a four-bar mechanism in a robotic assisted surgery system [25]

2.2 Force Measurement

In the proposed instrument, surface strains of the lead of the lead-screw mechanism is measured as a quantity for the normal grasping forces applied to the soft tissue at tip jaws. Two strain gages were embedded in the lead-screw mechanism (Fig. 1) to measure tension and compression strains in the lead of the lead-screw representing the normal grasping forces. Calibrations of the strain gage configurations were performed using known masses.

2.3 Kinematics of Grasping Mechanism

Pivotal mechanism are utilised at the tip jaws of most of the laparoscopic surgical instruments. The pivotal motion is commonly generated by linear displacement of the push rod and a mechanism to convert it to rotary displacement. In the proposed instrument the required linear movement is produced by a rotary actuator coupled with a lead-screw mechanism. Figure 3 describes the kinematics parameters of the tip mechanism for two different states where the tip is closed and tip jaws are at an arbitrary angle θ. In order to control the actuated laparoscopic instrument, it is necessary to obtain the relationship between the angular position of jaws (θ) and angular rotation of the gearbox shaft (ϕ). Considering Fig. 3, the following equations are derived:

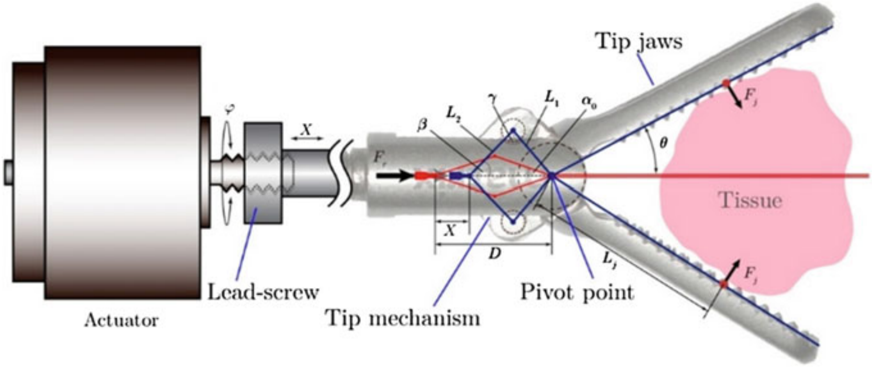


Fig. 3 Kinematics parameters of the tip mechanism for two different states of tip jaws

$$\alpha = \alpha_0 + \theta \tag{1}$$

$$L_1 \sin(\alpha) - L_2 \sin(\beta) = 0 \tag{2}$$

$$L_1 \cos(\alpha) + L_2 \cos(\beta) = D - X \tag{3}$$

where L_1 and L_2 are the constant geometrical parameters of the tip mechanism, α, β, and X are the variables of the mechanism and α_0 and D are the variables of the mechanism where the tip is at closed state $(\theta = 0)$ as described in Fig. 3. Combining Eqs. (2) and (3) yields:

$$\alpha + \beta = A \tag{4}$$

where A is as follows:

$$A = Arc\cos\left(\frac{(D - X)^2 - L_1^2 - L_2^2}{2L_1 L_2}\right) \tag{5}$$

Moreover, using Eqs. (2) and (4), α can be determined as follows:

$$\alpha = Arc\sin\left(\sqrt{\frac{L_2^2 \sin^2(A)}{L_2^2 \sin^2(A) + (L_1 + L_2 \cos(A))^2}}\right) \tag{6}$$

Using Eqs. (1), (5) and (6), and considering $X = {}^{\phi}\!/_{2\pi}$ for any angular position of the gearbox shaft (ϕ), the jaw's angular position is obtained as follows:

$$\theta = Arc\sin\left(\frac{\sqrt{\left[4L_1^2 L_2^2 - \left(\left(D - \frac{\phi}{2\pi}\right)^2 - L_1^2 - L_2^2\right)^2\right]}}{4L_1\left(D - \frac{\phi}{2\pi}\right)}\right) - \alpha_0 \tag{7}$$

To be able to control tip jaws to any specific angular position, the angular position of the gearbox shaft (ϕ) need to be derived as a function of the tip angular position (θ). This mathematical relation is derived as follows:

$$\phi = 2\pi(\sqrt{L_2^2 - L_1^2 \sin(\alpha_0 + \theta)^2} - L_1 \cos(\alpha_0 + \theta) - D) \tag{8}$$

According to the nonlinear relationship between θ and ϕ, the response of the tip with relatively closed jaws to the actuator shaft displacements is faster than that of the tip with relatively opened jaws. Strain gages applied to the lead rod results in the measurement of the tension and compression forces of the rod that is marked by F_r in Fig. 3. In order to determine the actual forces applied to the tissue at the tip jaws (F_j), the force propagation of the tip mechanism (the relation between F_r and F_j) should be defined.

It is worth noting that for the sake of modelling of the force propagation, the force applied to the tissue is assumed to act at a point lying at a distance L_j from the pivot point of the tip [12, 27, 28]. Neglecting the friction and balancing the forces and the moments around the pivot point leads to the derivation of the force propagation model as follows:

$$F_j = \frac{L_1}{2L_j} \times \frac{\sin(\gamma)}{\cos(\beta)} \times F_r \tag{9}$$

where γ is the angular variable of the mechanism (Fig. 3) and variables β and γ are defined using the following equations:

$$\beta = arc\sin(\frac{L_1 \sin(\theta + \alpha_0)}{L_2}) \tag{10}$$

$$\gamma = \pi - (\theta + \alpha_0) - \beta \tag{11}$$

The ratio of the forces in the push rod and at tip jaws (F_r/F_j) with respect to jaws angular positions indicates that applying a force to the push rod generates greater force at the jaws with relatively small angular position in comparison with the force that it generates at the jaws with relatively large angular position. Closing the tip jaws increases the normal forces of the jaws with the rate higher than that of the decrease in the angular position of the jaws.

The nonlinear relationship between θ and ϕ and also F_r/F_j for the allowable range of movements of the tip jaws is plotted in Fig. 4. As this figure indicates, applying a force to the push rod generates greater force at the jaws with relatively small angular position in comparison with the force that it generates at the jaws with relatively large angular position. Closing the tip jaws increases the normal forces of the jaws with the rate higher than that of the decrease in the angular position of the jaws.

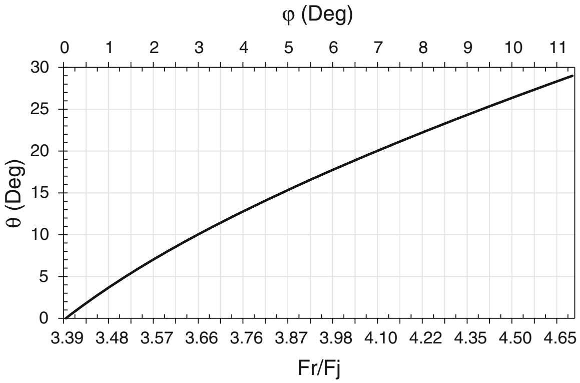


Fig. 4 Nonlinear relationships of θ with ϕ and F_r/F_j

3 Experimental Results

3.1 Stiffness Measurement

The Young's modulus is a measure of the stiffness of an elastic material and is a quantity used to characterize materials. An indentation experiment was conducted in this research to compute the effective Young's moduli of artificial tissue samples and study their mechanical properties relative together assuming linear elasticity. The assumption of linear elasticity relies on the applying of small indentation relative to the characteristic dimensions of the tissues. The effective elastic Young's modulus (E) corresponding to an indentation depth of δ may be calculated as follows [29]:

$$E = \frac{3F}{8d\delta} \tag{12}$$

where F is the reaction force and d is the diameter of a circular punch indenter applied to the soft tissue.

To conduct the experiment, a high-precision instrument was developed (Fig. 5) by retrofitting a micrometre with position resolution of 0.01 mm with a FSG-15N1A force sensor from Honeywell Sensing and Control (1985 Douglas Drive North Golden Valley, MN 55422 USA) with the force resolution of 1 gr. The instrument was used to apply precise indentations to the tissue samples and record the force responses using a U6 data acquisition (DAQ) module from LabJack Corporation (3232 S Vance St STE 100 Lakewood, CO 80227 USA) for three cycles per tissue sample. The average values were calculated and force-displacement relationships were obtained that are presented in Fig. 5.

The young modulus of the three selected soft tissues (Fig. 5) were calculated using Eq. (12) as 10.5, 17.7, and 26.0 kPa. These artificial tissue samples were deliberately chosen to be relatively close in terms of stiffness. They are also

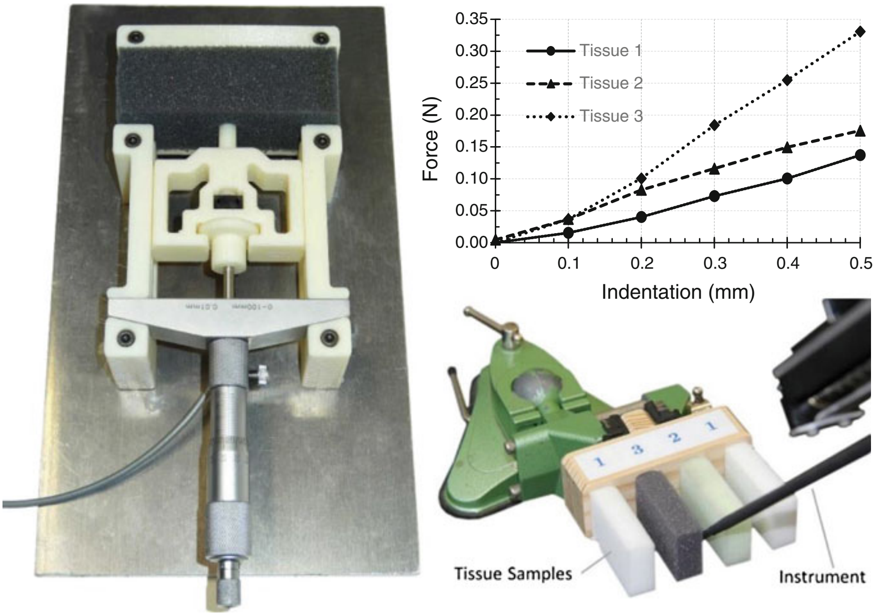


Fig. 5 The developed high-precision indentation device for stiffness measurement of soft tissue (*left*), force/displacement relationships (*right-top*) for the three selected artificial tissue samples made up of sponges (*right-bottom*)

selected to be representative of relatively softer tissues among human soft tissues with the range of young modulus from 0.1 to 241 kPa [30, 31]. These selections were made in order to properly evaluate the effectiveness of the proposed instrument in characterizing soft tissues with low and relatively close young moduli. These tissue samples were also examined by the surgeon collaborator to make a realistic choice of human soft tissues in the experimental studies.

3.2 Tissue Characterization Experiment

This experiment was conducted to evaluate the capability of the proposed instrument in non-invasive measurement of the grasping forces for tissue samples of varying stiffness. Three artificial tissue samples made up of sponge material that are identical in thicknesses but slightly different in stiffness were chosen for this experiment. In order to verify the accuracy of the measurement methodology as well as the correctness and the effectiveness of the post-processing algorithm, even in characterizing tissues with close stiffness, the tissues were purposely selected with slight variation in stiffness (Fig. 5), such that they would be hardly differentiated by direct exploration with one's fingers.

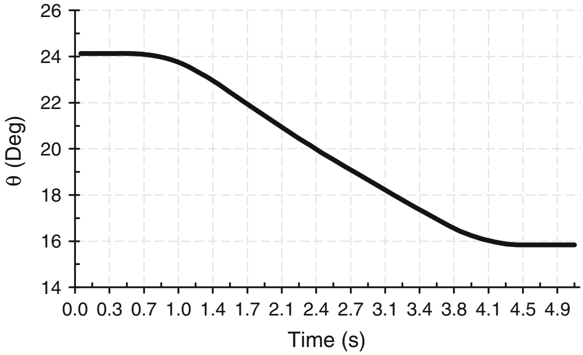


Fig. 6 Angular displacement of the tip jaws in the experiment

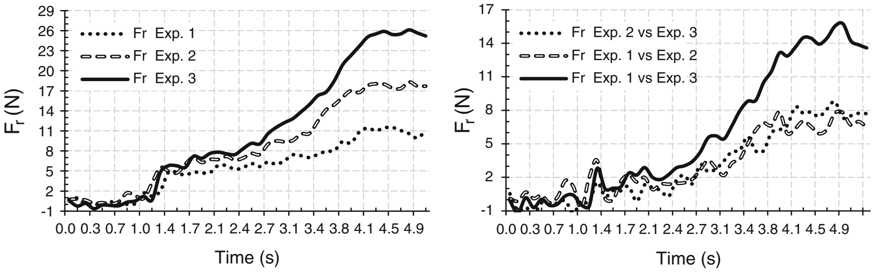


Fig. 7 Measured forces (*left*) and comparatives of the measured forces (*right*) applied to the push rod (F_r) in the experiment

In this experiment, according to the tip jaws angular displacements illustrated in Fig. 6, the tip jaws were closed from 24° to 16° while the tissue samples were grasped and the applied force to the push rod were measured using the strain gages applied to the lead-screw (Fig. 1). The initial positions of the tissue samples between jaws were held constant, therefore they all were incurred similar deformation but with different normal forces. The data acquired in this experiment are plotted in Fig. 7. This figure shows the measured forces applied to the push rod (F_r). In the real palpation, for the purpose of diagnosing the normal/infected tissues, usually comparative forces/stiffness of the tissues are of interest. Therefore, the comparatives between the measured forces applied to the push rod for the three different tissues are also calculated and are also plotted in Fig. 7. The normal forces applied to the tissues at tip jaws calculated using Eq. (9) as well as the comparatives between these forces are plotted in Fig. 8. According to the Figs. 6, 7 and 8, the jaws started to close at the initial angular position of 24° at time 0.7 s and reached the final position of 16° at time 4.3°.

Compression started at time 1.2 s with tip jaws in angular position of 23.2° where the recorded data showed a sudden increase in grasping forces. As the jaws were closing, the grasping forces were increasing but with higher rate for harder

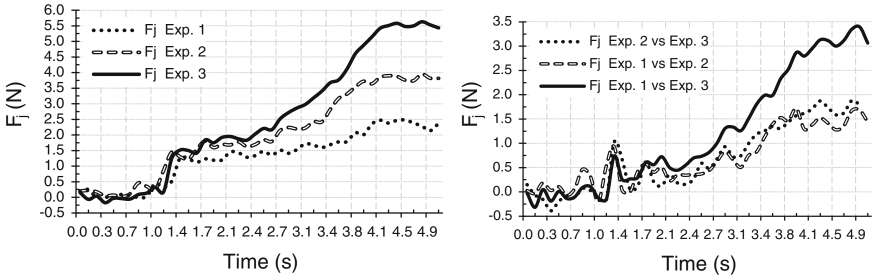


Fig. 8 Calculated forces (*left*) and comparatives of the calculated forces (*right*) applied to the artificial tissue samples (F_j) in the experiment

tissue than the rate for softer tissue. Once the jaws got to the final position at time 4.3 s, the forces remained constant to the maximum grasping forces. In this experiment, the measured peak forces applied to the push rod of the insert for three tissue samples from softer to harder are 12, 18, and 26 N, respectively.

These forces are, respectively, equivalent to 2.5, 3.8, and 5.6 N of the calculated normal forces at jaws directly applying to the tissue (Fig. 8). As it is also observed from the Figs. 7 and 8, the differences between increase rates of F_j and F_r increased as jaws were closing, got to the maximum difference at time 2.4 s where the jaws were at 20°, and decreased to its minimum value in this experiment when the tip was at its closest state of 16°. This phenomenon confirms the nonlinear relationship of F_j and F_r or the angular position of tip jaws (θ) with the angular displacement of the actuator shaft (ϕ) as it is also presented in Fig. 4. As it is shown in Figs. 7 and 8, there is a jump in all of the comparative forces at the time 1.4 s where the compression is started which is verify the effectiveness of the proposed instrument in correctly measuring the grasping forces for different tissue sample specifically at the time of the compression where the magnitude of the grasping forces are in the lower range. As it is also clear from the Figs. 7 and 8, the measured forces using the proposed instrument are easily distinguishable which results in ability to characterise soft tissues of varying stiffness. Therefore, it can be concluded that the proposed laparoscopic instrument has good accuracy and performance for the grasping force measurement and is able to distinguish between tissues of varying stiffness even with relatively close young moduli.

4 Conclusion and Future Works

An automated minimally invasive surgical instrument was introduced, the modelling and development issues were discussed, and experimental results were presented and analysed in this paper. The proposed surgery instrument has the capability of minimally invasively measuring normal tip interaction forces e.g. grasping and cutting. The instrument features non-invasive actuation of the tip and also the measurement of interaction forces without using any actuator and sensors at the jaws. The grasping direction in the proposed instrument can also be adjusted during the surgical procedure. The modularity feature of this force feedback-enabled minimally invasive instrument makes it interchangeable between various tool tips of all functionalities (e.g. cutter, grasper, and dissector) without loss of control and force measurement capability necessary to avoid tissue damage and to palpate and diagnose tissue and differentiate its stiffness during surgery. A high precision device for the measurement of young modulus of soft tissues were developed and utilised in this research. Experiments were conducted to evaluate capabilities of the proposed instrument in non-invasively measuring normal grasping forces. The result showed high accuracy and performance and verified the ability of the instrument in measuring normal grasping forces and in distinguishing between tissue samples even with slight differences in stiffness. The sterilizability of the instrument and especially the force sensing sleeve also needs improvements in future works before it can be used in surgery operating room.

Acknowledgement This research is funded by Australian Research Council, ARC Discovery-DP0986814, ARC LIEF-LE0668508, and ARC LIEF-LE0453629.

References

1. A.R. Lanfranco, A.E. Castellanos, J.P. Desai, W.C. Meyers, Robotic surgery: a current perspective. Ann. Surg. **239**, 14 (2004)
2. M. Moradi Dalvand, B. Shirinzadeh, J. Smith, Effects of realistic force feedback in a robotic assisted minimally invasive surgery system. Minimally Invasive Therapy and Allied Technologies (MITAT) (2013) (in press)
3. S.M. Sukthankar, N.P. Reddy, Towards force feedback in laparoscopic surgical tools, in *Proceedings of the 16th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 1994. Engineering Advances: New Opportunities for Biomedical Engineers.* IEEE (1994)
4. C.E. Reiley et al., Effects of visual force feedback on robot-assisted surgical task performance. J. Thorac. Cardiovasc. Surg. **135**, 196–202 (2008)
5. M. Fakhry, F. Bello, G.B. Hanna, A real-time compliance mapping system using standard endoscopic surgical forceps. IEEE Trans. Biomed. Eng. **56**, 1245–1253 (2009)
6. P. Lamata et al., Understanding perceptual boundaries in laparoscopic surgery. IEEE Trans. Biomed. Eng. **55**, 866–873 (2008)
7. P. Dubois, Q. Thommen, A.C. Jambon, In vivo measurement of surgical gestures. IEEE Trans. Biomed. Eng. **49**, 49–54 (2002)
8. J. Rosen, B. Hannaford, C.G. Richards, M.N. Sinanan, Markov modeling of minimally invasive surgery based on tool/tissue interaction and force/torque signatures for evaluating surgical skills. IEEE Trans. Biomed. Eng. **48**, 579–591 (2001)
9. S.K. Prasad et al., in *Medical Image Computing and Computer-Assisted Intervention*, pp. 279–286
10. J. Dargahi, M. Parameswaran, S. Payandeh, A micromachined piezoelectric tactile sensor for an endoscopic grasper: theory, fabrication and experiments. J. Microelectromech. Syst. **9**, 329–335 (2000)
11. U. Seibold, B. Kuebler, G. Hirzinger, Prototype of instrument for minimally invasive surgery with 6-axis force sensing capability, in *ICRA* (2005)

12. B. Kuebler, U. Seibold, G. Hirzinger, Development of actuated and sensor integrated forceps for minimally invasive robotic surgery. Int. J. Med. Robot. Comput. Assist. Surg. **1**, 96–107 (2005)
13. P. Valdastri et al., Integration of a miniaturised triaxial force sensor in a minimally invasive surgical tool. IEEE Trans. Biomed. Eng. **53**, 2397–2400 (2006)
14. M. Mahvash et al., Modeling the forces of cutting with scissors. IEEE Trans. Biomed. Eng. **55**, 848–856 (2008)
15. M.C. Yip, S.G. Yuen, R.D. Howe, A robust uniaxial force sensor for minimally invasive surgery. IEEE Trans. Biomed. Eng. **57**, 1008–1011 (2010). doi:10.1109/tbme.2009.2039570
16. J. Rosen, B. Hannaford, M.P. MacFarlane, M.N. Sinanan, Force controlled and teleoperated endoscopic grasper for minimally invasive surgery-experimental performance evaluation. IEEE Trans. Biomed. Eng. **46**, 1212–1221 (1999)
17. G. Tholey, J.P. Desai, A modular, automated laparoscopic grasper with three-dimensional force measurement capability, in *IEEE International Conference on Robotics and Automation.* IEEE (2007)
18. C.R. Wagner, N. Stylopoulos, R.D. Howe, The Role of Force Feedback in Surgery: Analysis of Blunt Dissection, in *Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems* (2002)
19. M. MacFarlane, J. Rosen, B. Hannaford, C. Pellegrini, M. Sinanan, Force-feedback grasper helps restore sense of touch in minimally invasive surgery. J. Gastrointest. Surg. **3**, 278–285 (1999)
20. A.J. Madhani, G. Niemeyer, J.K. Salisbury, The black falcon: a teleoperated surgical instrument for minimally invasive surgery, in *Proceedings on 1998 IEEE/RSJ International Conference on Intelligent Robots and Systems*, vol. 2. IEEE (1998)
21. J. Rosen, J.D. Brown, L. Chang, M.N. Sinanan, B. Hannaford, Generalized approach for modeling minimally invasive surgery as a stochastic process using a discrete Markov model. IEEE Trans. Biomed. Eng. **53**, 399–413 (2006)
22. S. Shimachi, Y. Hakozaki, T. Tada, Y. Fujiwara, Measurement of force acting on surgical instrument for force-feedback to master robot console, in *International Congress Series*, vol. 1256. Elsevier (2003)
23. M. Moradi Dalvand, B. Shirinzadeh, S. Nahavandi, F. Karimirad, J. Smith, in *Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering and Computer Science*, WCECS 2013, pp. 419–424 (2013)
24. M. Moradi Dalvand, B. Shirinzadeh, Forward kinematics analysis of offset 6-RR$\underline{\text{C}}$RR parallel manipulators, in *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science* **225**, 3011–3018 (2011)
25. M. Moradi Dalvand, B. Shirinzadeh, Motion control analysis of a parallel robot assisted minimally invasive surgery/microsurgery system (PRAMiSS). Robot. Comput.-Integr. Manuf. **29**, 318–327 (2013). doi:10.1016/j.rcim.2012.09.003
26. M. Moradi Dalvand, B. Shirinzadeh, Remote centre-of-motion control algorithms of 6-RRCRR parallel robot assisted surgery system (PRAMiSS), in *IEEE International Conference on Robotics and Automation (ICRA)*, IEEE (2012)
27. T. Hu, A. Castellanos, G. Tholey, J. Desai, Real-time haptic feedback in laparoscopic tools for use in gastro-intestinal surgery, in *Medical Image Computing and Computer-Assisted Intervention—MICCAI,* pp. 66–74 (2002)
28. M. Tavakoli, R. PateI, M. Moallem, in *Proceedings of IEEE International Conference on Robotics and Automation*, ICRA'04, pp. 371–376 (2004)
29. M.J. Uddin, Y. Nasu, K. Takeda, S. Nahavandi, G. Capi, An autonomous trimming system of large glass fiber reinforced plastics parts using an omni-directional mobile robot and its control, in *Proceedings: Eight International Conference on Manufacturing and Management: Operations Management and Advanced Technology: Integration for Success*. PCMM (2004)
30. S.B. Kesner, R.D. Howe, Position control of motion compensation cardiac catheters. IEEE Trans. Robot. **27**, 1045–1055 (2011)
31. S.G. Yuen, N.V. Vasilyev, P.J. del Nido, R.D. Howe, Robotic tissue tracking for beating heart mitral valve surgery. Med. Image Anal. **17**, 1236–1242 (2013)

Chapter 37
Perception and Trust Towards a Lifelike Android Robot in Japan

Kerstin Sophie Haring, Yoshio Matsumoto and Katsumi Watanabe

Abstract This paper reports the results from an experiment examining people's perception and trust when interacting with an android robot. Also, they engaged in an economic trust game with the robot. We used the physical distance to the robot, and questionnaires to measure the participants' character and their perception of the robot. We found influences of the subject's character onto the amount sent in the trust game and distance changes over the three interaction tasks. The perception of the robot changed after the interaction trials towards less anthropomorph and less intelligent, but safer.

Keywords Android robot · Eysenck personality questionnaire · Godspeed questionnaire · Human robot interaction · Robot perception · Trust game

1 Introduction

Robots no longer belong to the world of science fiction, they are reality and more sooner than later will be having a real impact on the way we live. In our schools, homes, workplaces, museums, hospitals, shops we are, and will continue to be, interacting with robots so it is crucial that we begin to examine this young and fast

K. S. Haring (✉) · K. Watanabe
Research Center for Advanced Science and Technology, The University of Tokyo, 4-6-1, Komaba, Meguro-ku, Tokyo 153-8904, Japan
e-mail: ksharing@fennel.rcast.u-tokyo.ac.jp

K. Watanabe
e-mail: kw@fennel.rcast.u-tokyo.ac.jp

Y. Matsumoto
Tsukuba Central 2, 1-1-1, Umezono, Tsukuba, Ibaraki 305-8568, Japan
e-mail: yoshio.matsumoto@aist.go.jp

H. K. Kim et al. (eds.), *Transactions on Engineering Technologies*,
DOI: 10.1007/978-94-017-9115-1_37,

changing field. Today, the production of sophisticated robots enables researchers to examine the interaction between people and robots for the first time.

Human Robot Interaction (HRI) research faces a number of key challenges. On the engineering side, there are technological hurdles and on the human-factor side, the many influences forming social experiences. Social aspects of Human Robot Interaction (HRI) are critical for the future scenario if robots are to become a part of people's life. To come to a broad understanding what forms interaction (with robots) these factors must be teased apart for proper studies.

Despite recent progress of robotics and robots showing more and more capabilities, our perceptions and expectations towards robots are more shaped by what we see in the news, videos or movies than by real interaction with an actual physically present robot [1]. Even in this study which was conducted in a technological advanced country (Japan), real contact to a robot was the exception rather than the norm. It has been shown that the physical embodiment [2], means the actual presence of a robot, does play a crucial role in HRI and the perception of a robot. This study addresses this issue by letting humans directly interact with a very human-like android robot. Previous studies examined the effect of robot design and appearance onto people's perception and expectation towards them [3, 4]. In these studies, the android robot triggered significantly different reactions from subjects compared to other robot types (e.g. humanoid robot, pet robots) (Fig. 1).

An android robot is a very specific robot type that is designed to look and attempts to act like a human. In addition to the humanoid shape it has detailed features like skin, eyelashes, hair, etc. For this study, the female version of the android robot Geminoid was used. In a short glace (under 2 s.), human observers tend to recognize this android robot as a human (70 % of subjects failed to distinguish human and android). Android robots also give humans an eerie feeling, which was defined as "uncanny valley" [5]. It is stated that, as the appearance of robots becomes more human, they seem more familiar, until a point is reached at which the response from the observer quickly drops from positive to a strong revulsion.

To examine the perception towards this potential "uncanny" robot we decided to conduct a study in which subjects met an android robot "in person" for the first time.

Another interesting and challenging research question is how much people trust such a life-like robot. Under laboratory conditions, one way to measure trust is a so-called economic trust game [6], which allows us to empirically quantify trust in relationships in a reliable and standardized way. With these trust games, interaction behavior can be measured between humans. In this study, we applied this game to quantify the trust of humans towards the android robot. Our application of the game-theoretic approach to human robot interaction provides a model of empirical measurement of trust towards robots. This method can be applied to any type of robot and future data can be easily compared.

Social exchange is influenced by many factors. Recent research developments have looked directly at subject's decisions from the perspective of personality

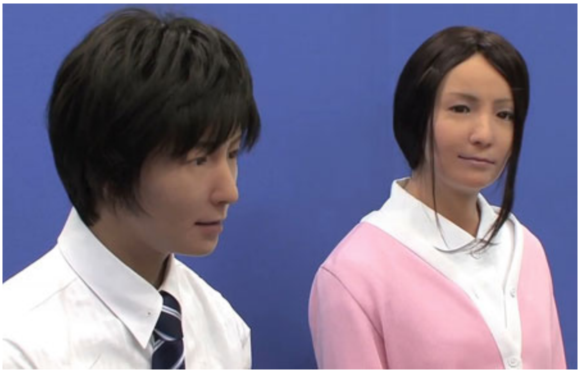


Fig. 1 The very life-like android robot Actroid in the male and female version. The appearance of this robot is so close to a real human that, on a picture, it is very difficult to distinguish it from a real human. This experiment used the Actroid-F, the female version of the android robot

psychology and suggested that individual character and personality type influence subject's decisions in laboratory experiments [7–9, 10, 11].

Investigation of specific personality traits has shown [7] that extravert personality type subjects (directed towards the objective world) exhibited stronger than expected feelings of trust with the amount sent in the trust game. Related research in the field of HRI [12] suggests that people's preferences for robot appearance and behavior may be related to their personality traits. It has been indicated that more introverted individuals tended to prefer mechanoid robot appearance and extraverts preferred more humanoid robots.

Only very recently researchers started to use robots to assess cooperative intentions of novel partners by using nonverbal signals with the help of robots and a variant of the prisoner's dilemma game [13, 14]. Another study focused on the physical appearance of the robot face and stated that participants acted as if they attributed complex human-like motivations to the shown robots. Participants did not have direct contact with a robot but were presented a photo-based survey [15].

Surveys are quick ways to measure human perception. However, there are some limitations to this method as participants might be biased. It is therefore crucial that also objective methods through observation of the subject's behaviors are applied. An example for such observations would be the times a subject needs to fulfill a task or spends with the robot and also the distance he or she keeps away from the robot.

A study examining the complex relationship of proxemics to robots found that 60 % of the participants chose distances to the robot that were comparable to normal human–human social interaction distances [16]. It was also found that the experience with owning pets and robots decreased personal space and the directions of the robot's gaze had an effect on proxemics behavior [17].

Factors like prior experiences with robots, prior relationships with non-human agents like pets [16], and the subject's personality [18–20] may influence the interaction. A crucial factor is the physical presence of the robot [2] rather than just having a picture or video of the robot.

This led to the construction of a complex experiment that intended to evaluate changes in participant's perception in relation to their personality traits, trust in relation to their personality traits and observations like proxemics and task times. The goal was to evaluate these factors under laboratory conditions with an actual physically present robot. We used previously validated scales and questionnaires that would enable future research to compare robot and personality types. Participants interacted with the robot three times, allowing them to get more familiar with the robot. It engaged them in a short conversation (e.g. introducing itself, asking the name) and then instructed a very simple task (move a box) in the first two trials and to touch its hand in the third. We conducted a 2 × 2 condition experiment and participants were assigned randomly to one of the two condition conditions: the robot turned the head toward the boxes when talking about them and a higher/lower payback in the trust games as described below. We hypothesized that:

1. Subjects would reduce the amount of space between them and the robot over the three trials.
2. The subject's performance in the trust game would be linked to their likeability toward the robot.
3. Participant's previous experience with non-human agents (e.g. pets, virtual agents, etc.) would decrease their distance to the robot [17].
4. Individuals with extravert personality traits would make higher offers in the trust game.

2 Materials and Methods

2.1 Participants

Fifty-six subjects were recruited from local universities in Japan. One subject was eliminated from the examination due to participant bias (demand characteristics), which left 55 subjects. Age ranged from 18 to 66 years (with the mean of 22.6 years). 37 participants were female, 18 male. None of the participants had experienced any interaction with an android robot before. The interaction context and questionnaire were in Japanese. The participants received monetary reimbursement for their participation. They were randomly assigned to the experimental conditions.

2.2 Experiment Flow

The experiment consisted of three main stages. In the first stage, the subject's basic demographic data, personality traits (Eysenck Personality questionnaire [21], and perception of the robot were evaluated (Godspeed [22]). To evaluate changes in participant's perception of the robot, the questionnaire was administered before the interaction tasks (showing pictures of android robots) and after the interaction tasks with the robot. Participants were also asked and if they ever owned a pet, their prior exposure to virtual agents (e.g. in computer games) and to robots. This was rated on a 5-point scale from 1 (never) to 5 (nearly daily).

The second stage consisted of three simple interaction tasks with the robot. In trial 1 and 2 the robot asked the subject to move a box to another position, in task 3 it asked to touch its hand. Before each task, the robot engaged the participant in a small conversation and then gave instructions of the task. When the task was completed, the robot thanked them for their cooperation and asked the subject to wait outside the room.

In the third stage an economic trust game was 'played' between the robot and participants in a similar context to that used during human–human interaction [6, 23]. In the two-player trust game player 1 (here always the subject) was endowed with a fixed amount of money (JPY1000, approximately $10), and given the option of sending any portion of the money to player 2 (the robot). The returning amount from the robot to the subject was manipulated by the researcher. Depending on the randomly assigned condition the payback would be either JPY200 less or more money than the subject had initially sent to the robot.

2.2.1 Eysenck Personality Questionnaire

The Eysenck Personality questionnaire [23] categorizes personalities in a systematic way, using the three factors of psychoticism, extraversion and neuroticism. It is also one of the few personality questionnaires that is validated in Japanese and other languages for a later direct comparison of intercultural studies. We used the Japanese version of the short-form Eysenck Personality Questionnaire—Revised (EPQ-R) [6]. The questions were presented in random order.

2.2.2 Robot Perception

The Godspeed questionnaire [22] measures five key concepts in HRI using 5-point scales. Anthropomorphism is the attribution of a human form and human characteristics. As mentioned, in a short glance for example (under 2 s.), humans tended to recognize the android robot as a human. Animacy is the perception of the robot as a lifelike creature. Perceiving something as alive allows humans to distinguish humans from machines. As it is emphasized in Piaget's framework [24], major factors of "being alive" are movement and intentional behavior. Likeability describes the first (positive) impression people form of others. Research suggests [25] that humans treat robots as social agent and therefore judge them in a similar way. Perceived intelligence states how intelligent and human-like subjects judge the behavior of the robot. The android robot we used here is mainly interacting with pre-programmed speech and head turns, and tele-operated by a researcher. According to Bartneck and colleagues [22], the subject's rating depends on the robot's competence. Perceived safety describes the perception of danger from the robot during the interaction and the level of comfort the subject's experience. The questions were presented in random order.

To compare the perception, we calculated the mean of every category of the 5-scale Godspeed questionnaire before and after the interaction trials. Before the experiment, subjects were shown two pictures of the android robot, after the experiment, they only had to fill in the questionnaire.

2.2.3 Distance and Touch Time Measurements

For the proxemics, we measured the distance from the robot to the position the subject put the chair during the interaction task. For that, the robot asked them to sit on a chair that was positioned at the end of the room in a way that participants had to pick it up from there and roll it to a position they wanted to sit down. The researchers readjust the chair position before the subject entered for the next task. The subjects did not know that their distance to the robot and the task times were measured during the experiment.

We also measured the times when the subject initiated touch as the robot verbally asked subjects to touch its hand. The time was measured from the first request until the subjects made physical contact with the robot's hand. All participants did touch the hand. If they hesitated, the robot asked again. In addition, we recorded how long the subjects were touching the hand (with minimum of 1 s).

2.2.4 Statistical Tests

The data was analyzed with the statistical software R. The data was summarized over means and we employed hypothesis tests. There were no outliers outside of three standard deviations of the mean.

3 Results

The reaction of people interacting with the android robot ranged from apathetic to exhibiting great enjoyment and excitement. Most participants were open to the experience with the robot and curious about it.

3.1 Distance to the Robot

The data supports the hypotheses 1, that participants come significantly closer to the robot in each trial (pairwise *t*-test, trial 1 vs. trial 2 $t(54) = 4.87$; $p < 0.001$, trial 2 vs. trial 3 $t(54) = 2.67$; $p < 0.05$, Bonferroni corrected). The mean absolute distances from the robot during the interaction tasks are presented in Table 1.

Table 1 Distance to the robot in the three interaction trials

Trial 1	Trial 2	Trial 3
128.2 cm	119.9 cm	116.2 cm

Table 2 Mean values of the Godspeed robot perception questionnaire before and after the interaction trials

	Before	After
Anthropomorphism	3.10	2.45
Animacy	2.83	2.85
Likeability	2.85	2.85
Perceived intelligence	3.49	2.85
Perceived safety	2.67	2.85

3.2 Perception Before and After the Trials

The data does not show any significant difference in animacy and likeability for differences before and after the task (paired *t*-test), but we found significantly different results for Anthropomorphism ($t(53) = 4.22$, $p < 0.001$), perceived intelligence ($t(53) = 7.55$, $p < 0.001$) and perceived safety ($t(53) = -1.99$, $p = 0.05$). We missed the data from one subject because the questionnaire was not filled in. The results are displayed in Table 2. Participants perceive the robot as much less anthropomorphic and intelligent after the interaction tasks, but also much safer.

3.3 Robot Perception and Trust Game

When looking at the robot perception and the trust game, we could not find any support for the hypotheses 2, but the amount sent in the trust game to the robot after the interaction tasks was higher, if the participants rated the perceived intelligence higher before the interaction trials (correlation, $r = 0.28$, $p = 0.03$). However, the amount sent in the trust game did not show any significant correlations with other categories of robot perception.

3.4 Personality and Trust Game

We examined the hypothesis 4 that extraverts would endow the robot with a higher amount in the trust game. The data showed that the more extravert a person was, the higher the amount sent in the trust game (correlation, $r = 0.44$, $p < 0.001$). Other character traits did not correlated with the amount sent in the trust game.

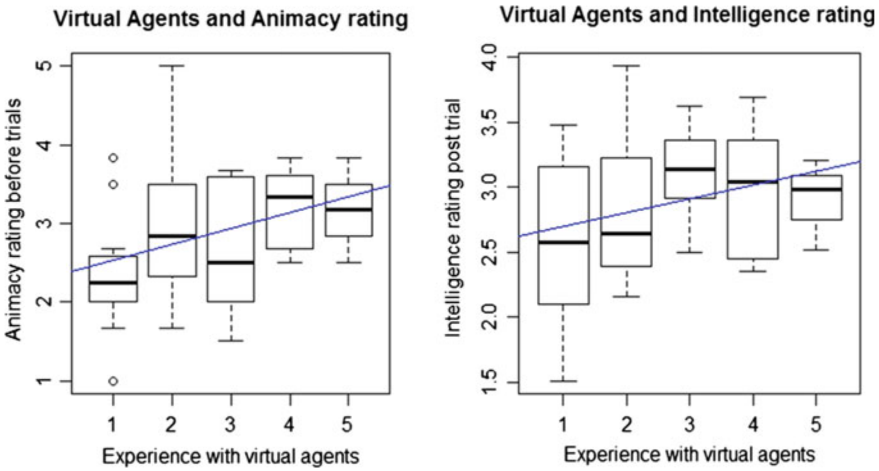


Fig. 2 The prior experience with virtual agents and the influence on the animacy rating before and the intelligence rating after the interaction trials

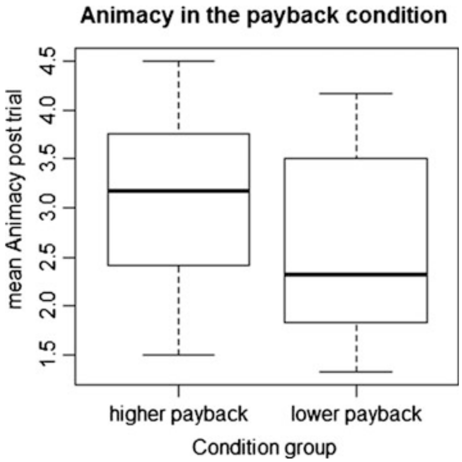
3.5 Exposure to Virtual Agents

We found that the more people were exposed to virtual agents in their daily life, the higher they rated anthropomorphism and animacy before the trials (but not after) and intelligence and safety after the trials (but not before). The perception of anthropomorphism and perceived safety followed the general trend, but significant differences were found in animacy ($F(1{,}52) = 5.21$, $p = 0.02$, one subject's data is missing because the questionnaire was not filled in) and perceived intelligence ($F(1{,}53) = 3.03$, $p = 0.08$) (Fig. 2).

3.6 Non-human Agents and Reward Condition

When subjects never had a pet, then the payback from the robot in the trust game had an influence on the perception of animacy (2-way ANOVA, $F(1{,}51) = 6.44$, $p = 0.01$), likeability ($F(1{,}51) = 5.63$, $p = 0.02$), perceived intelligence ($F(1{,}51) = 6.26$, $p = 0.01$) and safety ($F(1{,}51) = 6.69$, $p = 0.01$) (the trust game took place after the interaction tasks). If the (manipulated) payback from the robot was more than the amount the subject sent, the subjects who never had a pet rated these categories significantly higher than the subjects in the lower payback group.

Fig. 3 The difference in the rating of the animacy after the interaction trials in the two payback conditions of the trust game



3.7 Robot Perception and Reward Condition

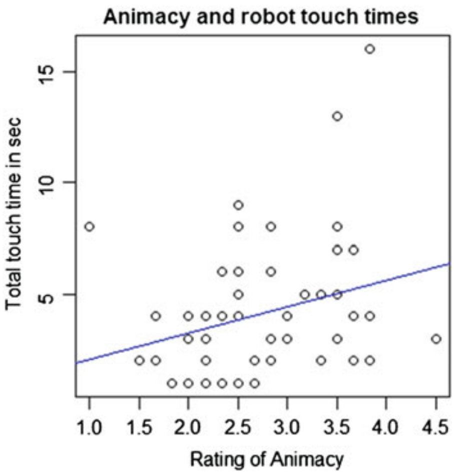
Subjects in the higher payback condition tended to rate the robot slightly more as a lifelike creature (animacy) than subjects in the lower payback condition, but it did not reach significance (ANOVA, $F(1,53) = 3.24$, $p = 0.07$) (Fig. 3).

3.8 Robot Perception and Touch Times

We observed some weak effects for the time it took until subjects touched the robot. It seems like that a higher likeability before the interaction tasks was correlated with a shorter time until people touched the robot. There was no correlation found in the data for a change in likeability after the interaction tasks. In humans, touch is an important channel for social communication. Yet, the scenario of someone requesting to touch the hand might be quite unlikely and therefore sound odd. To get comparative data here, we are now planning a follow-up experiment comparing the touch times to a human interaction partner.

For the total touch time, we removed one outlier from the data. One female participant held the robot's hand for nearly 80 s (mean touch was $M = 5.51$, $SD = 10.46$, outside of seven standard deviations of the mean). After excluding her, we split the data at the mean animacy rating of all the subjects and compared them. The statistical analysis showed that the total touch duration was longer if the animacy of the robot was rated higher ($F(1,49) = 4.06$, $p = 0.04$, one subject's data is missing because the questionnaire was not filled in, two touch time data were not measured correctly) (Fig. 4).

Fig. 4 The subjects touch the hand of the robot longer when the rating of the animacy before the interaction trials is higher



4 Discussion

The interaction with this highly realistic looking robot revealed that people get more familiar with the robot over time and their perception of the robot changed significantly when they actually interacted with it. We found that the lack of experience with non-human agents (pets) had an influence on the robot perception. The more extravert participants entrusted the robot with a higher amount of money.

When observing their distance to the robot there were clear tendencies that participants came closer to the robot. We think this shows that the subjects got more familiar with it. This effect was strong enough to go beyond the initial surprise effect this robot would have in the first encounter. Subjects also came significantly closer in the second and third interaction scenario. From the literature we know that people maintain shorter distances to somebody they feel close with [26, 27], but larger distances to somebody they dislike or that carries a physical stigma [26]. People maintain a personal space of roughly 1.2 m around themselves that is usually not violated by others [27]. In the present study, we observed roughly similar distances to this.

As expected, the participants rated the robot before the experiment as quite anthropomorph and intelligent. At the same time, they exhibited a kind of uncomfortable feeling when judging safety and pleasantness of the interaction. We think that at the beginning most of the subjects only knew from the description of the picture that they saw a robot. After interacting with the robot, they rated it as less anthropomorph and intelligent but at the same time as safer. This indicates how important the actual embodiment of the robot in direct interactions would be. It seems that the participants had high expectations in terms of human form and characteristics (anthropomorphism) and intelligence towards the robot that were not fulfilled. At the same time, after actually seeing and interacting with the robot, they felt more comfortable in its presence. This is congruent with a study stating that the exterior of a robot shapes the expectations towards it [4]. One explanation could be that the participants recognized the shortcoming of the robots after the interaction, which they could not judge from the picture shown before the interaction. As the robot is by far not as realistic as the very first impression in the picture suggested, they would also perceive it less as a thread and more as a "safer" agent. These results could indicate that the importance of the exterior is still underrated and that future robot designers should carefully consider not only the abilities of the robot, but also how it looks.

The subjects also sent more money to the robot if they perceived the robot as intelligent before the interaction tasks. The perception of intelligence might lead them to think that the robot might be able to increase both their outcome in the trust game.

As stated in the hypotheses 4, extravert people were more open and more likely to endow to robot with a higher amount in the trust game. This is consistent with the results of the previous study [7]. Also, the subjects paid a similar amount of money to the robot as on other studies before by using human agents [6]. We conclude therefore that the robot, in a sense, was perceived as a trustworthy and intelligent partner for such an economic game under laboratory conditions.

It seemed that a higher payback in the trust game influenced how life-like the participants perceive the robot and their impression with a higher payback made the robot slightly more "human" than "machine" to them, as they benefit from the robot's "decision". In particular, if subject's never had a pet, a higher payback in the trust game influenced the perception of the robot. It could be that people who never experienced the "rewarding" company of a pet rely here on more on the monetary outcome from the human robot relationship.

If the subjects had a prior higher exposure to virtual agents, they perceived the robot very life-like (animacy) at first but this perception changed after the interaction with the robot. This could indicate that the robot "in movement" was perceived less life human-like than the picture of it might suggest. Also, if the exposure to virtual agents was higher, the perceived intelligence was higher after the experiment. This could be that the subjects highly exposed to virtual agents did not expect a real robot to leave that level of competent impression. This could be an indication that future robots could adjust their interaction with people to their possible preferences and infer interaction strategies and models directly from the reaction of humans to maximize the positivity of the experience.

5 Future Work

There are several analyses and experiments of interest. We are planning to evaluate the video data towards how the subjects touched the robot. There have been studies examining how people touch strangers and what kind of touch expresses a certain emotion [28, 29].

To evaluate the data in comparison with a human interaction partner, we also plan to conduct an experiment with a human agent with the same interaction tasks. This will enable us to compare the distance, touch times, and the perception of a human agent compared to a robot. It would also be interesting to examine possible effect of cultural background. This experiment was performed also in Australia to compare the perception and trust of this android robot directly to data from subjects with a different cultural background. The data evaluation is currently ongoing.

The android robot Geminoid-F is a very specific and human-like robot. To compare the data to a different, more machine-like robot, we also plan to conduct this experiment with a humanoid robot that is not designed to look exactly like a human. We expect people to perceive and approach this robot in a different manner than the Geminoid-F.

Acknowledgment We thank the Service Robotics Group of the Advanced Industrial Science and Technology (AIST) for the provision of the robot and the technical support. We also would like to thank Toru Hosokawa from the University of Tohoku for the copy of the Japanese version of the Eysenck personality questionnaire. This work was supported in part by the Mitsubishi Shoji Scholarship and the Japan Science and Technology Agency (CREST).

References

1. Z. Khan, *Attitudes Towards Intelligent Service Robots*, vol. 17 (NADA KTH, Stockholm, 1998)
2. J. Wainer, D.J. Feil-Seifer, D.A. Shell, M.J. Mataric, in *The 15th IEEE International Symposium on Robot and Human Interactive Communication, ROMAN 2006*. The role of physical embodiment in human-robot interaction (IEEE, 2006), pp. 117–122
3. K.S. Haring, C. Mougenot, K. Watanabe, in *5th International Conference on Knowledge and Smart Technologies (KST)*. Perception of different robot design. Special session on "Fluency in communication between human, machine, and environment" (2013)
4. K.S. Haring, C. Mougenot, K. Watanabe, in *8th Annual Conference for Basic and Applied Human-Robot Interaction Research (HRI)*. The influence of robot appearance on assessment (Tokyo, 2013)
5. M. Mori, Bukimi No Tani (the uncanny valley). Energy **7**, 33–35 (1970)
6. J. Berg, J. Dickhaut, K. McCabe, Trust, reciprocity, and social history. Games Econ. Behav. **10**(1), 122–142 (1995)
7. K.J. Swope, J. Cadigan, P.M. Schmitt, R. Shupp, Personality preferences in laboratory economics experiments. J. Socio-Econ **37**(3), 998–1009 (2008)
8. J. Wischniewski, S. Windmann, G. Juckel, M. Brüne, Rules of social exchange: game theory, individual differences and psychopathology. Neurosci. Biobehav. Rev. **33**(3), 305–313 (2009)
9. P. Schmitt, R. Shupp, K. Swope, J. Mayer, Pre-commitment and personality: behavioral explanations in ultimatum games. J. Econ. Behav. Organ. **66**(3), 597–605 (2008)
10. H. Brandstätter, W. Güth, Personality in dictator and ultimatum games. CEJOR **10**(3), 191–215 (2002)
11. A. Ben-Ner, F. Kong, L. Putterman, Share and share alike? Gender-pairing, personality, and cognitive ability as determinants of giving. J. Econ. Psychol. **25**(5), 581–589 (2004)
12. M.L. Walters, K. L. Koay, D. S. Syrdal, K. Dautenhahn, and R. Boekhorst. "Preferences and perceptions of robot appearance and embodiment in human-robot interaction trials." Procs of New Frontiers in Human-Robot Interaction (2009)

13. D. DeSteno, C. Breazeal, R.H. Frank, D. Pizarro, J. Baumann, L. Dickens, J.J. Lee, Detecting the trustworthiness of novel partners in economic exchange. Psychol. Sci. **23**(12), 1549–1556 (2012)
14. K.S. Haring, Y. Matsumoto, K. Watanabe, in *Proceedings of The World Congress on Engineering and Computer Science 2013, WCECS 2013*. How do people perceive and trust a lifelike robot? Lecture notes in engineering and computer science, San Francisco, 23–25 October 2013, pp. 425–431
15. M.B. Mathur, D.B. Reichling, in *Proceedings of the 4th ACM/IEEE International Conference on Human Robot Interaction*. An uncanny game of trust: social trustworthiness of robots inferred from subtle anthropomorphic facial cues, (ACM, 2009), pp. 313–314
16. M.L. Walters, K. Dautenhahn, K.L. Koay, C. Kaouri, S.N. Woods, C.L. Nehaniv, R. te Boekhorst, D. Lee, I. Werry, in *Proceedings of Cog Sci 2005 Workshop: Toward Social Mechanisms of Android Science*. The influence of subjects' personality traits on predicting comfortable human-robot approach distances' (2005), pp. 29–37
17. L. Takayama, C. Pantofaru, in *IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2009*. Influences on proxemic behaviors in human-robot interaction (IEEE, 2009), pp. 5495–5502
18. M.L. Walters, K. Dautenhahn, R. Boekhorst, K.L. Koay, C. Kaouri, S. Woods, C. Nehaniv, D. Lee, I. Werry, in *IEEE International Workshop on Robot and Human Interactive Communication, ROMAN 2005*. The influence of subjects' personality traits on personal spatial zones in a human-robot interaction experiment (IEEE, 2005), pp. 347–352
19. B. Friedman, P.H. Kahn Jr, J. Hagman, in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. Hardware companions?: What online AIBO discussion forums reveal about the human-robotic relationship (ACM, 2003), pp. 273–280
20. T. Kanda, H. Ishiguro, M. Imai, T. Ono, in *IJCAI*. Body movement analysis of human-robot interaction (2003), pp. 177–182
21. T. Hosokawa, M. Ohyama, Reliability and validity of a Japanese version of the short-form Eysenck personality questionnaire-revised. Psychol. Rep. **72**(3), 823–832 (1993)
22. C. Bartneck, D. Kulić, E. Croft, S. Zoghbi, Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. Int. J. Soc. Robot. **1**(1), 71–81 (2009)
23. H.J. Eysenck, Manual of the Eysenck personality scales (EPS Adult). 23–24 (1991)
24. D. Parisi, M. Schlesinger, Artificial life and Piaget. Cogn. Dev. **17**(3), 1301–1321 (2002)
25. B. Reeves, C. Nass, *The Media Equation: How People Treat Computers, Television, and New Media* (Cambridge University Press, New York, 1997)
26. R. Kleck, P.L. Buck, W.L. Goller, R.S. London, J.R. Pfeiffer, D.P. Vukcevic, Effect of stigmatizing conditions on the use of personal space. Psychol. Rep. **23**(1), 111–118 (1968)
27. E.T. Hall, E.T. Hall, *The hidden dimension* (Anchor Books, New York, 1969)
28. M.J. Hertenstein, D. Keltner, B. App, B.A. Bulleit, A.R. Jaskolka, Touch communicates distinct emotions. Emotion **6**(3), 528 (2006)
29. M.J. Hertenstein, J.M. Verkamp, A.M. Kerestes, R.M. Holmes, The communicative functions of touch in humans, nonhuman primates, and rats: a review and synthesis of the empirical research. Genet. Soc. Gen. Psychol. Monogr. **132**(1), 5–94 (2006)
30. J.J. Edney, C.A. Walker, N. Jordan, Is there reactance in personal space? J. Soc. Psychol. **100**(2), 207–217 (1976)
31. A. Mehrabian, Inference of attitudes from the posture, orientation, and distance of a communicator. J. Consult. Clin. Psychol. **32**(3), 296 (1968)

Chapter 38
A Wearable Walking-Aid System For Visual-Impaired People Based on Monocular Top-View Transform

Qing Lin and Youngjoon Han

Abstract This paper presents a wearable system which can provide walking-aids for visual-impaired people in an outdoor environment. Unlike many existing systems that rely on stereo cameras or combination of other sensors, the proposed system aims to do the job by using just single camera mounted at user's belly. One of the main difficulties of using single camera in outdoor navigation task is the discrimination of obstacles with cluttered background. To solve this problem, this paper makes use of the inhomogeneous re-sampling property of top-view transform. By mapping the original image to a top-view virtual plane using top-view transform, background edges in the near-field are sub-sampled while obstacle edges in the far-field are oversampled. Morphology filters with connected component analysis are used to enhance obstacle edges as edge-blobs with larger size, whereas sparse edges from background are filtered out. Based on the identified obstacles, safe path can be estimated by tracking a polar edge-blob histogram on the top-view domain. To deliver the safe direction to the user, an audio message interface is designed. The system is tested in different outdoor scenes with complex road conditions, and its efficiency has been confirmed.

Keywords Audio guidance · Monocular vision · Obstacle detection · Polar edge-blob histogram · Top-view transform · Wearable walking-aids

Q. Lin · Y. Han (✉)
Electronic Engineering Department, Soongsil University, 511, Sangdo-Dong,
Dongjak-Gu, Seoul
e-mail: young@ssu.ac.kr

Q. Lin
e-mail: lqsdust@163.com

H. K. Kim et al. (eds.), *Transactions on Engineering Technologies*,
DOI: 10.1007/978-94-017-9115-1_38,

1 Introduction

Authoritative statistics have shown that about 1 % of the world population is visually impaired, and among them about 10 % is fully blind. One of the consequences of being visually impaired is the limitations in mobility. Therefore, many electronic travel-aid systems have been developed to provide assistance to blind people in a certain local environment. Electronic travel-aid systems can be categorized depending on how to sense the environment and how to inform the blind user [1]. In general, environment can be sensed through ultrasonic sensor, laser sensor, or camera, and users can be informed via auditory or tactile sense. In recent years, camera based travel-aid systems have won much attentions due to its advantages like large sensing area, rich sensing data as well as low cost.

Most existing vision-based travel-aids systems are developed using stereo vision methods. In these systems, stereo cameras are used to create a depth map of the surrounding environment. The distance information contained in this depth map is then quantized into certain kind of grid representation, which is converted into tactile or auditory sensing modalities so as to be perceived by the visual-impaired user. For Instance, TVS [2] and Tyflos navigator system [3, 4] quantize depth map into a regular grid representation, which is converted into vibration sensing on a 2-D vibration array attached on the user's abdomen. ENVS system [5] quantizes depth map to a rectangular block representation, which maps to electrical pulses that stimulate user's fingers. In [6], depth map is quantized to a polar grid representation, which is transformed into an acoustic sound space.

Although many stereo-vision based travel-aids systems have been proved to be effective under certain environment, some problems still exist. First of all, due to the high computation cost of getting a dense depth map, most of these systems tend to directly convey the quantized depth information to the user without doing any safe path estimation process. As a result, users have to estimate a safe path themselves by sensing and judging the transformed auditory or tactile pattern from the depth map. This makes the system less easy to use and requires much user training. In addition, the accuracy of depth map is largely dependent on stereo matching, which is a challenging task in cluttered outdoor scenes.

Despite stereo-vision based system, systems using only single camera were proposed as well. Compared with stereo cameras, single camera is more compact and easier to maintain. Some of these mono-vision based systems focused on identification of object pixels among background pixels. For example, in [7], region growing segmentation is used to discriminate obstacle pixels with the background. in NAVI system proposed by Sainarayanan et al. [8], a fuzzy learning vector quantization (LVQ) neural network is trained for the classification of object pixels and background pixels. Although the obstacle detection performances for these systems in simple indoor environment are encouraging, their performances may get deteriorate in outdoor environment with various illumination changes and complex background.

In this paper, we present a monocular vision approach to do obstacle detection and safe path estimation for assisting visual-impaired users to pass through a pedestrian path. In the proposed method, a camera is attached at blind user's belly and looks downward to the road in front. The basic idea for obstacle detection is to discriminate obstacle pixels with background pixels. In contrast to Sainarayanan's method that made use of pixel-wise features, edge-based features are explored to discriminate obstacles with road pavement background. By re-sampling the original image pixels and mapping to a top-view virtual plane, clutter edges from background in the near field are suppressed, while obstacle edges in the far field are enhanced. Morphology filters are then used to enhance this inhomogeneous re-sampling effect on connectivity and scale of edges, so that obstacle edge-blobs can be identified easily by connected component analysis. To find a safe path, a polar distortion model of obstacles are built on top-view domain, based on which a polar edge-blob histogram is calculated by scanning all the polar directions to check edge-pixel accumulations that lie on each polar direction. The part of histogram where the biggest valley appears is detected to find the largest area where no obstacle edge-blob exists. Compared with our previous work presented in [9], a joint tracking of polar histogram bin group and bounding polar angle is added to make the safe direction estimation more stable. The safe path directions are further transformed into a clock-face direction set, and delivered to the user via audio messages.

2 Object Detection on Top-View Domain

Top-view transform is in nature an inhomogeneous re-sampling and mapping process. In this section, the re-sampling and mapping process is re-formulated in horizontal and vertical directions, and the re-sampling effect on the scale and connectivity of edges is discussed.

The model of vertical direction re-sampling is illustrated in Fig. 1a. In Fig. 1a, $\boldsymbol{C_r}$ is the real camera center with $\boldsymbol{S_r}$ as its image plane, while $\boldsymbol{C_v}$ is the virtual top-view camera center with $\boldsymbol{S_v}$ as the virtual top-view plane. To figure out the re-sampling relationship between $\boldsymbol{S_r}$ and $\boldsymbol{S_v}$ plane, the only parameters that require are $\boldsymbol{\varphi}$ and $\boldsymbol{\theta}$. By the geometrical description in Fig. 1a, for each point $\boldsymbol{P_v}$ on the virtual top-view plane $\boldsymbol{S_v}$, its corresponding sampling point $\boldsymbol{P_r}$ on the real image plane $\boldsymbol{S_r}$ can be calculated based on their common projection point $\boldsymbol{P_g}$ on the ground plane. As in (1) shows, for each point $\boldsymbol{i}$ on top-view plane, its corresponding sampling point $\boldsymbol{h}$ on the real image plane can be obtained. The model of horizontal re-sampling is illustrated in Fig. 1b, for each row $\boldsymbol{W_k}$ in $\boldsymbol{C_r}$'s field of view on the ground plane, its length can be calculated according to the triangular similarity, and by comparing $\boldsymbol{W_k}$ with $\boldsymbol{C_v}$'s field of view on the ground plane, the sampling ratio can be computed. As in (2) shows, the sampling ratio $\boldsymbol{W_s}$ for each row on the top-view plane can be finally calculated.

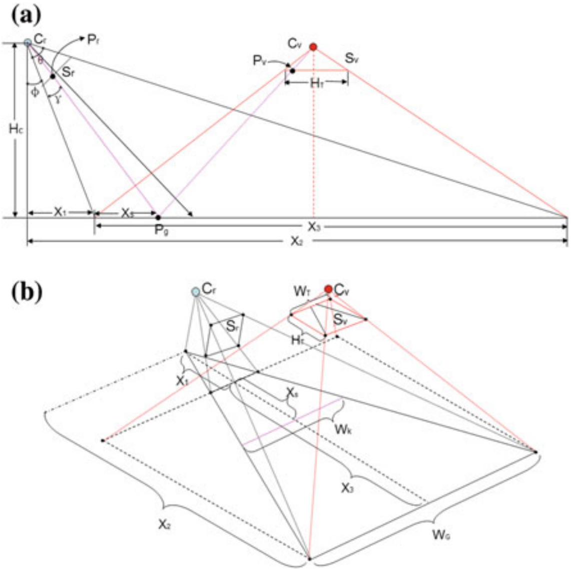


Fig. 1 Top-view re-sampling and mapping model. **a** Vertical direction re-sampling and mapping, **b** Horizontal direction re-sampling and mapping

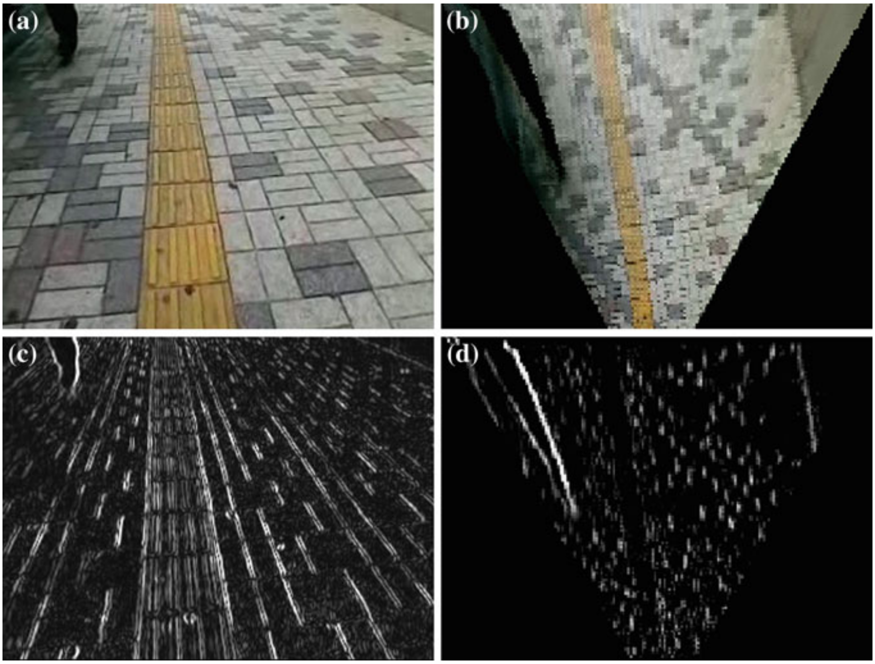


Fig. 2 Effect of top-view re-sampling and mapping. **a** Original image, **b** Top-view image, **c** Original edge map, **d** Top-view edge map

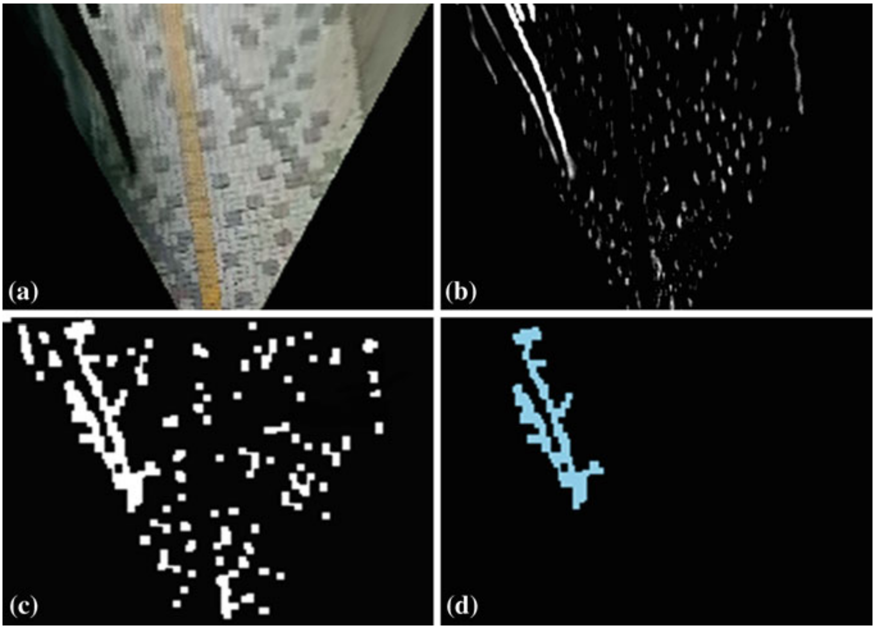
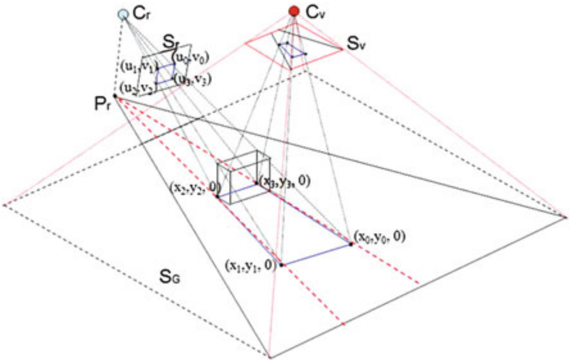


Fig. 3 Edge-blob extraction. **a** Top-view image, **b** Top-view edge map, **c** Morphology filtering, **d** Major edge-blobs

Fig. 4 Obstacle projection model on top-view domain



$$
\begin{aligned}
X_1 &= H_c \cdot \tan\phi, \quad X_2 = H_c \cdot \tan(\theta + \phi) \\
X_3 &= X_2 - X_1, \quad X_s = i \cdot X_3 / H_T (i = 0 \cdots H_T) \\
\gamma &= \arctan[(X_1 + X_s)/H_c] - \phi \\
h &= \begin{cases} H_I/2 - f \cdot \tan(\theta/2 - \gamma)(\gamma \leq \theta/2) \\ H_I/2 + f \cdot \tan(\gamma - \theta/2)(\gamma > \theta/2) \end{cases}
\end{aligned}
\tag{1}
$$

$$\begin{aligned} &W_G = X_3 \cdot (W_T/H_T), \quad X_k = X_1 + X_s \\ &W_k = W_G \cdot X_k/X_2, \quad W_{k\_p} = W_k \cdot (W_T/W_G) \\ &W_s = W_T/W_{k\_p} \end{aligned} \tag{2}$$

Figure 2 shows the effect of top-view mapping by comparing original edge map with top-view edge map. On original edge map in Fig. 2c, obstacle edges are mixed with clutter edges from the background, which makes it difficult to discriminate obstacle's edges with those pavement edges around. However, on top-view edge map in Fig. 2d, the top-view re-sampling process enhances the scale and connectivity of obstacle edges in the far field while suppresses clutter edges in the background. Compared with the edge map of original image in Fig. 2c, it is obviously much easier to discriminate obstacle edges on top-view domain in terms of their scale and connectivity. After top-view re-sampling and mapping, the obstacle edges are enhanced in terms of scale and connectivity. To further emphasize this effect, a combination of morphology operations and connected component analysis is used to extract edge-blobs with large size. These edge-blobs are regarded as candidate obstacle representations.

Here a 3 × 3 rectangular structure element is used to remove pavement edge segments with an opening operation, followed by a closing operation to fill the gaps inside remaining foreground pixels. A connected component labeling operation is then applied to group the connected foreground pixels into blobs. Blobs with size smaller than a pre-defined threshold are discarded. As shown in Fig. 3c, many small edge-blobs from pavement are eliminated by opening operation, and closing operation fixes the shape of foreground blobs. Finally, as shown in Fig. 3d, only two major edge-blobs are selected, which correspond to possible obstacle regions on the top-view.

3 Safe Path Estimation

Due to the top-view re-sampling effect, the shape of a generic obstacle with quasi-vertical boundaries will be distorted on the top-view domain. Here an interesting property of this distortion is that, an obstacle which rises up from the ground surface would be elongated in the direction of an imaginary connection line joining the camera's perpendicular projection on the ground and the base point of the obstacle in top-view images, as is shown in Fig. 4.

This geometric property means that, obstacle edges should also lie along this connection line on top-view image. Therefore, vertical obstacle's edges should lie on series of radial orientations with respect to the camera's projection point on the top-view plane. This vertical line distortion can be partly explained by the inhomogeneous re-sampling process, while it can also be derived from formula on inverse perspective mapping [10]. As is shown in (3), the point on the real image plane $\boldsymbol{S_r}$ is represented in (u, v), and point on the ground plane $\boldsymbol{S_G}$ is represented by

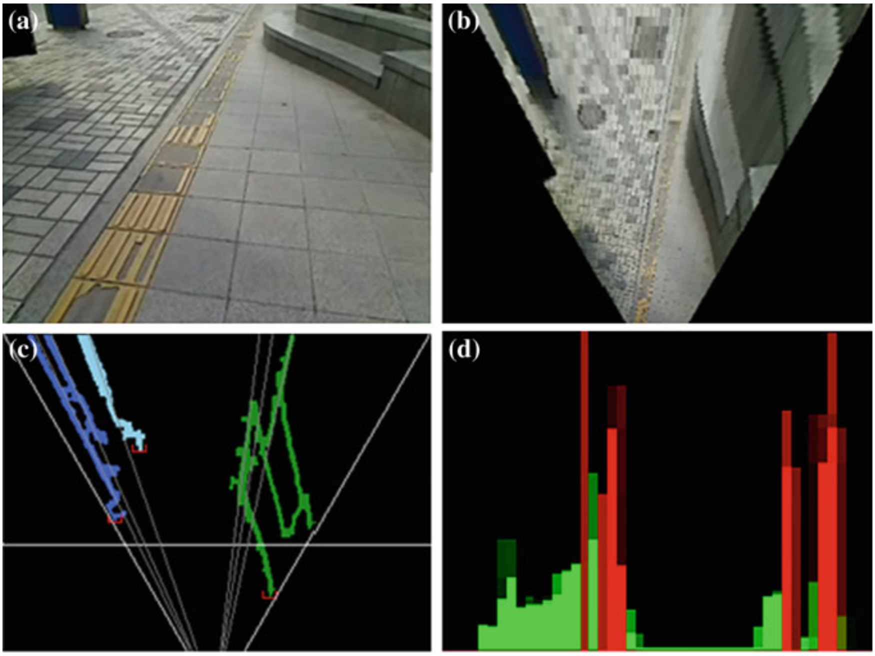


Fig. 5 Safe path estimation through polar edge-blob histogram. **a** Original image, **b** Top-view image, **c** Safe-area estimation, **d** Polar edge histogram

(*x*, *y*, *0*). Vertical lines on the image plane S_r can be represented by $v = k$, while *k* is a constant value, substituting this into (3), we can get (4), where c_1 and c_2 are constant terms. Finally, we can obtain (5), where (*l*, *d*) represents camera center's projection point P_r on the ground plane.

$$\begin{cases} x(u,v) = h \times \cot[(\theta - \partial) + u\frac{2\partial}{n-1}] \times \cos[(\gamma - \partial) + v\frac{2\partial}{n-1}] + l \\ y(u,v) = h \times \cot[(\theta - \partial) + u\frac{2\partial}{n-1}] \times \sin[(\gamma - \partial) + v\frac{2\partial}{n-1}] + d \end{cases} \tag{3}$$

$$\begin{cases} x = h \times \cot[(\theta - \partial) + u\frac{2\partial}{n-1}] \times c_1 + l \\ y = h \times \cot[(\theta - \partial) + u\frac{2\partial}{n-1}] \times c_2 + d \end{cases} \tag{4}$$

$$y - d = (c_2/c_1)(x - l) \tag{5}$$

Based on the obstacle projection model on top-view domain, here a polar edge histogram is constructed on the top-view plane for the estimation of safe path. As is shown in Fig. 5c, on the edge map, from the right boundary to the left boundary, polar directions are sampled with respect to the convergence point C, which corresponds to camera's perpendicular projection point on the ground plane. For each sampled polar direction, the number of edge-blob pixels that lie along this direction is counted. By accumulating all the sampled polar directions, a polar edge-blob histogram can be constructed as shown in Fig. 5d.

In polar edge histogram, the horizontal axis represents sampled polar directions in angles, and the vertical axis is the number of edge-blob pixels that lie along each sampled direction angle. The bins with high values indicate the directions where obstacles appear, while bins with zero values correspond to the directions where no obstacles exist. Therefore, safe-area should be estimated by the bins with zero values.

4 Safe Path Tracking and Delivery

Since the camera is mounted on user's body, the camera will show some swing motions due to the movement of human body. These swing motions will bring additional noise to the safe path estimation. To estimate the safe path more steadily, the largest valley bin group on polar edge-blob histogram is tracked.

A flowchart of safe path tracking is shown in Fig. 6. For tracking initialization, consecutive zero-value bins in frame ***t*** are grouped and sorted according to their group size. Then the largest bin group is selected as the tracking group in frame ***t***. In the following frame, the zero bin group that is closest to the tracking group in frame ***t-1*** is selected as the tracking group in frame ***t***. If the size of tracking group is smaller than a threshold, then tracking will be stopped and re-initialized from the beginning.

In addition to tracking on polar edge-blob histogram, the polar angles of the safe area boundaries on top-view domain are also tracked by means of a Kalman filter. When mapping the polar histogram to top-view domain, the histogram bin group under tracking can be represented by two bounding polar angles. The solid curve in Fig. 7 shows the measured value of the left bounding polar angle, this noisy value pattern is mainly caused by camera's shaking motion with user's rolling gait. The noise involved in this pattern can be approximated by Gaussian noise. Therefore, one-dimensional Kalman filters are used to find the stable estimation. The filtered value is shown by the dash curve in Fig. 7.

To delivered the safe path direction to blind user, the estimated directions are transformed into a "clock-face" direction representation. Clock face directions are broadly accepted as a common way to indicate directions in the blind people community. As in Fig. 8 shows, the safe path is mapped from top-view domain to original-view domain, which is divided by projecting the top-half clock face area (10 o'clock–2 o'clock) onto the center horizontal line in original image space. The center of the mapped free path on the center line is defined as the safe direction indicator. The clock face section it falls into determines the safe direction to be delivered to the user in form of audio messages as : "go 10 o'clock direction". The audio message is delivered to the user through a loudspeaker whenever the estimated direction changes from one clock face section to another.

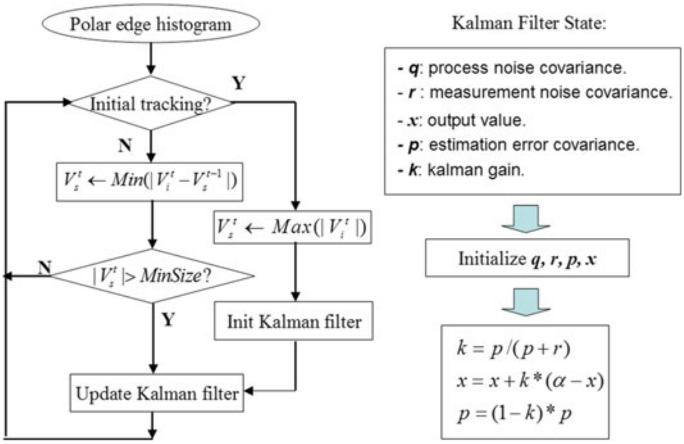


Fig. 6 Safe path tracking flowchart

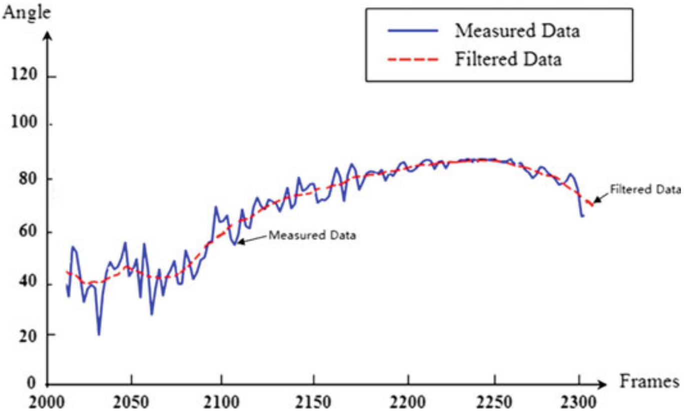


Fig. 7 Safe path direction tracking result

5 Experimental Results

To test the performance of the algorithm, we attached a camera on a belt and fix it at user's waist, pointing a little bit downward to the road ahead of user, as is shown in Fig. 9. The camera is simply a Logitech webcam, which captures color image at 320 × 240 resolution in 30 fps. The safe path algorithm is implemented using Visual C++ under MS Windows platform, which runs on a laptop computer with 1.8 GHZ CPU and 2 GB DDR memory. The webcam captures images of the road environment, and then processed by the path finding software which runs at the laptop computer carrying in user's backpack.

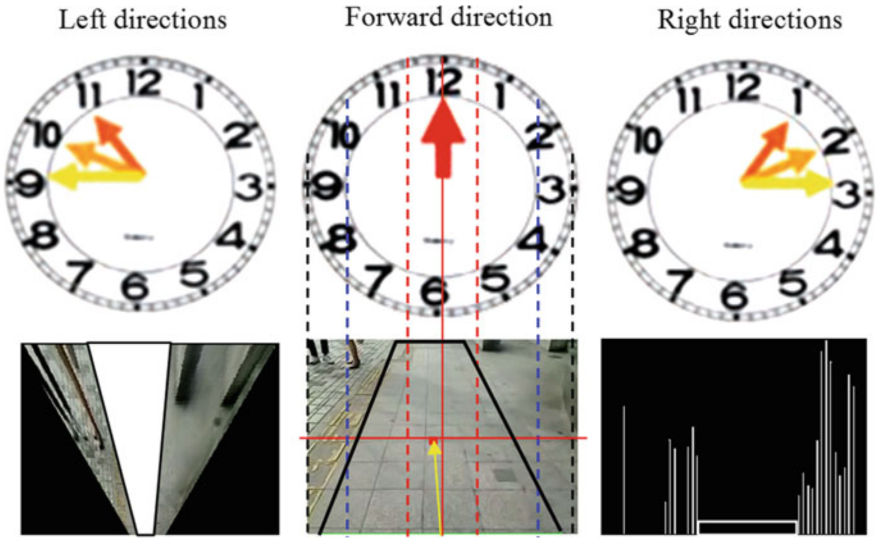


Fig. 8 Clock face direction representation

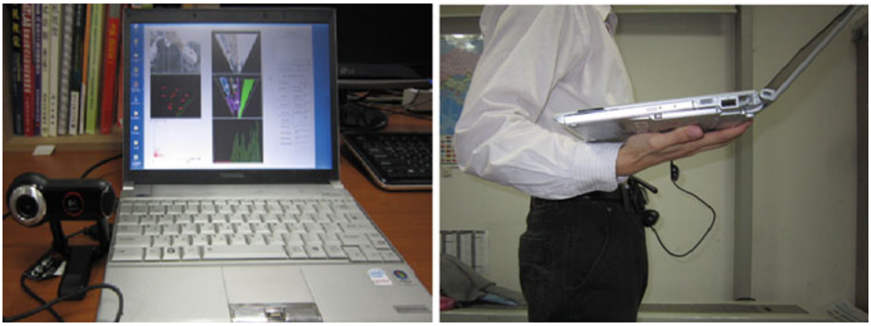


Fig. 9 Wearable prototype system

To make a top-view mapping, camera's downward viewing angle and its angular aperture should be measured. By using these two parameters and applying formula (1) and (2), a mapping table is made to store the mapping relationship from one point on the top-view domain to its corresponding location on the original-view domain. With this mapping table, top-view re-sampling and mapping can be done very efficiently.

The algorithm is tested on several outdoor pedestrian path scenes, with various roadside structures and cluttered road surface. A test scene sample is shown in Fig. 10, where processing results on original view and top-view are compared. In Fig. 10a, edge map on original image is examined by a vertical projection histogram. Due to the influence of clutter edges from the pavement, it is very difficult to

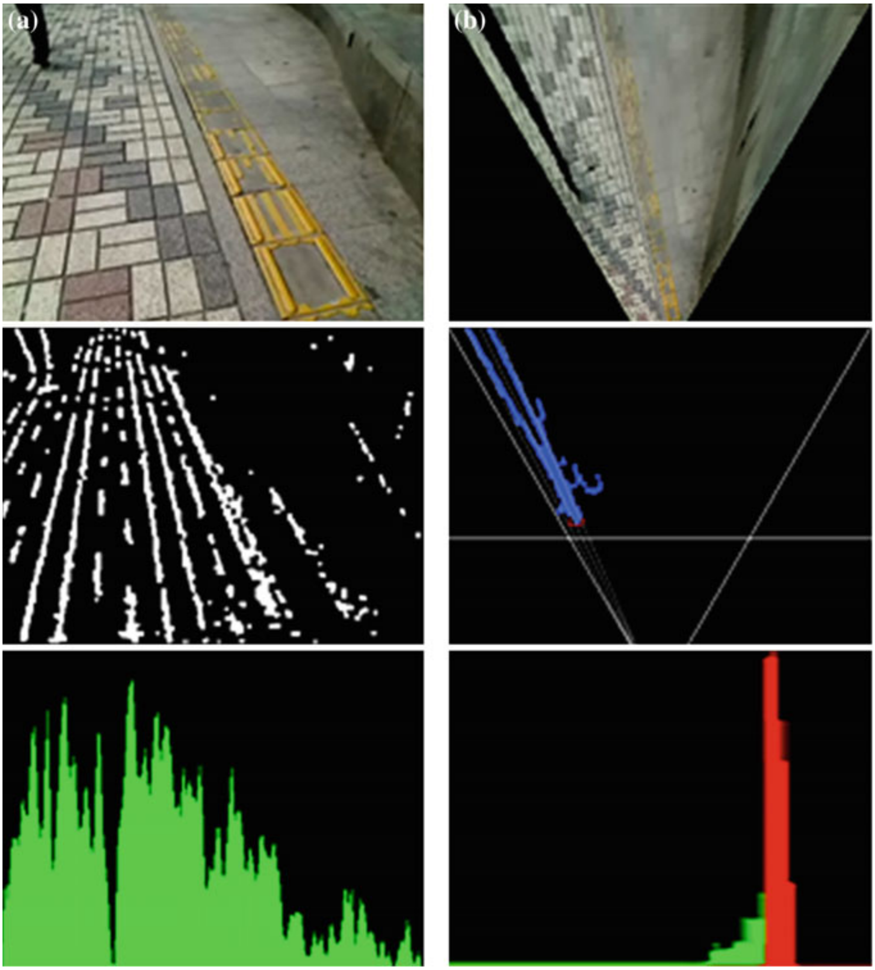


Fig. 10 Original-view and top-view processing comparison. **a** Original view processing, **b** Top-view processing

discriminate a person located in the upper-right corner with background clutters in the image. In contrast, clutter edges are suppressed while obstacle edges are enhanced on the top-view edge map shown in Fig. 10b, and the location of the person can be clearly identified on the polar edge-blob histogram.

To evaluate obstacle detection performance, the test scenes are divided into three sets, including pedestrian paths in open space scene, park scene and urban scene, some sample images from these three different test sets are shown in Fig. 11. All the critical obstacle positions are manually labeled on the top-view images of these test sets. A true positive (TP) detection is defined to be such that the detection corresponds with an actual obstacle, and the deviation should not

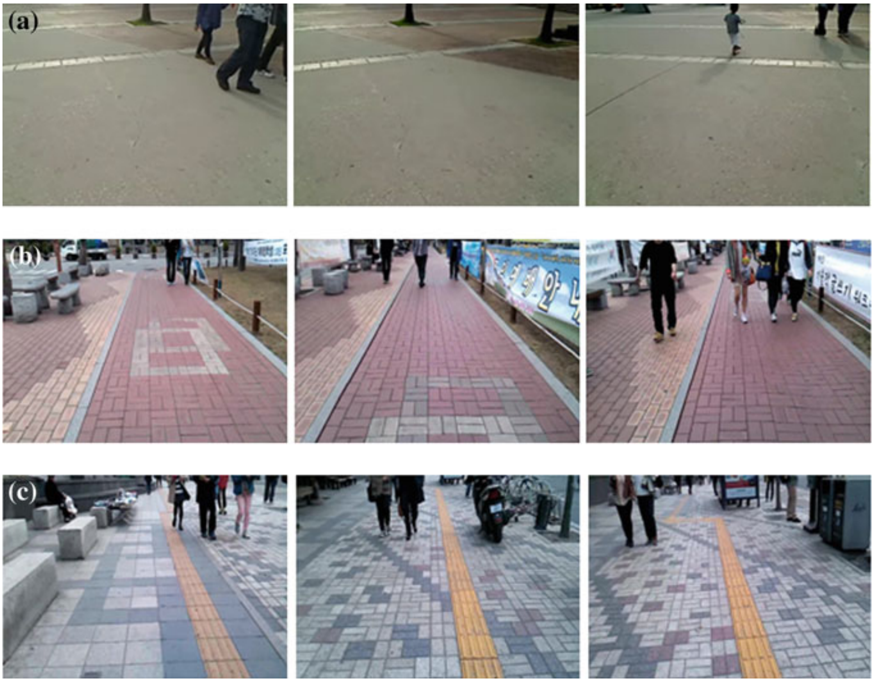


Fig. 11 Sample images from test sets. **a** Open space test set, **b** Park test set, **c** Urban test set

Table 1 Obstacle detection rate

Test sets	Obstacle	TP	FP	FN
Urban	365	314	32	12
Open	278	263	11	5
Park	212	193	13	8

exceed 20 % of the obstacle's size, otherwise it is considered as a false positive (FP), obstacles that have not been detected is false negative (FN).

Table 1 shows the detection results on three test sets. For safe path estimation, it is very critical to control the false negative rate for sake of safe navigation. Therefore, during testing, the algorithm parameters are tuned in such a way to achieve an acceptable true positive rate while keeping false negative rates as small as possible.

Since the proposed algorithm relies on radial distribution of edges on top-view domain, when strong background edges appear in similar radial patterns with that of obstacles on top-view, they may give rise to FP cases. Moreover, small planar obstacles in the near field may be sub-sampled heavily on top-view, which makes it difficult to discriminate with ground clutters. Therefore, small holes or stones on cluttered road surface may not be properly detected, which give rise to FN cases.

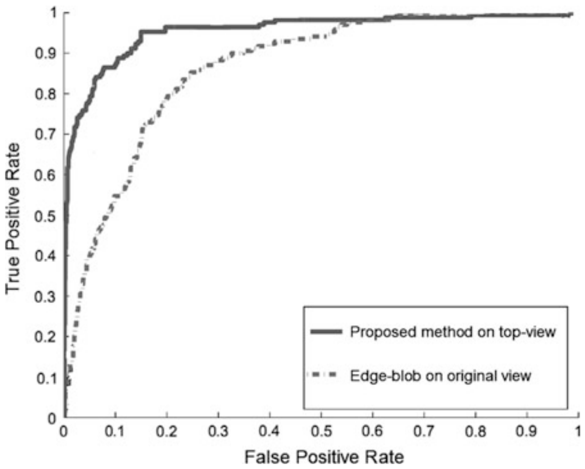


Fig. 12 ROC curve of top-view and original-view methods

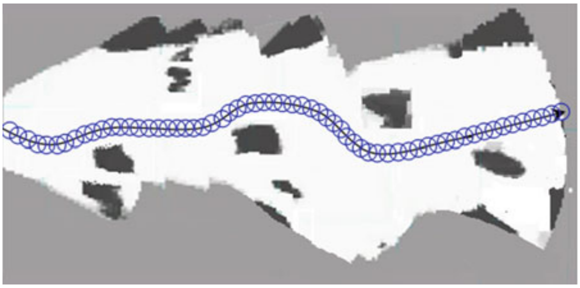


Fig. 13 Simulated walking trajectory on occupancy map

In the test, open space set achieves a high TP rate of 94.6 %, as this set involves mainly vertical obstacles like pedestrians, and less cluttered road surface. While in urban set, only 86 % TP rate is achieved, due to highly cluttered road surface as well as many planar obstacles in small size.

To show the effectiveness of the proposed algorithm on top-view domain, an obstacle detection method [11] using edge-blobs on original view is implemented and tested on the urban test set. The quantitative comparison between the two algorithms is shown in Fig. 12. The ROC curves are generated by varying the obstacle edge-blob extraction threshold in both algorithms. It can be observed that the proposed method on top-view has shown much better performance under complex background.

To evaluate the safe path estimation performance, simulated user walking trajectory is generated by using the estimated safe walking direction and user's walking speed. User's walking speed is recorded by using an inertial sensor attached on user's body. This simulated walking trajectory is then mapped to a top-view occupancy map generated from obstacle detection module. A segment of this synthesized map is shown in Fig. 13, which is generated by walking on a pedestrian path around our campus. Based on the simulation results, it is found that simulated walking trajectory is able to avoid salient obstacles with vertical edges However, as small planar obstacles in the near distance may be removed together with the clutter patterns from the pavement. It may cause problems for safe path estimation in this case. Under the experiment platform condition, the path finding algorithm can run at an average of 16.4 fps, which is fast enough to satisfy the real-time requirement of human navigation task.

6 Conclusion and Future Work

In this paper, a wearable walking-aid system for guiding the visual-impaired people in outdoor environment is proposed. Rather than using stereo cameras, the proposed system handles this problem with just single camera, which makes the system more portable and easy to configure. Compared with other single camera solutions, the proposed method takes advantage of a top-view re-sampling process to suppress and eliminate background edges. And by modeling obstacle projections on top-view domain, safe path can be estimated steadily by means of a polar edge histogram. The proposed algorithm can work efficiently in an outdoor side-walk environment, and provide valuable information to the visual-impaired user. Future work would include improving the audio user interface to make the prototype system work in a more user-friendly way.

Acknowledgment This research is supported by Next-Generation Information Computing Development Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science and Technology (No. 2012M3C4A7032182)). This work (Grants No. C0119991) is also supported by Business for Cooperative R&D between Industry, Academy, & Research Institute funded by Korea Small and Medium Business Administration in 2013.

References

1. D. Dakopoulos, N.G. Bourbakis, Wearable obstacle avoidance electronic travel aids for blind: a survey, IEEE Trans. Syst. Man Cybern. **40**(1), 25–35, (2010)
2. L.A. Johnson, C.M. Higgins, A navigation aid for the blind using tactile-visual sensory substitution, in Proceedings of 28th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 6268–6292, NewYork, 2006
3. N. Bourbakis, Sensing 3D dynamic space for blind. IEEE Eng. Med. Biol. Mag. **27**(1), 49–55 (2008)
4. D. Dakopoulos, S.K. Boddhu, N. Bourbakis, A 2D vibration array as an assistive device for visually impaired, in *Proceedings of 7th IEEE International Conference Bioinformatics and Bioengineering*, vol. 1 (Boston, 2007), pp. 930–937
5. S. Meers, K. Ward, A substitute vision system for providing 3D perception and GPS navigation via electro-tactile stimulation, in *Proceedings 1st International Conference Sensing Technology*, pp. 21–23, Palmerston North, 2005
6. A. Rodríguez, J.J. Yebes, P.F. Alcantarilla, L.M. Bergasa, J. Almazán, A. Cela, Assisting the visually impaired: obstacle detection and warning system by acoustic feedback. Sensors **12**(12), 17476–17496 (2012)

7. A. Hub, J. Diepstraten, T. Ertl, Design and development of an indoor navigation and object identification system for the blind,in *Proceedings of ACM SIGACCESS Accessibility Computing*, pp. 147–152, New York, Jan 2004
8. G. Sainarayanan, R. Nagarajan, S. Yaacob, Fuzzy image processing scheme for autonomous navigation of human blind. Appl. Softw. Comput. **7**(1), 257–264 (2007)
9. Q. Lin, Y. Han, Safe path estimation for visual-impaired people using polar edge-blob histogram, in *Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering and Computer Science*, WCECS 2013, pp 401–405, San Francisco, 23–25 Oct 2013
10. M. Bertozzi, M. Broggi, GOLD: a parallel real-time stereo vision system for generic obstacle and lane detection. IEEE Trans. Image Proc. **7**(1), 62–81 (1998)
11. J.H. Yu, H.I. Chung, H.S. Hahn, Walking assistance system for sight impaired people based on a multimodal transformation technique, in *Proceedings of ICROS-SICE International Joint Conference*, pp. 1639–1643. Tokyo, Aug 2009

Chapter 39
Design and Characterization of a Model Fatigue Testing Machine for Academic Laboratories

Austin Ikechukwu Gbasouzor, Clementina Ogochukwu Okeke and Lazarus Onyebuchi Chima

Abstract Many engineering machines and mechanical components are subjected to fluctuating stresses, taking place at relatively high frequencies and under these conditions failure is found to occur. This is "fatigue failure". And this led to the invention of a fatigue testing machine. In view of effective design that will not fail accidentally, this research is conceived. This testing machine will determine the strength of materials under the action of fatigue load. Specimens are subjected to repeated varying forces or fluctuating loading of specific magnitude while the cycles or stress reversals are counted to destruction. The first test is made at a stretch that is somewhat under the ultimate strength of the material. The second test is made at a stress that is less that than that used in the first. The process is continued, and results are plotted.

Keywords Fatigue failure · Fatigue load · Fluctuating stresses · High frequencies · SN diagram · Strength of materials

A. I. Gbasouzor (✉)
Department of Mechanical Engineering, Anambra State University, P. M. B. 02, Uli, Nigeria
e-mail: unconditionaldivineventure@yahoo.com

C. O. Okeke
Department of Computer Science, Anambra State University, P. M. B. 02, Uli, Nigeria
e-mail: ogoookeke@yahoo.com

L. O. Chima
Department of Mechanical Engineering, Nnamdi Azikiwe University, P. M. B. 5025, Awka, Nigeria
e-mail: chimalaz@yahoo.com

H. K. Kim et al. (eds.), *Transactions on Engineering Technologies*,
DOI: 10.1007/978-94-017-9115-1_39,

1 Introduction

Fatigue is a failure of material or machine due to the action of repeated or fluctuating stress on a machine member for some number of times.

This failure begins with a small crack. The crack will develop at a point of discontinuity in the material such as a change in crosssection, a keyway, a hole or a notch. Less obvious points at which fatigue failure are likely to begin are inspection or stamp marks, internal cracks or even irregularities caused by machining. Once a crack is initiated the stress concentration effect becomes greater and the crack progresses more rapidly. As the stressed area decreases in size, the stress increases in magnitude until finally, the remaining area fails suddenly. A fatigue failure therefore is characterized by two distinct regions. The first of those is due to the progressive development of crack, while the second is due to sudden fracture.

Unlike other failures, fatigue failure gives no visible warning in advance. It is sudden and totally dangerous. This sudden failure, which is dangerous, and can lead to not just minor accident but fatal accident and loss of lives triggered the quest to invent a machine that can test and give or predict the effect of fatigue on various metals such as aluminium, cast-iron, mild steel, etc.

In most testing of those properties of materials that relate to the stress–strain diagram, the load is applied gradually to give sufficient time for the strain to fully develop. Furthermore, the specimen is tested to destruction, and so the stresses are applied only once. Testing of this kind is applicable, then, to what are known as static conditions. Such conditions closely approximate the actual conditions to which many structural and machine members are subjected.

The condition frequently arises, however, in which the stresses vary or they fluctuate between levels. For example, a particular fibre or aluminium on the surface of the rotating shaft subjected to the action of bending loads undergoes both tension and compression for each revolution of the shaft. Since the shaft is part of an electric motor rotating at 1,400 rev/min, the aluminium metal is stressed in tension and compression 1,400 times each minute. If in addition, the shaft is also axially loaded (as it would be, for instance, by a helical or worm gear), an axial component of stress is superposed upon the bending component.

In this case, some stress is always present in any one metal, but now the level of stress is fluctuating. These and other kinds of loading occurring in machine members produce stresses that are called variable, repeated, alternating or fluctuating stresses.

Often, machine members are found to have failed under the action of repeated or fluctuating stresses; yet the most careful analysis reveals that the actual maximum stresses were below the ultimately strength of the material, and quite frequently even below the yield strength. The most distinguishing characteristic of these failure is that the stresses have been repeated a very large number of times. Hence, the failure is called “Fatigue Failure”.

When machine parts fail statically, they usually develop a very large deflection, because the stress has exceeded the yield strength and part is replaced before fracture actually occurs. Thus, many static failures give visible warning. It is sudden and total, and dangerous. It is relatively simple to design against a static failure because our knowledge is comprehensive. Fatigue is a much more complicated phenomenon, only partially understood, and the engineer seeking competence must acquire as much knowledge of the subject as possible.

1.1 Design Objectives

This aims at designing and constructing a fatigue-testing machine that is capable of testing the fatigue life of various samples of specimen from metals, such as mild steel, aluminum, brass, etc.

Also with the result of each test carried out with this machine, the fatigue life of various materials can be obtained and fatigue failure be guarded against in an optimum manner.

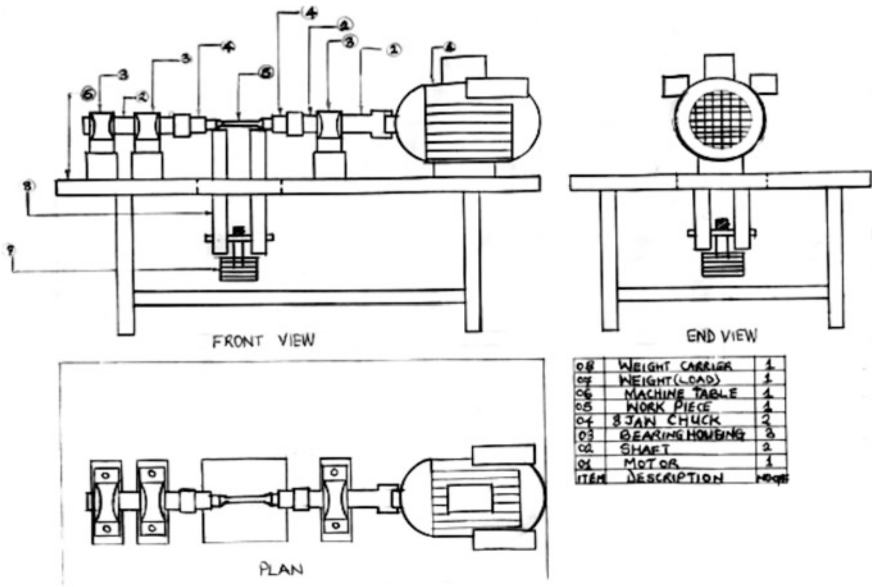
2 Methodology

2.1 General Description

The fatigue-testing machine is of the rotating beam type. The specimen functions as a single beam symmetrically loaded at two points. When rotated one-half revolution the stress in the fibres originally above the neutral axis of the specimen are reversed from compression to tension for equal intensity. Upon completing the revolution, the stresses are again reversed, so that during one complete revolution the test specimen passes through a complete cycle flexural stress. The fatigue testing machine consists of the following components: A. 4 HP electric motor, B. Bearing and its housing assembly C. Weight hanger assembly, D. Dead weight, E. Bearing spindling, F. Digital counter, G. Magnetic cyclic pick up (dynamo), H. Variable speed control I. Switch, J. Specimen, K. Drill chunks, L. The metal desk (Fig. 1).

2.2 Bearing and Its Housing Assembly

The particular bearing used in this project is the single, row, deep; groove bearing that can take both radial load and some thrust load. The bearing is shouldered in a housing made of cast iron to secure adequate support for the bearing and resist the maximum thrust load.



08	WEIGHT CARRIER	1
07	WEIGHT(LOAD)	1
06	MACHINE TABLE	1
05	WORK PIECE	1
04	3 JAW CHUCK	2
03	BEARING HOUSING	3
02	SHAFT	2
01	MOTOR	1
ITEM	DESCRIPTION	NOS

Fig. 1 The isometric views of the fatigue-testing machine

2.3 Weight Hanger Assembly

This consists of small bar made of cast iron, with a spring. The head of this bar is designed to be hanged on the loading hardness assembly. The spring is designed to absorb shock preventing any slight vibration of the housings on the dead weight. The choice of cast iron for the material for hanger is due to the fact that cast iron has the modulus of elasticity E as 60–90 MPa and can withstand the maximum weight of which the machine can carry.

2.4 Dead Weight

These are loads of different sizes and weight.

2.5 Bearing Spindle (Shaft)

The bearing spindle (shaft) are machined from a bar of high grade steel (310 stainless steel) and is thicker than specimen by (5/8 in.) to prevent the machine parts from experiencing fatigue due to the loading of the specimen. The stainless steel can withstand adverse temperature and corrosive environment, due to its high carbon content it does not fail-easily.

2.6 Digital Counter

The digital counter is a terminating device that gives out the equivalent value of number of revolution of motor in term of cycles. The digital counter registers 1 for each 140 revolution of the motor, so to obtain the number of cycles of stress, counter readings must be multiplied by 10,000. A total of 1,400,000 revolutions are turned up for each cycle of number. At 1,400 rpm, a complete cycle of numbers takes place in about 166 h, or 7 days if operating 24 h per day. In a very long test therefore the number of cycle must be recorded. The speed reduction unit is lubricated from time to time. This easy-to-read digital cycle counter is connected to a magnetic pick up device (dynamo). Push button control is provided to reset the display count at the start of a test.

2.7 The Cut-Off Switch

The cut-off switch is employed to set upon the circuit to the motor when the bearing experiences a 3° angle.

2.8 Specimens

The standard specimen, a tapered-end specimen has the length of 87.3 mm. They are machined to match the tapers within the spindles of the machine and held in place using drill-chuck. Stress as applied to the specimen by direct application of dead weight to ensure precise loading. Maximum fibre stress in a specimen having a 0.300 in. (7.62 mm) diameter is. By decreasing the diameter, the value of the maximum fibre stress can be increased. An easy-to-use reference table within the operator’s manual makes determination of the load weight needed to produce a particular stress a single circulation.

2.9 Drill Chucks

The metal desk on which the system is assembled on is made with 3 mm angle iron of mild steel and plain pan made of mild steel for the top. The length of the desk is 760 mm; the width and the height are respectively 210 and 190 mm. The desk is drilled at each point where parts are to fit with bolt and nut. A rectangular opening is found through which the hanger is passed; this is located at the top of the desk. The length and width of this opening are respectively 14.7 and 11.8 mm. The total weight of the desk is 75 kg.

2.10 Setting Up the Machine

The base should first of all be set leveled so that the weight will hang perpendicularly to the axis of the specimen. Wires in conduct from the power supply are run to the connecting block in the base of the machine and soldered to the plugs provided. The machine is equipped to operate from an AC power source at 240 volts, 50 cycles, and single phase. Motor and relay equipment to operate from a power source of different rating can be provided on special order. With mild steel legs, the machine is set on five rubber tyres. The weight hanger is of such length that it will clear the desk since it is mounted on the legs. A shock absorber in the hanger prevents the slight vibration of the housing from being impacted to the weight. The base is provided with holes so it can be bolted down if desired.

Assemble the housing with a sample of specimen provided in accordance with the direction given below, and start the machine to see if there is any misalignment. This will be indicated by noise and vibration. Examine the housing wind for oil. It should stand about midway of the windows when the machine is idle.

Adjust the speed to 1,400 rpm by means of a rheostat in the base of the machine. The machine should now be ready for a test.

2.11 Speed

The normal speed is 1,400 rpm. The speed depends on the voltage. If the voltage of the supply line exceeds the proper amount of sliding rheostat in the base of the machine should be adjusted so that the speed cannot exceed 1,400 rpm under ordinary conditions. Speeds in excess of the normal will cause no damage unless continued to long. In adjusting the speed take care that the contraction of the rheostat makes firm contact with the exposed potion of the winding.

If it desired to run at speed below 1,400 rpm, or gradually increase the speed from zero, a variable resistor can be placed in the motor lead wires by utilizing the plug type connector. This will be necessary only in exceptional cases where the slight starting torque exerted on the specimen may be objectionable. The inertia of the motor armature is such in comparison with that of the remote half of the spindle that there is sufficient delay in starting for most purposes.

3 Analysis and Discussion of Results

3.1 Specimen Dimension and Machine Capacity

For the solid specimen, the diameter of the test specimen is 7.62 mm.It is assumed that the machine would at times be required to test some high strength steel at stress levels in the neighborhood of the yield stress. Using stainless steel of 310 i.e. 45c steel at temperature of about 315 °C, ultimate tensile yields stress = 700 MPa.

Assuming Tresca's Yield Criterion, the shear yield stress is given as 350 mpa.

The torque required to start yield at the outer fibres of a specimen of such steel is $T = \tau_y \times \pi d^3$

$$\mathrm{T} = 350 \times \pi \frac{(7.62)^3}{16} = 30,406.25\,\mathrm{N\,mm}$$

∴ In this design, a machine capacity of 30,400 N mm is assumed.

3.2 The Torque-Transmitting Shaft

The torque-transmitting shaft, which comprises of the longer shaft, bolts, the chuck and the short shaft passing through bearing housing sustains the same fatigue load during a testing program as the specimen undergoing the tests, where T = torque = 30406.25 N mm

i.e. d = [16 × 30406.25/π × 87.5] = 12.09 mm

In the design, the minimum shaft diameter used is 25 mm. The reason is to make the shaft more rigid and to make it withstand heat and shock for many number of tests without failing easily due to fatigue.

3.3 Check for Stress Concentration Effect

At the position for the chuck, there is a possible or product stress raiser. It is the shaft shoulder.

3.4 Shaft Shoulder

$$\mathrm{D} = 25\,\mathrm{mm} \qquad \mathrm{d} = 16.2\,\mathrm{m} \qquad \mathrm{r} = \text{maximum fillet radius} = \ 5\,\mathrm{m}$$

$$\therefore \mathrm{D/d} = 25/16.2 = 1.54\,\mathrm{mm}$$

$$\mathrm{r/d} = 5.0/16.2 = 0.31\,\mathrm{m}$$

The stress concentration corresponding to the ratio above is 1.17

i.e. kt = 1.17

But the maximum nominal stress in sheer = $16\mathrm{T}/\pi\mathrm{d}^3$ = $16 \times 30406.25/\mathrm{Ti} \times (16.2)3 = 36.42\,\mathrm{N/m}^2$.

It is designed that no part of this shafting ever fails in service. With the little knowledge of torsion fatigue, we apply therefore factor of ignorance to some considerations.

A stainless steel of 310 or 45c8 steel is recommended for the shafting because of its strength and ability to prevent the machine parts from experiencing fatigue due to the loading of the specimen all through the test. This reason of choosing this material can be traced to the advantage of being highly resistance to fatigue owing to its low notch sensitivity, which is itself, a result of its high ductility.

45c8 steel or 310 stainless steel has these parameters;

$$\text{Yield strength } T_y = 350\,\text{mpa}$$

$$\text{Ultimate strength } S_{ut} = 700\,\text{mpa.}$$

Following a generally accepted thumb rule for $S_{ut} < 19{,}600$ Endurance limit Se' = $0.5S_{ut}$

$$\text{Endurance limit Se'} = 0.5 \times 700 = 350\,\text{mpa}$$

Using Tresca's Yield Criterion and assuming the validity of its application of fatigue loading, the endurance limit in shear is obtained as $T_e = S_e/2 = 175$ mpa.

At this point a factor of approximately 2 is applied and

$$T_e = T_e/2 = 175/2 = 87.5\,\text{mpa}$$

The minimum shaft diameter for the full machine load is given by the expression $d = \{16T/\pi T_d\}^{1/3}$

Applying the stress concentration factor,

$$T_{mass} = 1.17 \times 36.42 = 42.616$$

This value is very much less than the assumed endurance limit of 350 N/mm^2 and also than the design shear stress of 87.5 N/mm. It is therefore considered safe.

3.5 Chucks

Each chuck consists of two major parts, the hub made of mild steel and a circular plate of hardened medium carbon low alloy steel machined to take the end of the specimen. The two parts are fitted together with two aligning pins and four 5 mm diameter SAE grade 2 screws. The hub is shrink-fitted onto the shaft whose diameter is stepped up to 40 mm through a gene ales radius as shown in order to improve the fatigue strength of the joint.

3.6 Analysis of Shrink Fit

When cylinder is assembled by shrinkage as shown above, a contact pressure is developed at the interface. The pressure is a function of the digital interference and for materials of the same modulus E, is given by:

$$\mathrm{P} = \mathrm{E}\delta/\mathrm{b}\left[\left(\mathrm{a}^2 - \mathrm{b}^2\right)/2\mathrm{b}^2 - \left(\mathrm{b}^2 - \mathrm{a}^2\right)/\mathrm{c}^2 - \mathrm{a}^2\right]$$

For the case of shrinkage a hub onto a solid shaft, (a = 0) the expression reduces to:

$$\mathrm{P} = \mathrm{E}\delta/2\mathrm{b}\left[1 - \mathrm{b}^2/\mathrm{c}^2\right]$$

In the design b = 10 mm, c = 16.2

$$\mathrm{P} = 200 \times 10^3\delta/2 \times 10[1 - 10/16.2] = 10,000[1 - 0.617] = 3830\delta\ \mathrm{N/mm}^2$$

An interface of approximately 0.00762 or 7.62×10^{-3} mm is recommended. This will result in an interface pressure of

$$\mathrm{P} = 3830 \times 7.62 \times 10^{-3}\,\mathrm{mm} = 29.19\,\mathrm{N/mm}.$$

This value of stress is safely below the endurance limit and the design stress,

Total area on which interface pressure acts = π × 40 mm = 125.66 mm^2

∴Total effective normal force = 125.66 × 29.19 = 3668.02 N

Assuming a low value of the coefficient of static friction say $\mu = 0.08$, Limiting tangential force = 0.08 × 3668.02 = 293.44 N.

3.7 Short Shaft or Driven Shaft

3.7.1 Determination of Bending Stress

2nd moment of inertia $\mathrm{I} = \pi\mathrm{d}^4/64 = \pi/64 \times (25)^4 = 19174.76\,\mathrm{mm}^4$

Using the relation $\delta_b = (\mathrm{M/I})\ \mathrm{y}$

where $y = \frac{d}{2} = 12.5\,\mathrm{mm}$

$$\mathrm{M} = 30406.25\,\mathrm{N\,mm}$$

The maximum bending stress is obtained as

$$\delta = \frac{30406.25 \times 12.5}{19174.76} = 19.82\,\mathrm{N/mm}^2$$

The maximum bending stress is also safe for the machine.

Table 1 Angular twists of the various segments of the shafting

Shaft diameter d (mm)	Length, l (approx) (mm)	Twist angle θ (rads)
25 longer shaft	150	1.4861×10^{-3}
25 short shaft	70	6.93504×10^{-4}
16.2 (gauge length for long shaft)	40	2.248×10^{-3}
16.2 (gauge length for short shaft)	40	2.248×10^{-3}
7.62 specimen	87	9.986×10^{-2}

3.8 Deflection of Driven Shaft Under Peak Load

The deflection is given by

$$\delta_b = \mathrm{M\,L}^2/3\mathrm{EI}$$

where L is the effective length = 150 mm

$$\delta_b = \frac{30406.25 \times (150)^2}{3 \times 200 \times 10^3 \times 19174.76} = 0.0575\,\mathrm{mm} \approx 0.06\,\mathrm{mm}$$

Therefore angular displacement = 0.06/180 = 3.96×10^{-4} rads.

This displacement is also negligible compared with the relative displacement of the ends of the torque-transmitting shaft under peak load. The analysis shows that the setting of the double eccentric shaft is accurately reflected by the angular twist of the shafting with negligible error; introduced by the elastic bending of both the torque arm and the crank, where l and d are in millimeters. The angular twists of the various segments of the shafting are computed, using the approximate effective length. The results are expressed in the Table 1.

From the table, angular twist is obtained by summing the angular twist of the separate segments.

$\sum\theta$ approximate = 0.106 = 0.11 rads.

3.9 Determination of Bending Stress

2nd moment of inertia $\mathrm{I} = \pi \mathrm{d}^4/64 = \pi/64 \times (25)^4 = 19174.76\,\mathrm{mm}^4$

Using the relation δ = (M/I) y

where $y = \frac{d}{2} = 12.5\,\mathrm{mm}$

$$\mathrm{M} = 30406.25\,\mathrm{N\,mm}$$

The maximum bending stress is safe for the machine.

3.10 Deflection Under Peak Load

The deflection is given by

$$\delta_b = \text{M L}^2/3\text{EI}$$

where L is the effective length = 250 mm

$$\delta_b = \frac{30406.25 \times (250)^2}{3 \times 200 \times 103 \times 19174.76} = \frac{1900390625}{1.1504856 \times 10^{10}} = 0.165\,\text{mm}$$

The corresponding angular displacement is given by

$$0.165/250 = 6.607 \times 10^{-4}\text{rads.}$$

3.11 Maximum Torque Transmitting Capacity

$$= 1.0 \times 293.44 = 2934.4\,\text{N/mm}$$

This value is far more than the machine shall ever encounter in service.

3.12 Screws

Four 5 mm diameter screw in 48.26 mm PCD.

Shear load on each screw $= \frac{1\times 30406.25}{4\times 48.26=157.51\,\text{N}}$

Average shear stress $= \frac{157.51}{\frac{\pi}{4}\times 5^2} = 8.02\,\text{N/mm}^2$

3.13 Estimation of the Angular Twist of the Entire Torque-Transmitting Shaft Under Full Capacity Load

This exercise is necessary in order to be certain that the maximum angular displacement obtainable from the double eccentric shafts is adequate for providing the twist corresponding to maximum load. The analysis is made assuming. The use of the use of the largest shaft diameter of (25 mmϕ) and that the, specimen steel remains elastic unto the full load, The general idea is to ensure that adequate allowance is made for the case of plastic deformation where strains are larger.

The angular twist on a length L of a shaft of diameter d subjected to a torque T is given by the expression:

$$\theta = \frac{T}{G} \times L \times \frac{32}{\pi d^4} \theta$$

Where G is the modulus of rigidity = 80 × 10^3 N/mm
Substituting for T and G, and the expression reduces to

$$\theta = \frac{30406.25}{80 \times 10^3} \times L \times \frac{32}{\pi d^4} = 3.87 \frac{L}{d^4} \text{ rads.}$$

3.14 Bearing

A total of three bearing are incorporated into the design. They are of the light series. They are chosen for rigidity and to reduce deflection. The details are given below:

3 bearings, 25 mm bore, single row, deep groove ball bearing, self-aligning for the bearing housing for the three bearings.

The bearing in housing No 1 sustains a maximum load of

$$\frac{M}{L} = \frac{30406.25}{250} = 121.625\,\text{N}.$$

Hence average shear stress $= \frac{\frac{1}{2}(121.625)\times 4}{\pi \times (25)^2} = 0.12\,\text{N/mm}^2$

The bearing in housing No. 2 and 3 sustains a load of

$$\frac{M}{L} = \frac{30406.25}{150} = 202.71\,\text{N}$$

Also average shear stress $= \frac{\frac{1}{2}(202.75)\times 4}{\pi \times (25)^2} = 0.21\,\text{N/mm}^2$.

3.15 Bearing Housing

The three main bearing housing Nos. 1–3 are robust construction and are perfectly made of cast steel together with the weds as flanges. The casting are then rebored and forced rigidly onto the horizontal metal plate after which the final counter boring and reaming operators are performed with a high degree of accuracy. This ensures the maintenance of the axial symmetry of the bearing that was mounted in them.

Retaining rings are chosen according to availability and size, no emphasis is being placed on axial load capacity since in the design, and there are no axial loads.

3.16 Analysis of Dynamic Forces

Dynamic force of vary large magnitude are called into play by the rotation and acceleration of the unbalanced masses of the shafts. The effect here is to develop approximate analytical solution from which maximum dynamic forces are then computed.

3.17 Design of Weight Hanger

The weight hanger is designed to carry dead weight that will induce stress on the specimen. The weight hanger is designed to gain support from the specimen. This implies that the weight hanger is on the specimen. The weight hanger has these components:

Brass buchion: That has a direct contact with the specimen. It is simply on it that the hanger is simply supported from. It is two in number and separated apart. It is of brass buchion to create room for rotation without loosing its parts.

Arms: The hanger has arms that are 200 mm long. This arm is extended to a flat pan, which is the base for the weight hanger assembly. The arm is bolted firmly to the flat pan.

The spring and leg of hanger: This is the load carried. It has a spring to absorb shocks or impact force from getting to the specimen. Even when the machine is switched on there are some vibrations that are generated due to critical speed of these elements do not have an effect on the weight applied.

3.18 Analysis of the Design Calculations of the Weight Hanger

The weight hanger is designed under some assumed specifications. The analysis begins from determining the critical speed.

3.19 Critical Speed of the Machine

Critical speed is the speed the shaft attains and becomes unstable with deflection increasing without upper bonds. When as shaft is turning eccentricity causes a centrifugal force deflection which can be resisted to some extent by the shaft flexural rigidity EL As long as the deflection are small, no harm is done in the cause of the design, we are determining the critical speed based on the shaft and ensembled attachments mass. The critical speed of the fatigue testing machine design is given by:

$$\omega_c = (\pi/L)^2\sqrt{\frac{EI \times g}{Ay}}$$

$$\mathrm{L} = 110\,\mathrm{mm} = 0.11\mathrm{m}$$

$$\mathrm{E} = 200\,\mathrm{K\,N/mm^2} = 200 \times 10^3\,\mathrm{N/mm^2}$$

$$I = \frac{\pi}{64} \times d^4 = \frac{\pi}{64} \times 25^4 = 19174.76\,\mathrm{mm}^4$$

$$\mathrm{g} = 9.8\,\mathrm{m/s^2}$$

$$\mathrm{a} = \text{area of shaft} = \frac{\pi}{4} \times \delta^2 = \frac{\pi}{4} \times 25^2 = 490.9\,\mathrm{mm}^2$$

$$\gamma = \mathrm{p} \times \mathrm{L} = 7850 \times 0.11 = 863.5\,\mathrm{kg/m^2}$$

If $\Phi = \mathrm{A}\gamma \quad \gamma = 863.5\,\mathrm{kg\,m^3}$and $\mathrm{A} = 0.491\,\mathrm{m}^2$
Then $\phi = 423.87$ kg

$$\omega_c = (\pi/L)^2\sqrt{\frac{EI \times g}{\phi}}$$

$$\omega_c = (\pi/110)^2\sqrt{\frac{200 \times 10^3 \times 9.8 \times 19174.76}{423.87}}$$

$$= 8.1567 \times 297767.067 = 242.88\ \mathrm{rad/s} = 14572.8\,\mathrm{rad/mm}$$

This is greater than the natural speed of speed of the motor. This now approves the critical speed to be acceptable, to reduce its effect on the machine, the deflection is calculated first before continuing the vibration.

With L = 110 m, F = 293.44 N, I = 19174.76 mm^4 and E = 200 × 10^3
Static deflection of the shafting is gotten by:

$$\delta = \frac{FL^2}{3EI} = \frac{293.44 \times (110)^2}{3 \times 200 \times 10^3 \times 19174.76} = 0.034\,\mathrm{mm}$$

Frequency of the transverse vibration

$$F_n = \frac{0.4985}{\sqrt{\delta}} = \frac{0.4985}{\sqrt{0.034}} = 2.71\,\mathrm{Hz}$$

This is less than the frequency of the rotating speed of the motor.
Therefore, the machine is free from much vibration.

3.20 Determination of Spring for Damping Vibration

Taking Minimum weight to be carried to 100 N

Maximum weight to be carried to 500 N

Spring index C = 6, Factor of safety, Fs = 1.25,

Yield strength $\tau_y = 700$ mPa, Endurance limit, $\tau_e = 350$ mpa, Modulus of rigidity G = 80 KN /mm^2, Maximum deflection $\delta_f = 30$ mm.

3.21 Determination of Size of Spring

Let d = diameter of wire

D = mean diameter of spring D = C · d

We have that

Mean load $W_m = W_{max} + W_{min}/2 = 500 + 100/2 = 300$ N

Variable load W_r = max−min/2 = 500−100/2 = 200 N

Shear stress factor k_s = I + ½ c = I + ½ × 6 = 1.083

Wahls stress factor,

$$\mathrm{k} = \frac{4c-1}{4c-4} + \frac{0.615}{6} = \frac{4\times 6-1}{4\times 6-4} + \frac{0.615}{6}$$
$$= \left(24 - {}^{1}\!/_{24} - 4\right) + \left({}^{0.615}\!/_{5}\right) = 1.252\,\mathrm{s}$$

We know that menu shear stress τ_m is given by the equation

$$\tau_m = \mathrm{k} \times 8\mathrm{W_m} \times \mathrm{D}/\pi \mathrm{d}^3 = 1.083 \times 8 \times 300 \times 6\mathrm{d}/\pi\mathrm{d}^3 = \left(4964.1/\mathrm{d}^2\right)\mathrm{N/mm}^2$$

Variable shear stress is gotten from the equation

$$\tau_v = \mathrm{k}\ \times 8\mathrm{W_V} \times \mathrm{D}/\pi\mathrm{d}^3 = 1.2525 \times 8 \times 200 \times 6\mathrm{d}/\pi\mathrm{d}^3 = \left(3827.36/\mathrm{d}^2\right)\mathrm{N/mm}^2$$

Also we know that factor of safety (F.s) is equal to:

$$1/\mathrm{F.s}(\tau_m - \tau_v/\tau_y)/(\tau_y 2\tau_v/\tau_c)$$

$$\frac{1}{1.25} = \frac{\frac{4964.1}{\mathrm{d}^2} - \frac{3827.36}{\mathrm{d}^2}}{700} + 2 \times \frac{\frac{3827.36}{\mathrm{d}^2}}{350}$$
$$\frac{1}{1.25} = \frac{\frac{1136.74}{\mathrm{d}^2}}{700} + \frac{\frac{7654.72}{\mathrm{d}^2}}{350}$$
$$\frac{1}{1.25} = \frac{1.6239}{\mathrm{d}^2} + \frac{21.8706}{\mathrm{d}^2}$$
$$\frac{1}{1.25} = \frac{23.4945}{\mathrm{d}^2} \quad \mathrm{d}^2 = 23.4945 \times 1.25 \text{ and } \mathrm{d} = 5.42\,\mathrm{mm}$$

3.22 Diameter of Spring

(a) Mean diameter D = c · d = 6.d = 6 × 5.42 = 32.52 mm
(b) Outer diameter of the spring D_o = D + d = 32.52 + 5.42 = 37.94 mm
(c) Inner diameter of the spring D_i = D + d = 32.52–5.42 = 27.1 mm

3.23 Number of Turns of Spring

Let n = number of turns of spring and deflation of spring

$$\delta = \frac{8 \cdot W \cdot D^3 \cdot n}{G \cdot d^4}$$

$$30 = \frac{8 \times 500 \times (32.52)^3 .n}{8 \times 10^3 \times (5.42)^4}$$

$$n = \frac{30 \times 80 \times 103 \times (5.42)^4}{8 \times 500 \times (32.52)^3}$$

$$= 15.06 = 15 \text{ turns.}$$

Using squared and grounded ends of spring, the total number of turns of spring for the weight hanger will be: n1 = n + 2 = 15 + 2 = 17 turns.

3.24 Free Length of the Spring

The free length (4) of a spring is the length of the spring in the free or unloaded condition. It is given by:

L_f = solid length + maximum compression + clearance between adjacent coils.

$$n^1 d + \delta_{max} + 0.15 d_{max}$$

$$17 \times 5.42 + 30 + 0.15(3D) = 92.14 + 30 + 4.5 = 126.64\,\text{mm.}$$

3.25 Stiffness of Spring

This is the load required per unit deflection of the spring.

$$K = \frac{W_{max}}{\delta} = \frac{500}{30} = 16.67\,\text{N/mm.}$$

3.26 Pitch of the Coil Spring P

The pitch of a spring coil is defined as the axial distance between adjacent coils in uncompressed state.

$$P = \frac{Free\ length}{n^1 - 1} = \frac{126.64}{17 - 1} = 7.915\,\text{mm}.$$

3.27 Hanger Rod

Diameter of the hanger rod will be equal to the internal diameter of the spring minus clearance or allowances.

$$\text{Di} = 27.1\,\text{mm}$$
$$\text{Rod diameter} = \text{Di} - 0.1\ \text{Di} = 27.1 - 0.1(27.1) = 27.1 - 2.71 = 24.39\,\text{mm}.$$

3.28 Hanger Rod Head (Top)

This is a circular shape at the top of the rod to mesh on the top of spring to room for compression of the spring. Diameter of the hanger rod top is equal to outer diameter plus allowance.

$$\text{D}_\text{o} = 37.94\,\text{mm}$$
$$\text{Rod head diameter } = \text{D}_\text{o} + 0.1\ \text{D}_\text{o} = 37.94 + 3.794 = 41.734\ \text{mm}.$$

3.29 Test for Brittle Aluminium

For a test conducted on the machine using brittle aluminium specimen, the following result on Table 2 was obtained and when plotted on S–N graph, the graph (Fig. 2) was obtained.

Table 2 Test results on brittle aluminium specimen

Stress (MPa)	320	300	260	200	150
Life (cycle)	10	10^2	10^3	10^4	10^5

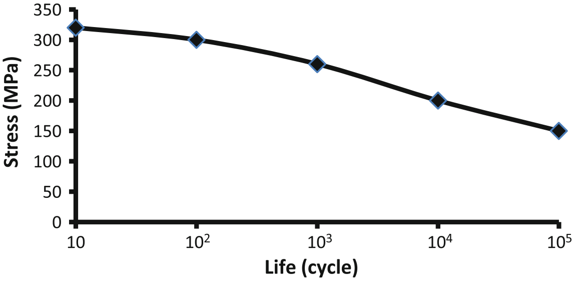


Fig. 2 An S–N graph of a test conducted on the machine using brittle aluminium

4 Conclusion

The study and test conducted so far shows that fatigue failure cannot be predicted accurately since material failure under fatigue are affected not by just reversal loading alone but also the number of revolution (cycle per minute) and fluctuating stress and other factors such as temperature, atmospheric condition, both internal and external defect on material subjected under fatigue stress. Such defect includes notch, inclusion, stress concentration and non-homogeneity.

At such, fatigue failure is sudden and total, hence dangerous and leads to major accident characterized by loss of lives, valuable goods and devices. Thus all precautions and measures should be taken to checkmate this failure since it cannot be curbed entirely or predicted in-to-to.

Chapter 40
The 2007–2009 Financial Crisis and Credit Derivatives

Mmboniseni Mulaudzi, Mark Petersen and Janine Mukuddem-Petersen

Abstract We discuss the relationship between investor payoffs and credit derivatives such as credit default swaps (CDSs) and mortgage-related collateralized debt obligations (CDOs). In this regard, we investigate the role that the interplay between these components played in the global financial crisis (GFC). More specifically, we develop a stochastic model for investor payoffs from investment in CDO tranches that are protected by CDSs. In a continuous-time framework, this model enables us to solve a stochastic optimal credit default insurance problem that has investor consumption and investment in structured mortgage products as controls. Finally, we provide numerical results involving mezzanine CDO tranches being hedged by CDSs and explain their link with the GFC.

Keywords Collateralized debt obligation · Credit default swap · Credit derivatives · Credit risk · Global financial crisis · Systemic risk

1 Introduction

The period prior to the 2007–2009 GFC was characterized by financial product development intended to achieve objectives such as offsetting a particular risk exposure (such as mortgage default) or obtain financing. Examples pertinent to this

M. Mulaudzi (✉)
Department of Decision Sciences, University of South Africa,
P. O. Box 392, Pretoria 0003, South Africa
e-mail: mulaump@unisa.ac.za

M. Petersen · J. Mukuddem-Petersen
Faculty of Commerce and Administration, North West University,
Private Bag X2046, Mmabatho 2735, South Africa
e-mail: mark.petersen@nwu.ac.za

J. Mukuddem-Petersen
e-mail: janine.mukuddempetersen@nwu.ac.za

H. K. Kim et al. (eds.), *Transactions on Engineering Technologies*,
DOI: 10.1007/978-94-017-9115-1_40,

crisis include the pooling of subprime mortgages into mortgage-backed securities or collateralized debt obligations (CDOs) for investment via securitization and a form of credit default insurance known as credit default swaps (CDSs). In particular, CDO issuance grew from an estimated $20 billion in Q104 to its peak of over $180 billion by Q107, then decreased to under $20 billion by Q108. Further, the credit quality of CDOs declined from 2000–2007, as the level of subprime and other non-prime mortgage debt increased from 5 to 36 % of CDO assets. In addition, CDOs and portfolios of CDSs called synthetic CDOs enabled a theoretically infinite amount to be wagered on the finite value of mortgages. In this regard, buying a CDS to insure a CDO ended up giving the seller the same risk as if they owned the CDO when the CDO market imploded. This boom in credit derivatives was accompanied by more complexity (compare with the IDIOM hypothesis postulated in [13]). This process increased the number of agents—such as mortgage brokers, specialized originators, special purpose vehicles and their due diligence firms, managing agents and trading desks as well as investors, insurances and repo funding providers—related to mortgage originations. The disconnect from the underlying mortgages resulted in these agents relying on indirect information that included FICO scores on creditworthiness, appraisals, organizational due diligence checks as well as computer models of rating agencies and risk management desks. Instead of spreading risk this provided the ground for fraudulent acts, misjudgments and finally market collapse.

Some relevant literature about the GFC and its relationship with credit derivatives is given below. Subprime mortgage-related problems were exacerbated by CDO distribution methods, off-balance sheet vehicles, derivatives that resulted in negative basis trades moving CDO risk as well as derivatives that created additional long exposure to subprime mortgages (see, for instance, [3] and [7]). Determining the extent of this risk is also difficult because the effects on expected mortgage losses depend on house prices as the first order risk factor. Simulating the effects of this through the chain of interacting securities is very difficult (see, for instance, [2]). On the other hand, [12] (see, also, [2] and [5]) shows that credit risk transfer through the derivatives market resulted in the origination of inferior quality mortgages by originators. We believe that mortgage standards became slack because securitization gave rise to moral hazard, since each link in the mortgage chain made a profit while transferring associated credit risk to the next link (see, for instance, [8] and [12]). The increased distance between originators and the ultimate bearers of risk potentially reduced originators' incentives to screen and monitor mortgagors (see [2]). The increased complexity of residential mortgage-backed securities (RMBSs) and markets also reduces the investor's ability to value them correctly (see, for instance, [8]). To our knowledge there is no studies about investor payoffs and subprime credit derivatives other than our conference paper (see, for instance, [10]).

1.1 Preliminaries About Credit Default Insurance

Credit default insurance provided by monolines is subject to capital requirements designed to ensure that there are sufficient funds to cover obligations. CDSs not sold by monolines are largely unregulated derivatives and as such do not have to meet the same collateral requirements. A market for CDSs exists, where multiple CDSs can be sold on a given underlying asset. Here, prices reflect the risk associated with the underlying asset. Since they replace the underlying assets, CDSs can be used to create *synthetic* CDOs, which serve in place of actual RMBS tranches. The volume of subprime mortgage exposures in CDOs can thus exceed (and did come to exceed) the amount of subprime structured mortgage products (SMPs).

The credit default insurance that we consider are related to CDSs. These swaps are financial instruments used as a hedge and protection for SMP holders, in particular CDO tranche investors, from the risk of default. As the nett worth of banks and other financial institutions deteriorated because of losses related to subprime mortgages, the likelihood increased that those providing the protection would have to pay their counterparties. This created uncertainty across the system, as investors wondered which companies would be required to pay to cover mortgage defaults. CDSs may seem to be the final bearer of credit risk resulting from housing-price declines. An overview is provided in Fig. 1.

In Fig. 1, an agreement is made between a protection buyer and seller to exchange credit risk of a CDO in the form of a CDS. The protection buyer makes periodic CDS premium payments in **1A** for a specific period of time, as protection against losses from the CDO. These losses are a consequence of bankruptcy, defaults, asset restructuring etc. The CDS premium payment is equal to the market price of the CDS which is a measure of the value of the credit risk that is being exchanged. The protection seller receives these periodic payments and profits from these while the CDO's price is stable. In the case of a credit event, a settlement between the buyer and seller is made. In this regard, the seller pays out the value of the CDS to the buyer as in **1B**. The risk-taking of the protection buyer is the same as selling a security short. This is classified as "short risk". On the other hand, the risk that the protection seller takes, is the same as owning a loan or security and is referred to as "long risk" (see, for instance [4, 6 and 11]).

1.2 Main Questions and Outline of the Paper

In this section, we state the main problems and provide an outline of the paper.

1.2.1 Main Questions

The main problems to emerge from the above discussion may be formulated as follows.

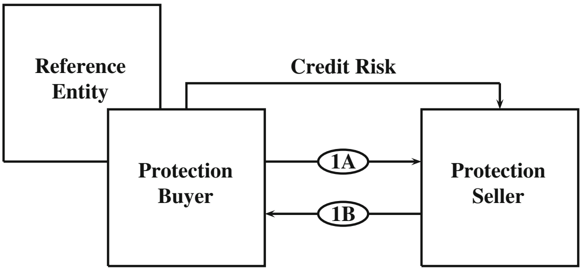


Fig. 1 Diagrammatic overview of credit default swaps—investor hedging against credit risk

Question 1. (*Investor Payoff Model with Hedgeable and Non-hedgeable Credit Risk*): How can we construct a stochastic dynamic model for investor payoff that incorporates credit derivatives? (see Sect. 2.3).

Question 2. (*Downside of Credit Derivatives*): How does our model in Question 1 relate to problems experienced with credit derivatives in the GFC? (see Sect. 2).

Question 3. (*Stochastic Optimal Credit Default Insurance Problem*): Which decisions about the rate of consumption by investors and value of investment must be made in order to attain an optimal investor payoff in the presence of SMPs and credit default insurance? (see Theorem 1 in Sect. 2).

Question 4. (*Quantitative Illustration of the GFC*): How do the numerical results on credit derivatives obtained in our paper explain the GFC in a quantitative way? (see Sect. 2).

1.2.2 Outline of the Paper

In short, we solve an optimal control problem that depends on a stochastic dynamic model for investor payoff that incorporates credit default insurance and SMP dynamics (see Sect. 2.3 of Sect. 2). In particular, we are able to set-up an optimal credit default insurance problem that seeks to establish the optimal rate of consumption, k_t^*, over a random term, $[t, \tau]$, and terminal payoff for SMP investors at τ (see Theorem 1 in Sect. 2). In terms of the GFC, we consider problematic issues such as the reduction of incentives for monitoring SPEs, incentives to destroy value, credit derivative market opacity, industry self-protection and systemic risk as well as the mispricing of credit (see Theorem 1 in Sect. 2).

2 CDSs Hedging Credit Risk from CDO Tranches

In this section, we construct a stochastic model for investor payoff that incorporates CDSs. The economic agents involved are investor banks that purchase subprime CDOs and protection from monoline insurers.

2.1 Structured Mortgage Products and Their Losses

In this section, we consider subprime CDOs and their losses. We suppose that a investor bank can invest in riskless Treasuries whose price at time t given by T follows the process

$$d\mathrm{T}_t = r^{\mathrm{T}}\mathrm{T}_t dt, \quad \text{for some } r^{\mathrm{T}} \geq 0. \tag{1}$$

The investor can also invest in a risky CDO whose price at t is given by P and follows the process

$$d\mathrm{P}_u = \mathrm{P}_u[r^P du + \sigma dZ_u], \tag{2}$$

where the CDO rate $r^P > r^{\mathrm{T}}$ and σ are constants. In addition, Z_t is a standard Brownian motion with respect to a filtration, $(\mathcal{G}_t)_{t \geq 0}$, of the probability space $(\Omega, \mathbb{G}, (\mathcal{G}_t)_{0 \leq t \leq \tau}, \mathbb{P})$.

Also, the investor is subject to an insurable credit risk modeled as a compound Poisson process, in which N is a Poisson process with deterministic parameter $\phi(t)$ and the CDO loss process, S. Assume that N is independent of Z, which is the Brownian motion of the CDO process. Also, the random CDO loss amount S is independent of N.

2.2 Investor Payoff Under Credit Default Insurance

In this section, we present a stochastic differential equation describing the dynamics of investor payoff under mortgage securitization as well as the risks that can be associated with components of this equation.

2.2.1 Model for Investor Payoff Under Credit Default Insurance

At time t, let Π_t and ψ_t be the investor payoff and the amount that the investor invests in CDOs, respectively. The investor earns an exogeneous income rate of $\imath(t)$ and consumes at a rate of k_t at t. For a CDO portfolio with unhedged credit risk we have that

$$\begin{aligned} d\Pi_u^u &= [r^{\mathrm{T}}\Pi_u + (\mu - r^{\mathrm{T}})\psi_u + \imath(u) - k_u]du + \sigma\psi_u dZ_u - S(\Pi_u, u)dN_u, \\ u &\geq t, \Pi_t = \pi, \end{aligned} \tag{3}$$

where $\mu = r^P - r^c$ for the transaction costs rate, r^c.

Next, we consider a CDO portfolio in which credit risk is hedged via CDSs. In this case, if the investor suffers a loss, S, from CDO default, then it is paid C at time t. The result is that

$$\begin{aligned} d\Pi_u = {} & [r^{\mathrm{T}}\Pi_u + (\mu - r^{\mathrm{T}})\psi_u + \imath(u) - k_u]du + p(u)du + \sigma\psi_u dZ_u \\ & - [S(\Pi_u, u) - C_u(S(\pi_u, u))]dN_u, u \geq t, \Pi_t = \pi, \end{aligned} \tag{4}$$

where the CDS premium payment leg rate and default payment leg are given by

$$p(u) = -(1 + \lambda(u))\phi(u)\mathbb{E}^{\mathbb{P}}[\mathrm{C}_u(S)] \quad \text{and} \quad \mathrm{C}_u(S(\Pi_u, u)), \tag{5}$$

respectively.

2.2.2 Numerical Results for Investor Payoff Under Credit Default Insurance

In this section, a motivating example is provided by the relationship between the insurer AIG (CDS protection seller) and Merrill Lynch (CDO tranches buyer). The latter's major losses in 2008 were attributed in part to the drop in value of its unhedged CDO portfolio after AIG ceased offering CDSs on CDOs. The loss of confidence of trading partners in Merrill Lynch's solvency and its ability to refinance its short-term debt eventually led to its acquisition by the Bank of America. Subsequently, values for investor payoffs are considered for 2000 to 2008. The simulation is obtained via the Euler-Maruyama numerical method. First, we consider the dynamics of investor payoff where credit risk from CDO tranches is unhedged.

Next, we present investor payoff dynamics where credit risk from CDO tranches is hedged by CDSs.

Figures 2 and 3 reflect investor payoff dynamics when credit risk from CDO tranches was unhedged and hedged by CDSs. If we consider the components of (3) and (4), low interest rates prevailed in 2001–2004 preceded by an increase in interest rates by the Federal Reserve Bank that scuppered the ability of mortgagors to refinance. Defaults increased dramatically in the U.S. in late 2006 and triggered a global financial crisis from 2007 onwards. The downturn of the housing market caused mortgage losses to increase significantly. Notice that losses were significantly less when CDO tranches were hedged than when they were not (compare Figs. 2, 3). The region inside the red circle in Fig. 3 bears testimony to an increase in protection seller payments to compensate for counterparty defaults. Simultaneously, the investor's payoff decreased significantly due to subprime bond losses as well as mortgage defaults, foreclosures, etc.

Figures 2 and 3 also support the hypothesis of [3] where it was observed that new issuance of CDOs came to an abrupt halt in early 2007. This took place subsequent to the implosion and re-pricing of credit risk in the capital markets.

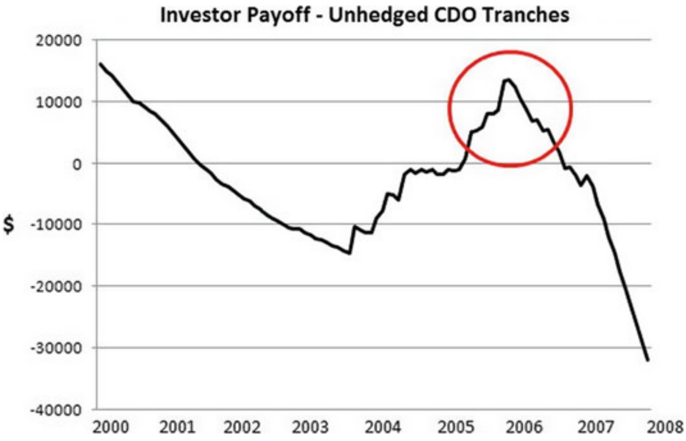


Fig. 2 Investor payoff–unhedged credit risk from CDO tranches

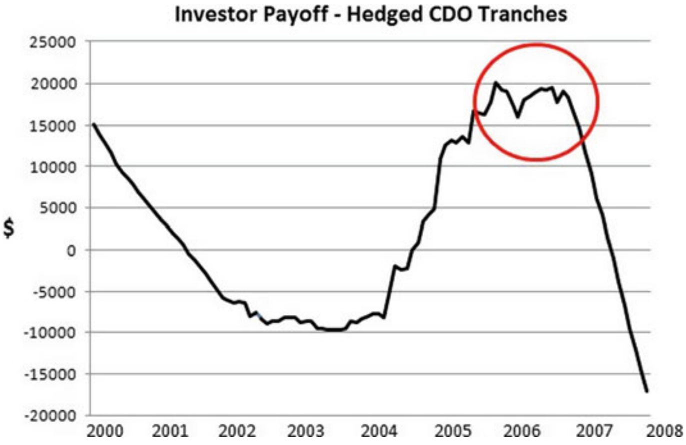


Fig. 3 Investor payoff–hedged credit risk from CDO tranches

Here it was found that the market inefficiencies were substantial, given the size of the CDO market and the magnitude of CDO fees. During the GFC, CDOs were used to arbitrage a substantial price discrepancy in the mortgage markets and to convert existing mortgages that are priced accurately into new fixed income mortgage-related instruments that are overvalued. Also, the aforementioned figures are related to [7] that uses data on privately-securited subprime mortgages to examine study the increase in defaults after 2007.

2.3 Optimal Credit Default Insurance

In this section, we solve an optimal credit default insurance problem related to the stochastic model of investor payoff with hedged credit risk given by (4). Because of the unpredictable shutdown of CDO markets, the solution to the insurance problem is determined for a random term $[t, \tau]$.

2.3.1 Statement of the Optimal Credit Default Insurance Problem

We let a set of control processes (laws), $\mathscr{A}$ which is adapted to investor payoff, Π, have the form

$$\mathscr{A} = \{(k_t, \psi_t, C_t) : \text{ measurable w.r.t.} \mathscr{G}_t;\ (4) \text{ has unique solution}\}. \tag{6}$$

The objective function of the stochastic optimal credit default insurance problem is given by

$$J(\pi, t) = \sup_{\mathscr{A}} \mathbb{E}^{\mathbb{P}} \left[\int_t^\tau \exp\{-\delta^r(u-t)\} U^{(1)}(k_u) du + \exp\{-\delta^r(\tau - t)\} U^{(2)}(\Pi_\tau) \right], \tag{7}$$

where, for the first- and second-order differential operators, D and D^2, respectively, we have

$$DU^{(1)}(.) > 0, D^2U^{(1)}(.) < 0, DU^{(2)}(.) > 0 \text{ and } D^2U^{(2)}(.) < 0.$$

Here, $U^{(1)}$ and $U^{(2)}$ are increasing, concave utility functions and $\delta^r > 0$ is the rate at which the utility functions for consumption, k, and terminal payoff, Π_τ, are discounted. Of course, in principle, one can formulate any utility function. The question then is whether the resulting Hamilton-Jacobi-Bellman equation (HJBE) can be solved (smoothly) analytically? In the sequel, we obtain an analytic solution for the choice of power utility functions.

We are now in a position to state the stochastic optimal credit default insurance problem for the investor's consumption rate, k, and terminal payoff, Π_τ, for an adjustment term, $[t, \tau]$.

Problem 1 (*Optimal Credit Default Insurance*): Suppose that the admissible class of control laws, $\mathscr{A} \neq \emptyset$, is given by (6). Moreover, let the controlled stochastic differential equation for the Π-dynamics be given by (4) and the objective function, $J : \mathscr{A} \to \Re^+$, by (7). In this case, we solve

$$\sup_{\mathscr{A}} J(k_t, \psi_t, C_t),$$

and the optimal control law $(k_t^*, \psi_t^*, \mathrm{C}_t^*)$, if it exists,

$$(k_t^*, \psi_t^*, C_t^*) = \arg\sup_{\mathscr{A}} J(k_t, \psi_t, C_t) \in \mathscr{A}.$$

The optimal credit default insurance problem determines the optimal consumption rate, k^*, and investor's optimal investment in mortgages, ψ^*, over a random interval. In this regard, Theorem 1 provides the general solution to this problem (see Problem 1). We note that the objective function in (7) is additively separable in $U^{(1)}$ and $U^{(2)}$ which is not necessarily true for all investors. In our problem, we have a discount rate, δ^r, which is used to discount these utility functions. This discount rate is chosen by the investor and it is not the market discount rate. In the sequel, connections between specific solutions of the optimal credit default insurance problem and the GFC are forged.

2.3.2 Solutions to the Optimal Credit Default Insurance Problem

In this section, we determine a solution to Problem 1 in the case where the term $[t, \tau]$ is random. In order to find the optimal control processes, we use the dynamic programming method where we consider an appropriate HJBE. In the sequel, we assume that the optimal control laws exist, with the objective function, J, given by (7) being continuous twice-differentiable. Then a combination of integral calculus and Itô's formula (see, for instance, [15]) shows that J satisfies the HJBE

$$\begin{cases} \delta^r J = J_t + \max\limits_{k}[U^{(1)}(k) - kJ_\pi] + (r^T\pi + \imath(t))J_\pi \\ \qquad + \max\limits_{\psi}\left[(\mu - r^T)\psi J_\pi + \frac{1}{2}\sigma^2\psi^2 J_{\pi\pi}\right] \\ \qquad + \max\limits_{C}\left[\phi(t)\left\{\mathbb{E}^{\mathbb{P}} J(\pi - (S - C(S)), t) - J(\pi, t)\right\}\right. \\ \qquad \left. - (1 + \lambda(t))\phi(t)\mathbb{E}^{\mathbb{P}}[S(S)]J_\pi\right] + \omega_b(t)\left[U^{(2)}(\pi) - J(\pi, t)\right] \\ \qquad \lim\limits_{s\to\infty} \mathbb{E}^{\mathbb{P}}\left[\exp\left\{-\int\limits_t^s (\rho + \omega_b(u))du\right\} J(\Pi_s^*, s) | \Pi_t^* = \pi\right] = 0, \end{cases} \tag{8}$$

In the sequel, J_t, J_π and $J_{\pi\pi}$ denote first and second order partial derivatives of J with respect to the variables t and π. The objective function, J, is increasing and concave with respect to payoff, π, because the utility functions $U^{(1)}$ and $U^{(2)}$ are increasing and concave. In this case, $\omega_b(t)$ is the hazard rate for investor at time t (compare with the hazard rate analysis in [7]). During the GFC, the hazard rate was very high due to dysfunction in the CDO market. It is important to note that the HJBE (8) can be deduced by using the methods contained in [15]. As a consequence, the integrability and regularity conditions that arise in our paper are covered by these contributions. For instance, in our case, we can use the verification theorems in [15] to show that if our objective function, J, has a smooth solution as well as the related HJBE, $\widehat{J}$, then under our regularity conditions, $J = \widehat{J}$.

Theorem 1 (Optimal Credit Default Insurance): *Suppose that the objective function, $J(\pi, t)$, solves the HJBE* (8). *In this case, a solution to the stochastic optimal credit default insurance problem is*

$$\psi_t^* = -\frac{\mu - r^{\mathrm{T}}}{\sigma^2}\frac{J_\pi(\Pi_t^*, t)}{J_{\pi\pi}(\Pi_t^*, t)}, \tag{9}$$

in which Π_t^* is the optimally controlled payoff under credit default insurance. Also, the optimal consumption rate, $\{k_t^*\}_{t \geq 0}$, solves the equation

$$D_k U^{(1)}(k_t^*) = J_\pi(\Pi^*, t), \tag{10}$$

where D_k represents the ordinary derivative with respect to k.

Proof The proof is completed via standard arguments about static optimization (see, for instance, [15]). □

2.3.3 Optimal Accrued Premiums

We recall that the CDS accrued premium is the amount owing to the protection seller for investor's credit default protection for the period between the last premium payment and default at τ. This premium has a direct influence on optimal CDS represented by C^*. For instance, from insurance theory (see, for instance, [1] and the extension to continuous-time in [9]), we have that the optimal CDS process is related to classical insurance theory. Analogous to [9] where deductibles were discussed, we can show that in the continuous-time setting, optimal CDS is accrued premium CDS. In this regard, we assume that $0 \leq \mathrm{C} \leq S$. Taking our lead from insurance theory and the assumption that $p(u)$ is proportional to the nett CDS premium for a portfolio with mass of type-A CDOs, λ, the optimal CDS contract takes the form

$$C(S) = \begin{cases} 0, & \text{if } S \leq \eta; \\ S - \eta, & \text{if } S > \eta. \end{cases} \tag{11}$$

Some features of the aforementioned CDS contract are as follows. If $S \leq \eta$, then it would be optimal for the investor not to buy CDS protection. If $S > \eta$, then it would be optimal to buy CDS protection. In the sequel, the maximization of the CDS contract purchased by the investor is now reduced to the problem of determining the optimal accrued premium, η.

Proposition 1 (Optimal Accrued Premium): *The optimal CDS contract is either no protection or per-loss accrued premium CDSs, in which the accrued premium, η, varies with time. In particular, at a specified time, the optimal accrued premium, η_t^*, solves*

$$J_\pi(\Pi_t^* - \eta_t, t) = [1 + \lambda(t)]J_\pi(\Pi_t^*, t). \tag{12}$$

No CDSs contract is optimal at time t if and only if

$$J_\pi(\Pi_t^* - ess\, sup\, S(\Pi_t^*, t), t) \leq [1 + \lambda(t)]J_\pi(\Pi_t^*, t). \tag{13}$$

Proof Again the proof is completed via standard arguments about static optimization (see, for instance, [15]). □

In order to determine an exact (closed form) solution for the stochastic optimization problem in Theorem 1, we are required to make a specific choice for the utility functions $U^{(1)}$ and $U^{(2)}$. Essentially these functions can be almost any function involving k and π, respectively. However, in order to obtain smooth analytic solutions to the stochastic optimal credit default insurance problem, in the ensuing discussion, we choose power utility function and analyze the result.

From Proposition 1, we deduce that the optimal CDS contract coincides with the optimal accrued premium, η^*. In this regard, η^* is attained when the marginal cost of decreasing or increasing η is equals to the marginal benefit of the CDS contract. Moreover, if $\lambda = 0$ then the optimal accrued premium should be zero, i.e., $\eta^* = 0$. In this case, if the investor holds no type-A CDOs, λ,—which is indicative of a high PD for reference mortgage portfolios—then it may be optimal for the investor to purchase a CDS contract which protects against all such losses. However, full protection may also introduce high costs in the event that the protection seller fails to honor its obligations. In particular, during the GFC, many investors that purchased CDS contracts promising to cover all losses, regretted making this decision when the protection sellers were unable to make payments after a credit event. Notwithstanding this, certain investors that bought CDS contracts that only pay when the losses exceed a certain level set by the protection seller found protection beneficial. In particular, they did not experience the same volume of losses as those who purchased full protection (see, for instance, [2]).

2.4 Optimal Credit Default Insurance with Power Utility

For a choice of power utility, we have that

$$\overline{U}^{(1)}(k) = \frac{k^\alpha}{\alpha} \quad \text{and} \quad \overline{U}^{(2)}(\pi) = \gamma\frac{\pi^\alpha}{\alpha}, \tag{14}$$

for some $\alpha < 1, \alpha \neq 0$, and $\gamma \geq 0$. The parameter γ represents the weight that the investor gives to terminal payoff versus the consumption rate and can be viewed as a measure of its propensity to retain earnings. This leads to the following result.

Proposition 2 (Optimal Credit Default Insurance with Power Utility)*: Let the power utility functions be given by* (14) *and assume that the investor's CDO losses, S, are proportional to the investor's payoff under mortgage securitization so that*

$$S(\pi, t) = \varphi(t)\pi,$$

for some deterministic S and severity function, $\varphi(t)$, where $0 \leq \varphi(t) \leq 1$. Under power utility, the objective function may be represented by

$$\bar{J}(\pi, t) = \frac{\pi^{\alpha}}{\alpha}\vartheta(t), \vartheta(t) > 0, \tag{15}$$

where $\vartheta(t)$ solves the differential equation

$$\vartheta^{'} + G(t)\vartheta + (1 - \alpha)\vartheta^{\frac{\alpha}{\alpha-1}} = -\gamma\omega_b(t), \tag{16}$$

with $G(t)$ having the form

$$\begin{aligned} G(t) = & -\delta^r + \frac{(r^{\mathrm{T}}\pi + \iota(t))\alpha}{\pi} + \frac{1}{2}\frac{(\mu - r^{\mathrm{T}})^2}{\sigma^2(1-\alpha)}\alpha \\ & + \frac{\phi(t)}{\pi^{\alpha}}(\mathbb{E}^{\mathbb{P}}[(\pi - \eta^*)^{\alpha}] - \pi^{\alpha}) - (1 + \lambda(t))\phi(t)\frac{\alpha}{\pi}\mathbb{E}^{\mathbb{P}}[S - \eta^*] - \omega_b(t) \end{aligned}$$

In this case, the investor's optimal rate of consumption is given by

$$k_t^* = \vartheta^{\frac{1}{\alpha-1}}\pi, \tag{17}$$

and the investor's optimal investment in CDOs is

$$\psi_t^* = \frac{(\mu - r^{\mathrm{T}})}{\sigma^2(1-\alpha)}\pi. \tag{18}$$

Furthermore, under power utility, the optimal accrued premium is given by

$$\eta_t^* = \min\{[1 - (1 + \lambda(t))^{\frac{1}{\alpha-1}}], \varphi(t)\}\pi. \tag{19}$$

Proof The proof follows from Theorem 1 and Proposition 1 as well as (8). Furthermore, a consideration of [14, Chapter 5, Sect. 3] yields a unique solution to (4) under power utility. □

In Proposition 2, the optimal controls in (17), (18) and (19) are expressed as linear functions of the investor's optimal profit under mortgage securitization, Π^*. In this case, we see that the optimal consumption rate, k^*, is independent of the frequency and severity parameters ϕ and φ, of the aggregate CDO losses, S, respectively. These results are true because the power utility function exhibits constant relative risk aversion which means that

$$-\frac{\pi D^2\overline{U}^{(2)}(\pi)}{D\overline{U}^{(2)}(\pi)} = 1 - \alpha.$$

Here, we see that if the relative risk aversion increases, the amount invested in CDOs decreases which may be indicative of the fact that the mass of type-A CDOs, λ, is low at that time. The expression for ϑ in (16) reveals that not only the objective function, $\bar{J}$, is affected by the horizon τ, but also the optimal consumption, k^*. Moreover, the investor's optimal investment in CDOs, ψ^*, is affected by the time horizon τ via the optimal consumption rate, k^*, which impacts on the investor's profit. In addition, the expression for ϑ in (16) shows that k^* depends on the frequency and severity parameters, ϕ and φ, of the CDO losses, S, respectively. Furthermore, the investor's optimal investment, ψ^*, is affected by mortgage losses that indirectly involves k^*. From Proposition 2, it is clear that the amount invested in CDOs, ψ, depends on the profit, Π. Reference mortgage portfolio defaults will cause a decrease in the investor's profit under mortgage securitization, which will later affect the consumption rate, k. In particular, this may cause a liquidity problem in the secondary mortgage market since μ may decrease as a result of this effect on k.

If profits, Π, decrease, it is natural to expect that some investors will fail as in the GFC. For instance, both the failure of Lehman Brothers investment bank and the acquisition in September 2008 of Merrill Lynch and Bear Stearns by Bank of America and JP Morgan, respectively, was preceded by a decrease in profits from securitization. A similar trend was discerned for the U.S. mortgage companies, Fanie Mae and Freddie Mac, who had to be bailed out by the U.S. government at the beginning of September 2008.

3 Conclusion and Future Directions

In this paper, we constructed a stochastic dynamic model for investor payoff that incorporates credit derivatives (see Question 1). This model related to problems experienced with credit derivatives in the GFC such as the reduction of incentives for monitoring SPEs, incentives to destroy value, credit derivative market opacity, industry self-protection and systemic risk as well as the mispricing of credit (see Question 2). In continuous-time, we obtained optimal investor payoff in the presence of SMPs and credit default insurance with consumption, SMP value and credit default insurance as controls (see Question 3). Finally, we were able to explain elements of the GFC in a quantitative way via numerical results involving credit derivatives (see Question 4).

In future, it is important that we increase the sophistication of our model by incorporating interest rate and credit risk more effectively. Also, our model has to accommodate dealing with real financial market interest rates. Structuring the securitization and pricing its outcome or for explaining the economic mechanism behind the recent crisis. In this regard, we have to account for important issues such as moral hazard in expanding mortgage portfolios, incomplete information among market players about their counterparties, myopia in decision making in subprime mortgage market and monetary policy incentives boosting the growth of the subprime market.

References

1. K.J. Arrow, Uncertainty and the welfare economics of medical care. Am. Econ. Rev. **53**, 941–973 (1963)
2. Y. Demyanyk, O. Van Hemert, Understanding the subprime mortgage crisis. Available at SSRN: http://ssrn.com/abstract=1020396. Accessed 19 Aug 2008
3. Y. Deng, S.A. Gabriel, A.B. Sanders, CDO market implosion and the pricing of subprime mortgage-backed securities. J. Hous. Econ. **20**(2), 68–80 (2011)
4. F.J. Fabozzi, X. Cheng, R.-R. Chen, Exploring the components of credit risk in credit default swaps. Finance Res. Lett. **4**, 10–18 (2007)
5. C.H. Fouche, J. Mukuddem-Petersen, M.A. Petersen, M.C. Senosi, Bank valuation and its connections with the subprime mortgage crisis and basel II capital accord. Discrete Dynam. Nat. Soc. 44 (2008). doi:10.1155/2008/740845
6. J. Hull, M. Predescu, A. White, The relationship between credit default swap spreads, bond yields and credit rating announcements. J. Bank. Finance **28**, 2789–2811 (2004)
7. J.B. Kau, D.C. Keenan, C. Lyubimov, V.C. Slawson, Subprime mortgage default. J. Urban Econ. **70**, 75–87 (2011)
8. K. Kendra, Tranche ABX and basis risk in subprime RMBS structured portfolios, Bloomberg, http://www.fitchratings.com/web_content/sectors/subprime/Basis_in_ABX_TABX_Bespoke_SF_CDOs.ppt. Accessed 20 Feb 2007
9. K.S. Moore, V.R. Young, Optimal insurance in a continuous-time model. Insur. Math. Econ **39**, 47–68 (2006)
10. M.P. Mulaudzi, M.A. Petersen, J. Mukuddem-Petersen, Credit derivatives and global financial crisis. *Lecture Notes in Engineering and Computer Sciences: Proceedings of the World Congress on Engineering and Computer Sciences 2013, WCECS 2013*, 23–25 October, 2013. San Frascisco, USA, pp 925–930 (2013)
11. J.P. Morgan, *Credit Derivatives Handbook: Detailing Credit Default Swap Products, Markets and Trading Strategies* (JPMorgan Corporate Quantitative Research, New York, 2006)
12. M.A. Petersen, M.P. Mulaudzi, I.M. Schoeman, J. Mukuddem-Petersen, A note on the subprime mortgage crisis: dynamic modeling of bank leverage profit under loan securitization. Appl. Econ. Lett. **17**(15), 1469–1474 (2009)

13. M.A. Petersen, M.C. Senosi, J. Mukuddem-Petersen, *Subprime Mortgage Models*. (Nova Science Publishers, New York, 2010) (ISBN: 978-1-61761-132-2 (ebook); ISBN: 978-1-61728-694-0 (Hardcover) 2011)
14. P. Protter, *Stochastic Integration and Differential Equations*, 2nd edn. (Springer, Berlin, 2004)
15. H.M. Soner, W.H. Fleming, *Controlled Markov Processes and Viscosity Solutions*, 2nd edn. (Springer, New York, 2008)

Chapter 41
Research Studies on Irreversible Relative Movements (Creeping) in Rolling Bearing Seats Regarding the Influential Parameters and the Remedies

Erhard Leidich and Andreas Maiwald

Abstract The effect of irreversible relative movements, i.e., creeping, in rolling bearing seats is a highly topical issue. Creeping leads to a continuous rotation of the bearing ring relative to the connection geometry (housing or shaft). Creeping may cause wear in the bearing seats, eventually resulting in bearing failure. Therefore, creeping must absolutely be avoided. This study focused on the influential factors that cause the creeping phenomena. In addition, methods to reduce the effects of creeping were defined. Finally, an algorithm for the determination of the critical creeping load was developed. This work enables the user to develop concrete information to optimise a bearing with respect to the bearing seat design and the selection of the roller bearing to prevent creep damage.

Keywords Bearing seat · Corrosion · Creeping · FE simulations · Irreversible relative movements · Roller bearing · Wear

1 Introduction

Roller bearings are used in many important applications. With respect to bearing design, the bearing seats are the primary focus. Issues develop in the bearing seats with increasing performance and higher dynamic stresses, which can lead to irreversible relative movements, i.e., creeping, between the bearing rings and the shaft or the housing. These relative movements are accompanied by the formation of corrosion in the bearing seat, which can lead to fractures of the shaft at the seat

E. Leidich (✉) · A. Maiwald
Department of Engineering Design, Chemnitz University of Technology,
09126 Chemnitz, Germany
e-mail: erhard.leidich@mb.tu-chemnitz.de

A. Maiwald
e-mail: andreas.maiwald@mb.tu-chemnitz.de

H. K. Kim et al. (eds.), *Transactions on Engineering Technologies*,
DOI: 10.1007/978-94-017-9115-1_41,

of the inner ring. Creeping can also result in wear, which can cause shaft displacements with serious consequences, e.g., for the tooth meshing in a gearbox. Various insurance companies describe the creeping of bearing rings (races) as the main damage focus in gearboxes in wind turbines. The follow-up costs for the operator, and ultimately for the manufacturer, are usually considerable. The topic is highly relevant, as there is a requirement for the growth of renewable energy sources, and wind power is an important source of renewable energy [1–4]. In this paper, a complex kinematic 3D finite-element (FE) multi-body simulation of a rolling bearing is presented. Using this simulation method, a detailed analysis of the creeping of the bearing ring (race) is presented for the first time, which introduces new opportunities to research in creeping processes. Furthermore, this method replaces complex and expensive experiments on the original bearing. The focus of these studies is on the outer rings of the radial bearings under a point load. Outer rings are typically provided with a clearance fit and are thus more prone to creeping than are the bearing seats with an interference fit.

2 Creeping

Creeping describes the flexing micro-movements of the bearing ring relative to the connection geometry. In contrast to tangential slip in shaft-hub connections, these movements occur without nominal torsional loading. In Fig. 1, a simplified diagram of the creeping process is shown. As shown in the diagram, even a pure normal load (bearing rests) leads to a wave-like deformation of the bearing ring, which is shown as a plate. The rotation of the bearing creates a combined load of normal and tangential forces on the bearing ring. This combined load generates a slip wave in the bearing seat, which leads to a continuous shift, Δ, between the contacting surfaces.

3 Fe Kinematic Simulation to Investigate Creeping

In the following, the computational implementation of the kinematic simulation is briefly explained. Figure 2 (left) shows the 3D model of the bearing NU 205, which is simulated using approximately 120,000 elements. To simplify the model structure without significant loss of precision, only the outer ring, with the rolling elements, and the housing are modelled. The housing and the outer ring are modelled as independent elastic solids. The rolling elements are rigid bodies. They are coupled by a prismatic joint with a central reference point (Fig. 2, right). Thus, each rolling element has only one degree of freedom in the radial direction. This radial degree of freedom is eliminated by the loads of the rolling elements, which can be calculated according to DIN ISO 281 [5]. By rotation of the reference point, the rolling elements move circumferentially and slide without friction over the

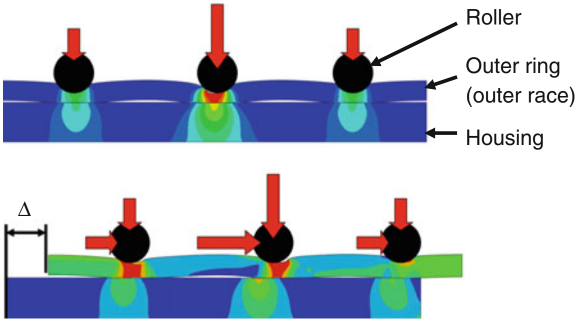


Fig. 1 Simplified representation of creeping movement at the plate. Normal load (*above*) and combined normal and tangential load (*below*)

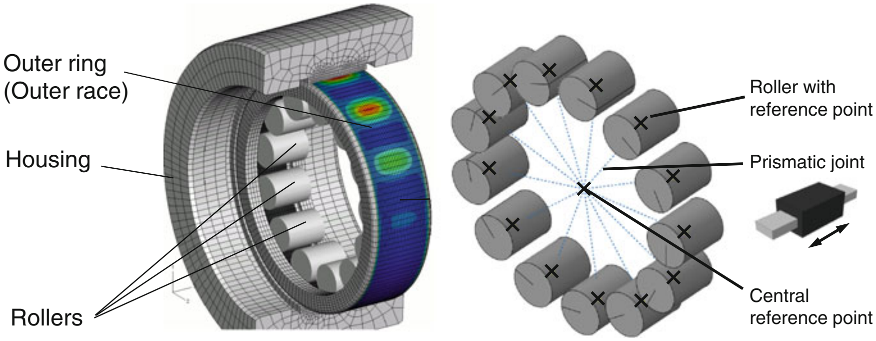


Fig. 2 Model of the simulated bearing NU 205 (*left*) and substitution of the cage (*right*)

contact surface of the outer ring. For all simulations (unless otherwise specified) the coefficient of friction in the bearing seat is set to $\mu = 0.3$. This value corresponds to an oil-lubricated steel-steel contact (100Cr6 E vs. 42CrMo4 +QT) after the "running-in phase", and has been determined experimentally in [3]. A detailed description of the complete simulation methodology can be found in [4]. For the FE analysis, version 6.10-3 of the ABAQUS software was used.

To limit the use of the complex, computationally intensive 3D FE simulations in the bearing design, another calculation model (SimWag, [4, 6]) can be used. The model provides a rough estimate of the bearing load when approaching the onset of creeping and is elucidated in Chap. 8.

To measure the intensity of creeping, the creeping torque, T_C, is evaluated in the simulated bearing ring. The creeping torque describes the torque in the circumferential direction, which is required to prevent the macroscopic relative movement (creeping) between the housing and the bearing ring (Fig. 3). A high creeping torque corresponds to a large creeping intensity of the bearing ring and is therefore classified as negative and converse.

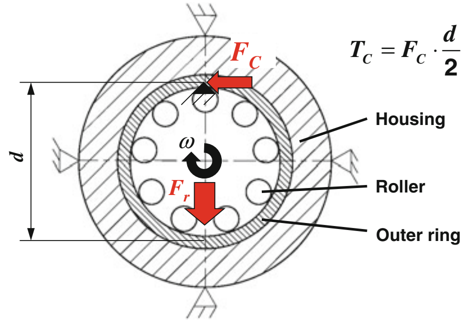


Fig. 3 Model for the determination of the creeping force, F_C, with respect to the creeping torque, T_C

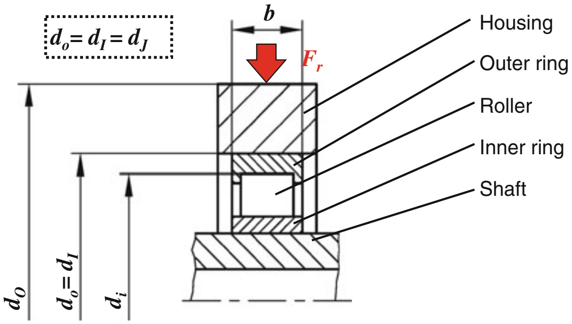


Fig. 4 Geometric parameters and definition of the joint diameter, d_J

Figure 4 shows all of the geometric parameters for the simulation results presented.

The thickness ratio, Q, of the inner and the outer diameter is used to describe the wall thickness of the cylindrical components as follows:

$$Q_b = \frac{d_i}{d_o}, \quad Q_h = \frac{d_I}{d_O} \tag{1}$$

To characterise the interference, or clearance fit, it is normalised by the joint diameter, d_J. By definition, the normalised clearance fit, $\xi*$, has negative values and is expressed as follows:

$$\xi \triangleq \xi^* = \frac{\Delta d}{d_J} \tag{2}$$

The surface area-normalised radial load, p_r, which is analogous to the bearing stress, is determined as the bearing load divided by the projected surface area of the bearing seat, A_{proj}:

$$p_r = \frac{F_r}{A_{proj}}, \quad \text{where } A_{proj} = b \cdot d_o, \tag{3}$$

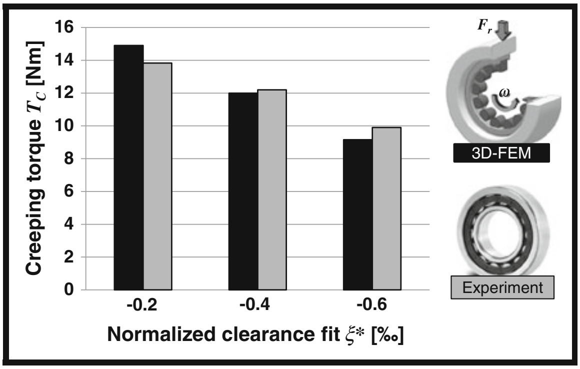


Fig. 5 Comparison of the creeping torque between the results of 3D FE simulations and experiments for the example of the normalised clearance fit, ξ^*, between the housing and the outer ring. (Reference data: Bearing NU 205, normalised radial load $p_r = 18$ MPa, coefficient of friction between the outer ring and housing $\mu = 0.3$, elastic modulus of the housing $E = 210$ GPa)

Using the current FE model, creeping movements are realistically simulated with the aid of a FEM for the first time. The simulation results correlate with the experimental data in [3, 7]. This correlation is demonstrated by the comparison of experimental and simulated results shown in Fig. 5. Thus, the simulation enables the user to perform extensive parameter variations to investigate the effects of creeping, which were not previously possible. Thus, the use of cost-intensive experiments is no longer necessary.

4 Parameter Analysis of the Creeping Behaviour of Roller Bearings

In Fig. 6, the influence of the bearing housing on the creeping behaviour for different values of the modulus of elasticity, E, the thickness ratio with respect to the housing stiffness, Q_h, and the coefficient of friction between the outer ring and the housing, μ, is illustrated.

The results show that with a decreasing modulus of elasticity, E, the creeping torque, T_C, also decreases. This behaviour occurs because a softer housing is better able to absorb the creeping motion of the bearing ring (Fig. 7). Thus, the bearing seat area between two loaded rolling elements (see Fig. 1) is exposed to a local contact pressure increase. Therefore, in this creeping-critical bearing seat area (see Chap. 2), higher local contact shear stresses are transferable.

This leads to a reduction or prevention of the creeping of the bearing ring.

The use of elastic thin-walled housing structures (lower housing wall thickness) only marginally reduces the torque.

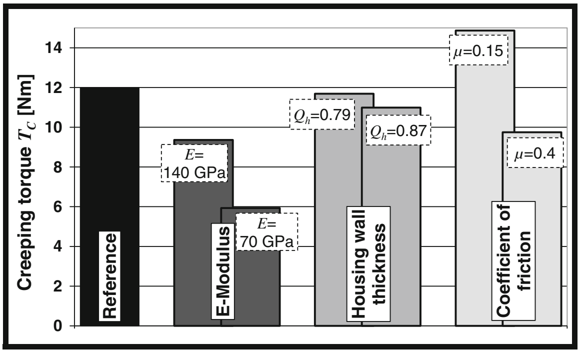
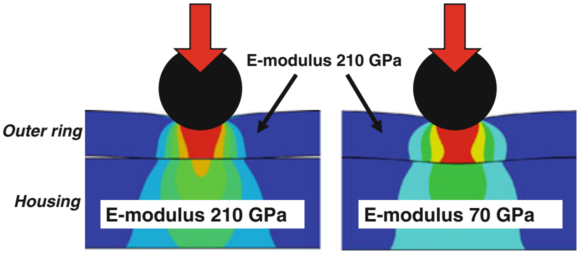


Fig. 6 Change of the creeping torque due to variations in the modulus of elasticity of the housing E, the housing wall thickness Q_h, and the coefficient of friction between the outer ring and the housing μ. (Reference data: Bearing NU 205, normalised radial load $p_r = 18$ MPa, $\mu = 0.3$, normalised clearance fit $\xi^* = -0.4$ ‰, $E = 210$ GPa, $Q_h = 0.69$)

Fig. 7 Qualitative representation of the absorption of the bearing ring deformation under normal load conditions by varying the housing stiffness through the modulus of elasticity E



The influence of the modulus of elasticity, in contrast to the housing wall thickness, is much higher. As expected, a higher coefficient of friction in the bearing seat leads to a reduction of the creeping torque. Therefore, arrangements to increase the coefficient of friction (e.g., surface coating) are recommended. Possible solutions are presented in [8].

Figure 8 shows the influence of the bearing parameters on the creeping behaviour by comparing the results with different thickness ratios of the outer ring, Q_b, the numbers of rollers, Z, and the surface area-normalised radial loads, p_r. The results indicate that thick bearing rings ($Q_b = 0.85$) have less inclination to creep than thin ones because of the increased stiffness of thick-walled bearing rings. The thick-walled bearing ring exhibits less deformation (see Chap. 2), and the creeping process is inhibited.

In addition, reduction of the radial load leads to a decrease in the creeping torque. This relationship is based on the fundamental processes of creeping, i.e., creeping occurs primarily from the formation of slip in the bearing seat as a result of the tangential deformation of the loaded bearing ring (see Chap. 2). If the load is small, the tangential deformation and thus the creeping torque are also small.

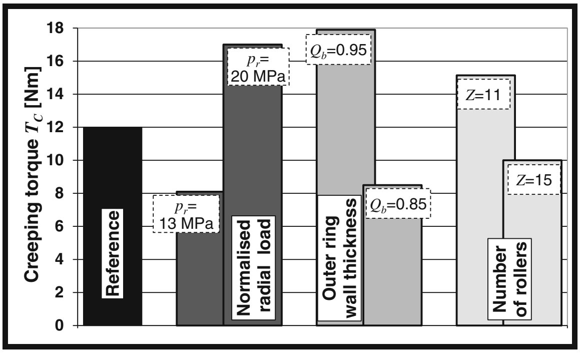


Fig. 8 Change of the creeping torque with variations of the thickness ratio of the outer ring Q_b, the number of rollers Z, and the surface area-normalised radial load p_r. (Reference data: Bearing NU 205, $p_r = 18$ MPa, coefficient of friction in the bearing seat $\mu = 0.3$, normalised clearance fit, $\xi* = -0.4$ ‰, modulus of elasticity of the housing, $E = 210$ GPa, $Z = 13$, $Q_b = 0.90$)

Fig. 9 Change of the creeping torque due to variation of the normalised clearance fit ($\xi* < 0$) and the normalised interference fit ($\xi \geq 0$) between the housing and the outer ring

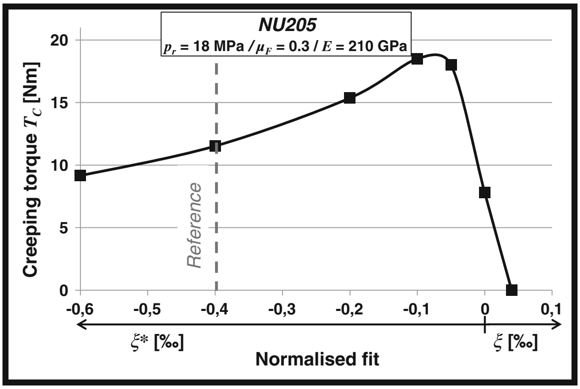


Figure 8 also shows that the increase in the number of rollers, with an identical bearing load, results in a decreased creeping torque. This behaviour is caused by the reduced roller loads, which reduces the tangential deformation of the bearing ring.

In Fig. 9, the effect of the normalised interference fit, ξ, and the clearance fit, $\xi*$, between the housing and the outer ring on the creeping behaviour is shown. The results indicate that an increase of the normalised clearance fit from $\xi^* = -0.1$ ‰ to $\xi^* = -0.6$ ‰ results in a reduction of the creeping torque.

In [3, 7], the simulated creeping behaviour is verified experimentally. The effect of the clearance fit can be explained by the modification of the load zone (Fig. 10).

Increasing the clearance fit (and/or the bearing clearance) leads to a reduction of the load zone, i.e., fewer rollers are used to transfer the bearing load. The load transfer is thus based on a smaller contact surface area and a higher contact

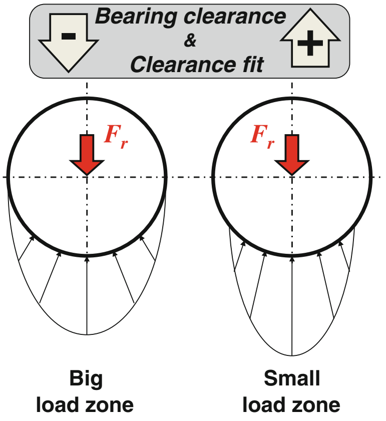


Fig. 10 Different load zones in relation to the clearance fit

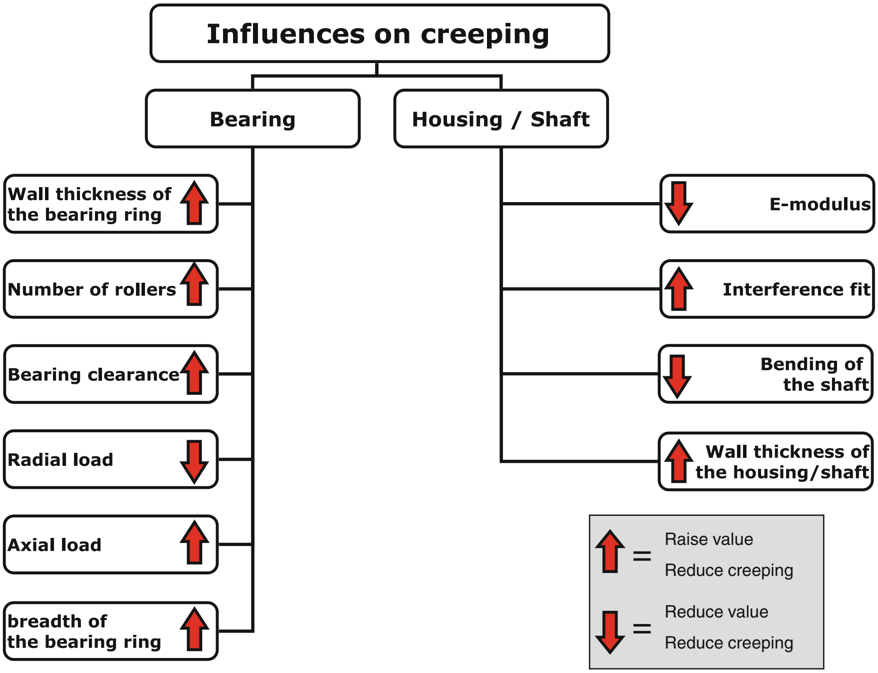


Fig. 11 Effects of the relevant bearing parameters on creeping [4]

pressure, thus increasing the transferable local contact shear stress. The reduction in stress leads to a reduction in or the prevention of creeping.

In addition to the parameter analysis presented, other influences on creeping were determined in [4, 9]. Those studies developed recommendations for bearing seat design. An overview of those solutions is shown in Fig. 11.

5 Comparison of the Creeping Propensities of Various Bearing Types

In determining the creeping propensities in addition to the variation of the bearing parameters (Chap. 4), the assessment of the different bearing types is similarly important. Therefore, commercially available single-row radial bearings are evaluated for their creeping limit and are subsequently compared. To determine the creeping limit, the radial load of the bearing, p_r, is continuously reduced until the creeping torque is equal to zero. The normalised radial load at the creeping limit $p_{r,lim}$ serves as a comparative value between the different bearing types. For the investigations of the bearing type, geometrically identical bearings (d_j and $b =$ constant) were selected and simulated under identical conditions. In Fig. 5, the results for all of the investigated bearings are compared. The results indicate that under the given conditions, the tapered roller bearing 30205 and the angular contact ball bearing 7205 have the highest creeping limit and thus the slightest creeping propensity. This behaviour is due to the bearing type specific additional axial force, which generates higher transferable shear stresses due to the necessary additional axial contact. The tapered roller bearing also exhibits the slightest increase of the creeping torque after transgression of the creeping limit $p_{r,lim}$. This behaviour suggests that harmful creeping damage is expected only at very high bearing loads. The cylindrical roller bearing NU205 and spherical roller bearing 20205 have lower creeping limits and are therefore more susceptible to creep. The worst is the deep groove ball bearing 6205 (Fig. 12).

6 Analysis of the Influence of the Bearing Size on the Creeping Behaviour

The influence of the bearing size on the creeping behaviour is another important aspect in the correct bearing selection by the design engineer. Therefore, different sizes of cylindrical roller bearings were simulated under identical conditions. To assess the creeping propensities of the different bearing sizes, the normalised radial load at the creeping limit $p_{r,lim}$ was used again (see Chap. 5). The results in Fig. 13 indicate that bearings up to $d_J = 180$ mm have an increasing limit load, $p_{r,lim}$, with increasing size. As a result, larger bearings tend to not be as vulnerable as smaller bearings in the same design. One exception is the largest simulated cylindrical roller bearing NU29/530. Due to the thin-walled bearing rings of NU29/530, it has a significantly lower limit load $p_{r,lim}$. A size effect is therefore present, but no general statement about the trend can be made.

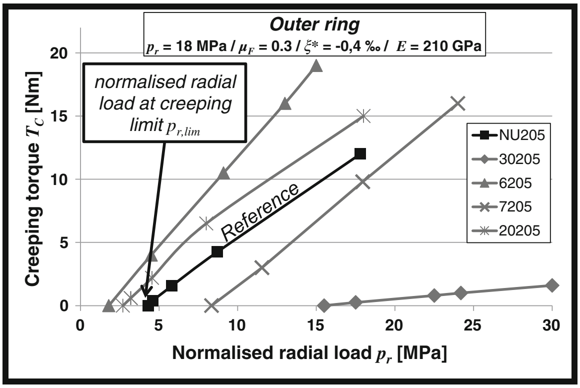


Fig. 12 Change of the creeping torque due to variation of the normalised radial load for different bearing outer rings

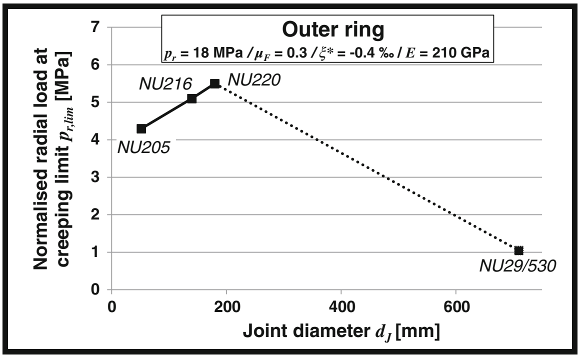
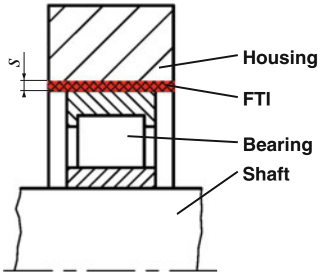


Fig. 13 Change of the normalised radial load at the creeping limit $p_{r,lim}$ due to variation of the joint diameter d_J for cylindrical roller bearings

7 Additional Constructive Capabilities to Reduce Creeping

As a possible remedy against walking (based on the results presented in Chap. 4, i.e., that the creeping moment decreases by a reduction in the modulus of elasticity), a flexible thin-film interlayer (FTI) between the housing and the bearing ring was simulated (Fig. 14). The housing is made of a partitioned continuum representing the two material areas of the FTI and regular housing material. Thus, the contact surfaces between the FTI and the housing are neglected, which would correspond to bonding of the contact materials.

The simulation results presented in Fig. 15 indicate that as the modulus of elasticity of the FTI decreases, the creeping torque also decreases. With the use of polyamide as the FTI, the reduction can be as large as 40 %. The opposite trend for the polyamide-FTI with the larger wall thickness, s = 300 μm (compared to s = 200 μm), is due to the reduced tangential stiffness for the thicker FTI. The polyamide is strongly deformed by the operating loads, and as a result of the model structure (locking of the outer ring in the circumferential direction, cf. Fig. 3), the creeping torque increases. The simulation of magnesium as the FTI material



Material	E-Modulus of the FTI E_{FTI} [GPa]
Magnesium	42
Polyamide	4

Fig. 14 Schematic diagram of the FTI (*left*) and the material characteristics of the FTI (*right*)

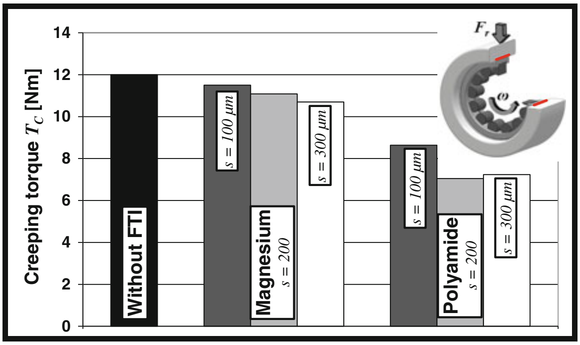


Fig. 15 Comparison of the creeping torque for varying wall thickness *s*, and material properties of the FTI. (Reference data: Bearing NU 205, normalised radial load $p_r = 18$ MPa, coefficient of friction in the bearing seat $\mu = 0.3$, normalised clearance fit, $\xi^* = -0.4$ ‰, modulus of elasticity of the housing $E = 210$ GPa)

indicates the limitations of the FTI, achieving only a marginal reduction of the creeping torque.

For heavily loaded bearings and critical creeping applications, a form-fitted connection between the bearing ring and the housing, including the shaft, is essential. Figure 16 shows the circumferential forces between the outer ring and the housing, which must be absorbed by a form-fitted creeping lock. The creeping lock was researched as a rigid element (the reference) and two elastic spring elements. Each spring element can be described by a linear spring characteristic curve, which is defined by a spring rate, $c = F_C/\Delta l$. The spring deflection, Δl, is thus equivalent to the relative movement between the housing and the outer ring (global slip). The results indicate that a soft form-fitted creeping lock has to take substantially lower loads than a rigid one. Further studies on the subject are part of the ongoing German research project Abhilfemaßnahmen Wandern (remedies for creeping) [10].

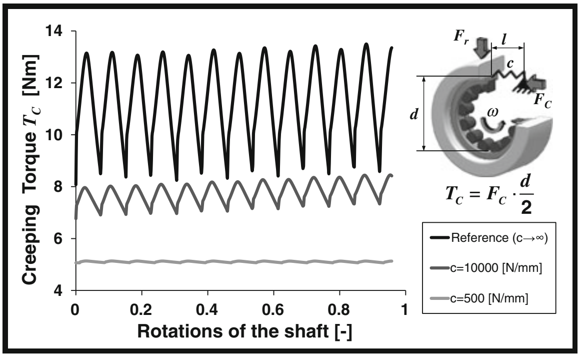


Fig. 16 Creeping torque as a function of the elasticity of the form-fitted creeping lock between the outer ring and the housing. (Reference data: Bearing NU 205, normalised radial load $p_r = 18$ MPa, coefficient of friction in the bearing seat $\mu = 0.3$, normalised clearance fit $\xi^* = -0.4$ ‰, modulus of elasticity of the housing $E = 210$ GPa)

8 Calculation Methods to Determine the Radial Load at the Creeping Limit

A fundamental goal for the practical implementation of the results is the development and verification of a calculation model to determine the critical creeping load of a bearing. The model should provide a rough estimate of the bearing load when approaching the onset of creeping. Implementation of this model should limit the use of complex, computationally intensive 3D FE simulations in the bearing design. The following algorithm is used to calculate the critical creeping load of roller bearings. It states that creeping starts when a load-induced slip zone extends over the entire width of the bearing ring. Figure 17 shows three different slip zones with variation of the radial load of the bearing. The left part of the picture shows that, due to the low load only slip on the edges of the joint, the bearing does not creep. In the middle of Fig. 17, the initial joint broad-slip zones are spreading due to higher loads; the bearing creeps are minor here. On the right site, the large slip zones indicate a strongly creeping bearing.

In [4, 6, 9], a guide for the approximate analytical calculation of the creeping conditions in the bearing seat is presented. Unfortunately, slip cannot be calculated using analytical equations. Therefore, the calculation of the inception of slip is performed with the radial stress ($\sigma_{rr}(F_{i,}\varphi)$) and the shear stress ($\tau(F_{i,}\varphi)$). However, this method allows only a rough approximation of the creeping limit.

To provide a more precise method of determining the real slip, the FE-based calculation tool *SimWag* was created. *SimWag* is based on the freeware finite element (FE) program Z88, which can be used to perform 3D contact simulations. This tool permits a considerable extension of the structures that can be simulated compared with the analytical 2D solution (implementation of the bending influence on the inner ring and the axial bearing forces, taking into account real bearing

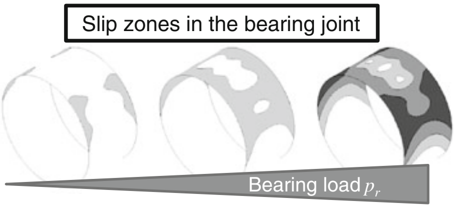


Fig. 17 Formation of slip zones by increasing the radial load

geometries, etc.). *SimWag* thus allows the user to calculate exact results and enables the user to construct bearing seats without creeping. The program is fully automated, so FE-knowledge is not required.

The calculation model *SimWag*, along with other algorithms (calculation of the creeping torque), have been programmed by the German research association Forschungsvereinigung Antriebstechnik e.V. (FVA).

9 Summary

In this paper, numerical analyses of the creeping behaviour of roller bearings are presented. With the aid of various complex 3D finite element kinematic simulations, the creeping mechanism is determined. Creeping describes flexing micromovements (slip) of the loaded bearing ring. This slip leads to a substantial continuous rotation of the bearing ring relative to the connection geometry (housing or shaft). Through the studies performed, important influencing parameters that cause or encourage creeping were determined.

Furthermore, approaches for geometric and constructive remedies for reducing or eliminating creeping are discussed. Finally, an algorithm for the analytical determination of the creeping critical load of the outer ring is presented.

Acknowledgments The authors thank the German organisations Forschungsvereinigung Antriebstechnik e.V. (FVA), Bundesministerium für Wirtschaft und Technologie (BMWi) and Arbeitsgemeinschaft industrieller Forschungsvereinigungen “Otto von Guericke” e.V. (AiF) for the financial support of the projects FVA 479 I, IGF-Nr. 15652 BG and IGF-Nr. 16987 BG.

References

1. E. Leidich, V. Walter, A. Maiwald, Relativbewegungen von Wälzlagerringen. J. Antriebstechnik **11**, 70–76 (2009) (Vereinigte Fachverlage, Mainz)
2. E. Leidich, A. Maiwald, A. Gacka, Wenig wandern, länger leben. J. Antriebstechnik **06**, 18–21 (2012) (Vereinigte Fachverlage, Mainz)
3. T. Babbick, Wandern von Wälzlagerringen unter Punktlast. Doctoral thesis, TU Kaiserslautern (2012)
4. E. Leidich, B. Sauer, A. Maiwald, T. Babbick, Beanspruchungsgerechte Auslegung von Wälzlagersitzen unter Berücksichtigung von Schlupf- und Wandereffekten. J. Nr. 956 (2010) (Forschungsvereinigung Antriebstechnik e.V, Frankfurt/M)

5. DIN ISO 281, Wälzlager—Dynamische Tragzahlen und nominelle Lebensdauer (Beuth Verlag, Berlin, 2009)
6. E. Leidich, A. Maiwald, FE simulations of irreversible relative movements (creeping) in rolling bearing seats. *Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering and Computer Science 2013, WCECS 2013*, 23–25 October, 2013, San Francisco, USA, pp. 1030–1035 (2013)
7. E. Aul, Analyse von Relativbewegungen in Wälzlagersitzen. Doctoral thesis, TU Kaiserslautern (2008)
8. A. Maiwald, E. Leidich, Einflussfaktoren auf das tribologische Verhalten von biegefreien Wälzlagersitzen bei Relativbewegungen infolge Wandern. *Proceedings of the GfT Tribology-Meeting*, issue 51, Aachen (2010), pp. 31.1–31.19
9. A. Maiwald, Numerische Analyse des Wanderverhaltens von Wälzlagerringen. Doctoral thesis, TU Chemnitz (submitted 2013)
10. E. Leidich, B. Sauer, T. Schiemann, S. Thiele, Definition und Auslegung von konstruktiven und tribologischen Abhilfemaßnahmen gegen tangentiale Wanderbewegungen von Wälzlagerringen. Frankfurt/M.: Forschungsvereinigung Antriebstechnik e.V., Research project FVA 479 IV, 2012, running

Chapter 42
Erlang Distribution and Exponential Distribution Models in Wireless Networks

Lela Mirtskhulava, Giorgi Gugunashvili and Mzia Kiknadze

Abstract Two mathematical models of wireless networks are analyzed in the given chapter. We demonstrate that the Erlang family provides more flexibility in modeling that exponential family, which only has one parameter. For this purposes one model has special Erlang distribution and second one is using exponential distribution. In practical situations, the Erlang family provides more flexibility in fitting a distribution to real data that the exponential family provides. The Erlang distribution is also useful in queueing analysis because of its relationship to the exponential distribution. To demonstrate the applicability of the Erlang distribution, we consider two queueing models, represented radio channels where the interarrival times between failure have the Erlang Distribution for FIRS model and Exponential distribution for second model.

Keywords Erlang distribution · Interarrival time between failures · Probabilistic approach · Queueing model

L. Mirtskhulava (✉) · G. Gugunashvili
Department of Computer Sciences, Ivane Javakhishvili Tbilisi State University,
13, University str, 0186 Tbilisi, Georgia
e-mail: lela.mirtskhulava@tsu.ge

G. Gugunashvili
e-mail: g.gugunashvili@gmail.com

M. Kiknadze
Department of Computer Engineering, Georgian Technical University,
76, Kostava str, 0175 Tbilisi, Georgia
e-mail: m.kiknadze@gtu.ge

H. K. Kim et al. (eds.), *Transactions on Engineering Technologies*,
DOI: 10.1007/978-94-017-9115-1_42,

1 Introduction

An accurate estimation of network performance is vital for the success of a network of any kind. Networks, whether voice or data, are designed around many different variables [1]. Two of the most important factors that you need to consider in the network design are service and cost. Service is essential for maintaining customer satisfaction. Cost is always a factor in maintaining profitability [2–5]. One way that you can factor in some of the service and cost elements in network design is to optimize circuit utilization.

Also to a large extent, the success of a network depends on the development of effective congestion control techniques that allow for optimal utilization of a network's capacity. Performance modeling is necessary for deciding the type of congestion control policies to be implemented. Performance models in turn, require very accurate traffic models that have the ability to capture the statistical characteristics of the actual traffic on the network [6, 7].

The design of robust and reliable networks and network services is becoming increasingly difficult in today's world. The only path to achieve this goal is to develop a detailed understanding of the traffic characteristics of the network [8].

Managing performance of networks involves optimizing the way networks function in an effort to maximize capacity, minimize latency and offer high reliability regardless of bandwidth available and occurrence of failures. Network performance management consists of tasks like measuring, modeling, planning and optimizing networks to ensure that they carry traffic with the speed, capacity and reliability that is expected by the applications using the network or required in a particular scenario.

Networks are of different types and can be categorized based on several factors. However, the factors that affect the performance of the different networks are more or less the same [9].

If the underlying traffic models do not efficiently capture the characteristics of the actual traffic, the result may be the under-estimation or over-estimation of the performance of the network. This would totally impair the design of the network. Traffic Models are hence, a core component of any the performance evaluation of networks and they need to be very accurate. Depending upon the type of network and the characteristics of the traffic on the network, a traffic model can be chosen for modeling the traffic [10–12].

2 Traffic Models and Erlang Distribution

Special Erlang Distribution lends itself well to modeling packet interarrival time for a number of reasons. The first is the fact that The Erlang distribution is a continuous probability distribution with wide applicability primarily due to its relation to the exponential and Gamma distributions. The exponential function is a strictly decreasing function of t. This means that after an arrival has occurred, the amount of waiting time until the next arrival is more likely to be small than large.

To begin modeling we define T as fixed length of packet. Simple traffic consists of single arrivals of discrete entities, packets. This kind of traffic can be expressed mathematically as a Point Process. Point processes can be described as a Counting Process or Inter-Arrival Time (IAT) Process.

We also assume that T is the length of the messages, let n be the number of packets in the message; we define l to be a number of phases in the Erlang distribution of failures, i.e. there is the scheme of failures arrival, according to which the failures must go through l phases (stages), before they actually will arrive; $F(u) = 1(t - \tau_b)$—Distribution function (DF) of packet length with a cyclic check redundancy (CRC), where 1(t)—unit function, and $\tau_b = T/n$ block length; r—the number of allowed repetitions of transmitting a block, G(u)—Distribution function of recovery time; α—the distribution intensity of each phase, i.e. the duration of time intervals between subsequent moments of occurrence of failure follow the Erlang Distribution given by:

$$A(u) = \frac{\alpha(\alpha u)^{l-1} e^{-\alpha u}}{(l-1)!} \tag{1}$$

Our task is to find Distribution Function of transfer time of fixed length message depending on number of the packets included in it and number of their retransmissions for a given network characteristics (Fig. 1).

We denote by $\Phi_{\mathrm{j}}^{(kv)}(t, x, T)$ the probability that the message transfer of fixed length of T (consisting of n packets, each of length of τ_b) starting from jth block will be completed in a time less than t under the condition that: (1) at the $t = 0$ moment data channel (DCH) is in K phase according to failures and xth part ($x \in [0, \tau_{\sigma}]$) of jth block has been transferred without distortion; (2) An attempt was made to transfer the j—the block without distortion v times.

By definition:

$$\Phi_j^{(kv)}(t, x, T) = \begin{cases} \Phi_{\mathrm{j+1}}^{(kv)}(t, 0, T) & \text{under } x = \tau_b \\ 0 & \text{under } x > \tau_b \end{cases}$$
$$k = \overline{1, l}, \ v = \overline{1, r}, \ j = \overline{1, n}; \tag{2}$$

We denote by $\Phi(t, T)$ the DF of the probability of transfer time of fixed length message ($T = n\tau_{\sigma}$) and by $\tilde{F}(u)$—Df of random length message, then we have the following notation:

$$\Phi(t) = \int_0^{\infty} \Phi(t, u) d\tilde{F}(u) \tag{3}$$

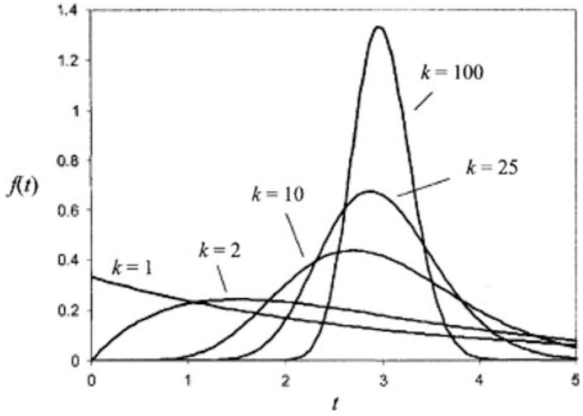


Fig. 1 A family of Erlang distributions with mean 3

where

$$\Phi(t,u) = \sum_{v=1}^{r}\sum_{k=1}^{l}\Phi_1^{(kv)}(t,0)/l \tag{4}$$

is DF od transfer time of fixed length message of u.

The model mentioned above can be described by the following system of equations:

$$\begin{aligned}\bar{F}(x)\Phi_j^{(kv)}(t,x) = &\int\limits_0^t e^{-\alpha u}\Phi_{j+1}^{k1}(t-u,0,T)F(x+u)du \\ &+\int\limits_0^t \alpha e^{-}\bar{F}(x+u)\Phi_j^{(k+1,v)}(t-u,x+u)du\end{aligned} \tag{5}$$

$$\begin{aligned}\Psi_j^{(lv)}(t,x) = &\int\limits_0^t e^{-\alpha u}\Phi_{j+1}^{(l1)}(t-u,0)d_uF(x+u) \\ &+\sum_{k=1}^{l}\sum_{p=0}^{\infty}\int\limits_0^t \alpha e^{-\alpha u}du\int\limits_0^{t-u} d_\upsilon F(x+u+\upsilon)\left[B_*^{(pl+k-1)}(\upsilon)+B_*^{(pl+k)}(\upsilon)\right] \\ &\Phi_j^{(k,v+1)}(t-u-\upsilon,0) \\ &v=\overline{1,r-1};\quad j=\overline{1,n};\end{aligned}$$

$$\begin{aligned}\Psi_j^{(lr)}(t,x) &= \int\limits_0^t e{-}^{\alpha u}\Phi_{j+1}^{(l1)}(t-u,0)d_uF(x+u) \\ &\quad + \int\limits_0^t (1-e^{-\alpha u})d_uF(x+u)\int\limits_0^{t-u}\Phi_j^{(1,1)}(t-u-\upsilon,0)dG(\upsilon)\end{aligned} \tag{7}$$

where $B_*^{(k)}(\upsilon)$ _k-fold convolution and $B(\upsilon) = 1 - e^{\alpha\upsilon}$;

$$\begin{gathered}\psi_j^{(kv)}(t,x) = \bar{F}(x)\Phi_j^{(kv)}(t,x,T); \\ \bar{F}(x) = 1 - F(x); \quad F(\tau_b^+) = 1; F(\tau_b^-) = 0\end{gathered}$$

Let us assume

$$\begin{gathered}\psi_j^{(kv)}(t,x) = \begin{cases}\Phi_{\mathrm{j}}^{(\mathrm{k}v)}(t,0), & x = 0 \\ \Phi_{\mathrm{j+1}}^{(\mathrm{k}v)}(t,0), & x = \tau_b \\ 0, & x > \tau_b\end{cases} \\ k = \overline{1,l},\ v = \overline{1,r},\ j = \overline{1,n};\end{gathered} \tag{8}$$

The boundary conditions have the form:

$$\Psi_{n+1}^{(kv)}(t,x) = \Phi_{n+1}^{(kv)}(t,0) = 1,\ k = \overline{1,l},\ v = \overline{1,r},\ j = \overline{1,n};$$

By integrating the product of the probabilities of events according to u, υ and summing the probabilities of incompatible events, we obtain (6).

Using the Laplace Transform in (6), we obtain:

$$\begin{aligned}\int\limits_0^\infty e^{-st}dt\int\limits_0^t e^{-\alpha u}d_uF(x+u)\Phi_{\mathrm{j+1}}^{(\mathrm{k1})}(t-u,0) &= \int\limits_0^\infty e^{-\alpha u}d_uF(x+u)\int\limits_u^\infty e^{-st}\Phi_{\mathrm{j+1}}^{(\mathrm{k1})}(t-u,0)dt \\ &= \int\limits_0^\infty e^{-\alpha u}d_uF(x+u)\int\limits_0^\infty e^{-s(z+u)}\Phi_{\mathrm{j+1}}^{(\mathrm{k1})}(z,0)dZ = \Phi_{\mathrm{j+1}}^{(\mathrm{k1})}(s,0) \\ &\int\limits_0^\infty e^{-(\alpha+s)u}d_uF(x+u) = e^{-(\alpha+s)(\tau_{\mathrm{b}}-x)}\Phi_{\mathrm{j+1}}^{(\mathrm{k1})}(s,0) \\ &\mathrm{xu} = \tau_6; \quad \mathrm{t-u=z}; \quad \mathrm{t=z+u}.\end{aligned} \tag{9}$$

$$\begin{aligned}\alpha\int\limits_0^\infty e^{-st}dt\int\limits_0^\infty e^{-\alpha u}\psi_j^{(k+1,v)}(t-u,x+u)du &= \int\limits_0^\infty e^{-\alpha u}du\int\limits_0^\infty e^{-st}\psi_j^{(k+1,v)}(t-u,x+u)dt\\ &= \alpha\int\limits_0^\infty e^{-\alpha u}du\int\limits_0^\infty e^{-s(u+z)}\psi_j^{(k+1,v)}(z,x+u)dZ\\ &= \alpha\int\limits_0^\infty e^{-(s+\alpha)u}du\int\limits_0^\infty e^{-sz}\psi_j^{(k+1,v)}(z,x+u)dZ\\ &= \alpha\int\limits_0^\infty e^{-(s+\alpha)u}d_u\psi_j^{(k+1,v)}(s,x+u)\\ &= \alpha\int\limits_0^\infty e^{-(s+\alpha)(\tau-x)}\psi_j^{(k+1,v)}(s,\tau)d\tau\\ &= \alpha e^{-(s+\alpha)u}\int\limits_x^\infty e^{-(s+\alpha)\tau}\psi_j^{(k+1,v)}(s,\tau)d\tau\end{aligned} \tag{10}$$

Denoting by:

$$\bar{\Psi}_j^{(kv)}(s,x) = \int\limits_0^\infty e^{-st}\psi_j^{(kv)}(t,x)dt; \tag{11}$$

$$\bar{\Phi}_j^{(kv)}(s,x) = \int\limits_0^\infty e^{-st}\Phi_j^{(kv)}(t,x)dx; \tag{12}$$

We obtain:

$$\begin{aligned}\psi_j^{(\mathrm{K}v)}(s,x) = {} & e^{-(\alpha+s)(\tau_b-x)}\Phi_{\mathrm{j+1}}^{(\mathrm{k1})}(s,0)\\ & + \alpha e^{-(\alpha+s)x}\int\limits_x^\infty e^{-(\alpha+s)\tau}\psi_j^{(k+1,v)}(s,\tau)d\tau\end{aligned} \tag{13}$$

$$e^{(s+\alpha)x}\psi_j^{(kv)}(s,x) = e^{-(\alpha+s)\tau_b}\Phi_{\mathrm{j+1}}^{(\mathrm{k1})}(s,0) + \alpha\int\limits_x^\tau be^{-(\alpha+s)\tau_b}\psi_j^{(k+1,v)}(s,\tau)d\tau. \tag{14}$$

Moving on to differential Eq. (14), we obtain:

$$\begin{gathered}\frac{d\bar{\Psi}_j^{(kv)}(s,x)}{dx} - (s+\alpha)\bar{\Psi}_j^{(k,v)}(s,x) + \alpha\bar{\Psi}_j^{(k+1,v)}(s,x) = 0 \\ \bar{\Psi}_j^{(kv)}(s,0) = \bar{\Phi}_j^{(kv)}(s,0); \quad j=\overline{1,n}, \quad k=\overline{1,l-1}, \quad v=\overline{1,r}\end{gathered} \tag{15}$$

$\psi_j^{(\mathrm{kv})}(s,x)$—monotonic and continuous function according to x $0^+ \leq x \leq \bar{\tau}_b$, in the range

$$\psi_j^{(kv)}(s,x) = 0, \quad \text{under} \quad \mathrm{x} \geq \tau_b^+$$

We apply the Laplace transform to (6). For this purpose, first of all, let $\mathrm{u+v=y}$; $\mathrm{dv=dy}$.

We rewrite (6) as follows:

$$\begin{aligned}\psi_j^{(lv)}(t,x) = &\int d_u F(x+u)e^{-\alpha u}\Phi_{\mathrm{j+1}}^{(11)}(t-u,0) \\ &+ \sum_{k=1}^{l}\sum_{p=0}^{\infty}\alpha e^{-\alpha u}du\int d_y F(x+y) \\ &\times [B_*^{(pl+k-1)}(y-u) - B_*^{(pl+k)}(y-u)]\Phi_{\mathrm{j}}^{(\mathrm{k},v+1)}(t-y,0)]\end{aligned} \tag{16}$$

and changing the order of integration in the second member, we get:

$$\begin{aligned}\psi_j^{(lv)}(t,x) = &\int_0^t e^{-\alpha u}\Phi_{\mathrm{j+1}}^{11}(t-u,0)du\mathrm{F(x+u)} \\ &+ \sum_{k=1}^{l}\sum_{p=0}^{\infty}\alpha\Phi_{\mathrm{j}}^{(\mathrm{k},v+1)}(s,0)\int_0^{\infty} e^{-sy}d_y F(x+y)\int_0^y e^{-\alpha u} \\ &\times \left[B_*^{(pl+k-1)}(y-u) - B_*^{(pl+k)}(y-u)\right]\end{aligned} \tag{17}$$

Applying the Laplace-Stieltjes, we get:

$$\begin{aligned}\bar{\psi}_j^{(lv)}(s,x) = &e^{-(\alpha+s)(\tau_b - x)}\psi_{\mathrm{j+1}}^{(11)}(s,0) \\ &+ \sum_{k=1}^{l}\sum_{p=0}^{\infty}\alpha\bar{\Phi}_{\mathrm{j}}^{(\mathrm{k},v+1)}(s,0)\int_0^{\infty} e^{-sy}d_y F(x+y) \\ &\times \int_0^y e^{-\alpha u}\left[B_*^{(pl+k-1)}(y-u) - B_*^{(pl+k)}(y-u)\right]du\end{aligned} \tag{18}$$

If we denote:

$$\varphi_1(s) = \int\limits_0^\infty e^{-sy} d_y F(x+u) = e^{-s(\tau_b - x)}$$

$$\varphi_2(s) = \int\limits_0^\infty e^{-sy} dy \int\limits_0^y \left[B_*^{(pl+k-1)}(y-u) - B_*^{(pl+k)}(y-u)\right] e^{-\alpha u} du$$

$$= \frac{1}{(s+\alpha)^2}\left(\frac{\alpha}{s+\alpha}\right)^{pl+k-1}$$

Thus, we have:

$$\bar{\Psi}_j^{(lv)}(s,x) = e^{-(\alpha+s)(\tau_b - x)} \Phi_{j+1}^{(l1)}(s,0) + \sum_{k=1}^{l} \sum_{p=0}^{\infty} \alpha \bar{\Phi}_j^{(k,v+1)}(s,0) \frac{1}{2\pi i} \int\limits_{C^*-i\infty}^{C^*+i\infty} \varphi_1(s-w)\varphi_2(w) dw \tag{19}$$

where C * is the abscissa of convergence of the improper integral, which lies in the domain of analyticity of under the integral sign function; $i = \sqrt{-1}$ or

$$\begin{aligned}\bar{\Psi}_j^{(lv)}(s,x) = {} & e^{-(\alpha+s)(\tau_b-x)} \Phi_{j+1}^{(l1)}(s,0) \\ & + \sum_{k}^{l-1} \Phi_j^{(k,v+1)}(s,0)\alpha^k \left[\frac{1}{2\pi i} \int\limits_{C-i\infty}^{C+i\infty} e^{-(s-w)(\tau_b-x)} \frac{(w+\alpha)^{l-1-k}}{(w+\alpha)^l - \alpha^l} dW\right] \\ & + \Phi_j^{(l,v+1)}(s,0)\alpha^l \left\{\frac{1}{2\pi i} \int\limits_{C-i\infty}^{C+i\infty} e^{-(s-w)(\tau_b-x)} \frac{dW}{(W+\alpha)\left[(w+\alpha)^l - \alpha^l\right]}\right\}\end{aligned} \tag{20}$$

Applying the Laplace-Stieltjes, we obtain

$$\bar{\Psi}_j^{(l,r)}(s,x) = e^{-(\alpha+s)(\tau_b-x)} \bar{\Phi}_{j+1}^{(l1)}(s,0) + g(s)\bar{\Phi}_j^{(1,1)}(s,0)\left(e^{-s(\tau_b-x)} - e^{-(\alpha+s)(\tau_b-x)}\right) \tag{21}$$

(20) and (21) contain [(l − 1)r + r]$^{\mathrm{j}}$ = lrj—equations, a number of unknowns is equal to 2jer,

$$\tilde{\psi}_j^{(kv)}(s,\tau_b) = \Phi_{\mathrm{j}}^{(\mathrm{k}v)}(s,0);$$
$$j = \overline{1,n}; \quad k = \overline{1,l-1}$$

By the method of deductions, we calculate (20):

$$\begin{aligned}\bar{\Psi}_j^{(lv)}(s,x) = {} & e^{-(\alpha+s)(\tau_b-x)}\bar{\Phi}_{j+1}^{(l1)}(s,0) \\ & + \sum_{k=1}^{l-1}\left[\bar{\Phi}_j^{(k,v+1)}(s,0)\alpha^k + \sum_{p=0}^{l} e^{-(s-w_p)(\tau_b-x)}(w_p+\alpha)^{l-1-k}\right. \\ & \left.\times\left(\prod_{\eta=1,\eta\neq p}^{l}(w_p-w_\eta)\right)\right]\bar{\Phi}_j^{(l,v+1)}(s,0)\alpha^l\sum_{p=0}^{l}e^{-(s-w_p)(\tau_b-x)}\left(\underset{\substack{\eta=1\\ \eta\neq p}}{\overset{l}{\Pi}}(w_p-w_\eta)\right)\end{aligned} \tag{22}$$

where $w_d = \alpha(e^{2\pi di/l} - 1),\ d = \overline{1,l}$—the zeros of the equation.

$$(w+\alpha)^l - \alpha^l = 0,\ \text{a } \mathrm{W}_0 = -\alpha;$$

To simplify further calculations, we introduce a new variable $\mathrm{y} = \tau_{\mathrm{b}} - \mathrm{x}$ (y—time remaining till the end of the block transfer). By substituting new variables we have

$$\frac{d\psi_j^{(kv)}(s,\tau_b-y)}{dy}\frac{dy}{dx} - (s+\alpha)\psi_j^{(kv)}(s,\tau_b-y) + \alpha\psi_j^{(k+1,v)}(s,\tau_b-y) = 0 \tag{23}$$

but taking into account that:

$$\psi_j^{(kv)}(t,x) = \psi_j^{(kv)}(t,\tau_b-y) = \tilde{\psi}_j^{(kv)}(t,y) = \bar{\psi}_j^{(\kappa v)}(s,y)$$

We have:

$$\frac{d\tilde{\psi}_j^{(kv)}(s,y)}{dy} + (s+\alpha)\tilde{\psi}_j^{(kv)}(s,y) - \alpha\tilde{\psi}_j^{(k+1,v)}(s,y) = 0 \tag{24}$$

$$\psi_j^{(kv)}(t,0) = \tilde{\psi}_j^{(kv)}(t,\tau_b) = \Phi_{\mathrm{j+1}}^{(kv)}(t,0) = 1; \quad t \geq 0.$$
$$\tilde{\psi}_j^{(kv)}(s,0) = \psi_{j+1}^{(kv)}(s,\tau_b) = \Phi_{\mathrm{j+1}}^{(kv)}(s,0).$$

$$\begin{aligned}\bar{\Psi}_j^{(lv)}(s,y) = e^{-(\alpha+s)y}\bar{\Phi}_{j+1}^{(l1)}(s,0) &+ \sum_{k=1}^{l-1}\left[\bar{\Phi}_j^{(k,v+1)}(s,0)\alpha^k \sum_{p=0}^{l} e^{-(s-w_p)y}\right. \\ &\left.\times (w_p+\alpha)^{l-1-k}\left(\prod_{\eta=1,\eta\neq p}^{l}(w_p-w_\eta)\right)\right] \\ &+\left[\bar{\Phi}_j^{(l,v+1)}(s,0)\alpha^l \sum_{p=0}^{l} e^{-(s-w_p)y}\left(\mathop{\Pi}_{\substack{\eta=1\\ \eta\neq p}}^{l}(w_p-w_\eta)\right)\right]\end{aligned} \tag{25}$$

$$\bar{\bar{\Psi}}_j^{(l,r)}(s,y) = e^{-(\alpha+s)y}\bar{\Phi}_{j+1}^{(l1)}(s,0) + g(s)e^{-sy}(1-e^{-\alpha y})\bar{\Phi}_j^{(1,1)}(s,0) \tag{26}$$

Applying the Laplace transform of the argument y (respectively, operator ω) к ((24) and (25)) и (26), we obtain:

$$\begin{gathered}(\omega+s+\alpha)\tilde{\bar{\psi}}_j^{(\kappa v)}(s,\omega) = \bar{\Phi}_{j+1}^{(kv)}(s,0) + \alpha\tilde{\bar{\Psi}}_j^{(k+1,v)}(s,\omega) \\ j=\overline{1,n}, \quad k=\overline{1,l-1}, \quad v=\overline{1,r}\end{gathered} \tag{27}$$

$$\begin{aligned}\tilde{\Psi}_j^{(lv)}(s,\omega) = \Phi_{j+1}^{(l1)}(s,0)(\omega+s+\alpha)& \\ +\sum_{k=1}^{l-1}&\left\{\alpha^k\Phi_j^{(k,v+1)}(s,0)\sum_{p=1}^{l}\left[(w_p+\alpha)^{l-1+k}\Big/\mathop{\Pi}_{\substack{\eta=1\\ \eta\neq p}}^{l}(w_p-w_\eta)\right]\Big/(\omega+s-w_p)\right\} \\ +\alpha^l&\Phi_j^{(l,v+1)}(s,0)\sum_{p=0}^{l}\left\{\left[1\Big/\mathop{\Pi}_{\substack{\eta=0\\ \eta\neq p}}^{l}(w_p-w_\eta)\right]\Big/(\omega+s-w_p)\right\} \\ v=\overline{1r-1},& \quad w_d=\alpha\left(e^{\frac{i2\pi d}{l}}-1\right), \quad d=\overline{1,l}, \quad w_0=-\propto\end{aligned} \tag{28}$$

$$\tilde{\Psi}_j^{(lr)}(s,\omega) = \Phi_{j+1}^{(l1)}(s,0)/(\omega+s+\alpha) + \{\alpha\bar{g}(s)/[(s+\omega)(s+\omega+\alpha)]\}\Phi_j^{(1,1)}(s,0) \tag{29}$$

$$\Phi_{(n+1)}^{(kv)}(s,0) = \frac{1}{s} \tag{30}$$

Here

$$
\begin{aligned}
\Psi_j^{(kv)}(t,x) =&\Psi_j^{(kv)}(t,\tau_b - y) = \tilde{\Psi}_j^{(kv)}(t,y)\\
\tilde{\bar{\Psi}}_j^{(kv)}(s,y) =& \int\limits_0^\infty e^{-st}\tilde{\Psi}_j^{(kv)}(t,y)dt\\
\tilde{\Psi}_j^{(kv)}(t,0) =&\Psi_j^{(kv)}(t,\tau_b) = \Phi_{j+1}^{(kv)}(t,0)\\
\tilde{\bar{\Psi}}_j^{(kv)}(s,0) =&\Phi_{j+1}^{(kv)}(s,0)\\
\tilde{\bar{\Psi}}_j^{(kv)}(s,\omega) =& \int\limits_0^\infty e^{-\omega y}\tilde{\bar{\Psi}}_j^{(kv)}(s,y)dy
\end{aligned}
$$

As a result of solutions of algebraic Eqs. (27), (28), (29) and taking into account (30) we find $\bar{\tilde{\psi}}_j^{(\kappa v)}(s,\omega) \quad (j = \overline{1,n};\ k = \overline{1,l};\ v = \overline{1,r})$ Defining the reverse conversion $\bar{\tilde{\psi}}_j^{(kv)}(s,\omega)$ and after substituting in it y = 0 (x = τ_b), we determine the value of. Let us solve the Eq. (27) by successive substitutions, starting from k = l − 1, k = l – 2, …, let l − m = k; m = l − k, then we have:

$$
\tilde{\psi}_j^{(kv)} = \frac{\sum_{c=1}^{l-k-1}\left[(\omega+s+\alpha)^{l-k-c}\alpha^{(c-1)}\right] + \alpha^{l-k-1} + S\alpha^{l-k}\tilde{\psi}_j^{(lv)}(s,\omega)}{S(\omega+S+\alpha)^{l-k}} \tag{31}
$$

Let us calculate the following Eqs. (28), (29) and (31), we give an example for this, where l = 2, r = 2, n = 1, j = 1,$\Phi_2^{(2\mathrm{i})}(s,0) = {}^1\!/_s$, I, i = 1,2.

$$
\tilde{\psi}_1^{(1i)}(s,\omega) = \frac{1 + S\alpha\tilde{\psi}_1^{(2i)}(s,\omega)}{S(\omega+S+\alpha)}, \quad i = 1,2.
$$

in accordance with (31)

$$
\tilde{\bar{\Psi}}_1^{(2,2)}(s,\omega) = \frac{1}{s(s+\omega+\alpha)}\left[\frac{\alpha g(s)}{(s+\omega)(s+\omega+\alpha)}\right]\bar{\Phi}_1^{(1,1)}(s,0)
$$

$$
\begin{aligned}
\tilde{\bar{\Psi}}_1^{(2,1)}(s,\omega) =&\frac{1}{s(s+\omega+\alpha)}\\
&+\alpha\bar{\Phi}_1^{(1,2)}(s,0)\left[\frac{1}{(w_1-w_2)(s+\omega-w_1)} + \frac{1}{(w_2-w_1)(s+\omega-w_2)}\right]\\
&+\alpha^2\bar{\Phi}_1^{(2,2)}(s,0)\left[\frac{1}{(w_0-w_1)(w_0-w_2)(s+\omega-w_0)}\right.\\
&\left.+\frac{1}{(w_1-w_0)(w_1-w_2)(s+\omega-w_1)} + \frac{1}{(w_2-w_0)(w_2-w_1)(s+\omega-w_2)}\right]
\end{aligned}
$$

$$\begin{aligned}\bar{\Phi}_1^{(21)}(s,0) = &\frac{1}{s}e^{-(s+\alpha)\tau_b} + \frac{1}{2}\left(e^{-s\tau_b} - e^{-(s+2\alpha)\tau_b}\right)\bar{\Phi}_1^{(12)}(s,0)\\ &+ \left[\frac{1}{2}e^{-s\tau_b} + \frac{1}{2}e^{-(s+2\alpha)\tau_b} - e^{-(s+\alpha)\tau_b}\right]\bar{\Phi}_1^{(2,2)}(s,0)\end{aligned}$$

Similar to:

$$\begin{aligned}\bar{\Phi}_1^{(2,2)}(s,0) &= A(s,T) + D(s,T)\bar{\Phi}_1^{(1,1)}(s,0);\\ \bar{\Phi}_1^{(1,2)}(s,0) &= A(s,T) + E(s,T)\bar{\Phi}_1^{(2,1)}(S,0);\\ \bar{\Phi}_1^{(1,2)}(s,0) &= A(s,T) + F(s,T)\bar{\Phi}_1^{(2,2)}(S,0);\end{aligned}$$

Conditional mean is equal to:

$$\begin{aligned}-\left|S\Phi_1^{(2,1)}(S,0)\right|'_{S=0} = \tau^{(21)}(0) = &\tau_b + \frac{1}{2}(1 - e^{-2\alpha\tau_b})\tau_1^{(1,2)}(0)\\ &+ \left(\frac{1}{2} + \frac{1}{2}e^{-2\alpha\tau_b} - e^{-\alpha\tau_b}\right)\tau_1^{(2,2)}(0);\end{aligned}$$

Solving the remaining equations, we finally obtain:

$$\Phi_1^{(1,1)}(S,0) = b(\alpha)K(\alpha)\frac{M(s)}{S}$$

$$\Phi_1^{(1,2)}(S,0) = \frac{b(\alpha)K(\alpha)e^{-s\tau_b}}{S}[1 + d(\alpha)g(s)M(s)]$$

$$\begin{aligned}\Phi_1^{(2,1)}(S,0) = &\frac{K(\alpha)e^{-s\tau_b}}{S}\{1 + \mathrm{b}(\alpha)\mathrm{e}(\alpha)\mathrm{e}^{-\mathrm{St_b}}\\ &\times [1 + d(\alpha)g(s)M(s)] + \mathrm{b}(\alpha)\mathrm{f}(\alpha)\mathrm{M(s)}\}\end{aligned}$$

$$\Phi_1^{(2,2)}(S,0) = \frac{K(\alpha)e^{-s\tau_b}}{S}[1 + b(\alpha)f(\alpha)M(s)]$$

$$\tau^{(1,1)}(0) = -\left|S\Phi_1^{(1,1)}(S,0)\right|'_{S=0} = b(\alpha)k(\alpha)M'(0)$$

$$\begin{aligned}\tau^{(1,2)}(0) &= -\left|S\Phi_1^{(1,2)}(S,0)\right|'_{S=0}\\ &= -\tau_b b(\alpha)k(\alpha)[1 - f(\alpha)M(0)] + b(\alpha)f(\alpha)k(\alpha)M'(0)\end{aligned}$$

$$\begin{aligned}\tau^{(2,1)}(0) = &- \left|S\Phi_1^{(2,1)}(S,0)\right|'_{S=0}\\ = &- k(\alpha)\tau_b[1 - 2b(\alpha)e(\alpha)] - k(\alpha)b(\alpha)\{e(\alpha)d(\alpha)[\tau_b + \tau_r]\\ &- \tau_b f(\alpha)\}M(0) + b(\alpha)k(\alpha)[e(\alpha)d(\alpha) + f(\alpha)]M'(0)\end{aligned}$$

$$\tau^{(2,2)}(0) = -\left|S\Phi_1^{(2,2)}(S,0)\right|'_{S=0}$$
$$= -\tau_b k(\alpha) - \tau_b k(\alpha)b(\alpha)f(\alpha)M(0) + b(\alpha)K(\alpha)f(\alpha)M'(0).$$

$$\mathrm{a}(\alpha) = 1/2 + e^{-\alpha\tau_b}[(1/2)e^{\alpha\tau_b} - 1]$$

where

$$b(\alpha) = 1 + \alpha\tau_b$$

$$c(\alpha) = 1/2 - e^{-\propto\tau_b}[(1/2)e^{-\propto\tau_b} - \alpha\tau_b]$$

$$d(\alpha) = 1 + e^{-\propto\tau_b}(1 - \alpha\tau_b)$$

$$e(\alpha) = (1/2)[1 - e^{-2\propto\tau_b}]$$

$$f(\alpha) = 1 - e^{-\propto\tau_b}$$

$$k(\alpha) = e^{-\propto\tau_b}$$

$$M(s) = \frac{[1 + e^{-s\tau_b}(a(\alpha)b(\alpha) + c(\alpha))]}{1 - e^{-2s\tau_b(d(\alpha)d(s)+f(\alpha))]}}$$

3 Model with Exponential Distribution

Under consumption (A) in the case of the arising errors (information distortion) the block transmission is repeated until the receiving of accurate information but no more r − 1 times. The channel is transferred for repair after r times repetition. The information transmission is received after the repair from distorted repetition.

The Radio channel works on information transmission, starting from jth block $(j = \overline{1,n})$ is described by the following equations:

$$\Phi_j^{(i)}(t) = \int_0^t dF_j(u)e^{-\alpha_j u}\Phi_{j+1}^{(1)}(t-u) + \int_0^t dF_j(u)[1 - e^{-\alpha_j u}]\Phi_j^{(i+1)}(t-u)$$

$$\Phi_j^{(r-1)}(t) = \int_0^t dF_j(u)e^{-\alpha_j u}\Phi_{j+1}^{(1)}(t-u) + \int_0^t dF_j(u)(1 - e^{-\alpha_j u})\int_0^{t-u} dG(\upsilon)\Phi_j^{(1)}(t-u-\upsilon)$$
$$(j = \overline{1,n}), \quad \Phi_{n+1}(t) = 1$$

Using the Laplace transform, we obtain:

$$\begin{aligned}
\bar{\Phi}_j^i(s) &= \bar{f}_j(s+\alpha_j)\bar{\Phi}_{j+1}^{(1)}(s) + \left[\bar{f}_j(s) - \bar{f}_j(s+\alpha_j)\right]\bar{\Phi}_j^{(j+1)}(s) \\
\bar{\Phi}_j^{(r-1)}(s) &= \bar{f}_j(s+\alpha_j)\bar{\Phi}_{j+1}^{(1)}(s) + \left[\bar{f}_j(s) - \bar{f}_j(s+\alpha_j)\right]\bar{g}(s)\bar{\Phi}_j^{(1)}(s) \\
\bar{\Phi}_{n+1}(s) &= \frac{1}{s}
\end{aligned}$$

The solution has the form:

$$\bar{\Phi}_1^{(1)}(S) = \frac{a^n(s)}{S[1-b(s)]^n}$$

Conditional mean time of task execution is calculated by the following equation

$$-T_c = s\bar{\Phi}_1^{(1)}(s)\left|\begin{matrix} 1 \\ s=0 \end{matrix}\right. = \left.\frac{a^n(s)}{[1-b(s)]^n}\right|'_{S=0}$$

4 Conclusion and Future Work

In this chapter, we investigated two queueing models, represented as wireless system, where time intervals between failures have Erlang distribution in first model and Exponential distribution in second one. We present some advantages of the Erlang model we proposed for mobility modeling. We show the generality of such model, which can be used to model not only interarriaval time between neighboring failures but also other time variables in wireless networks and mobile computing systems.

References

1. L. Mirtskhulava, Mathematical model of prediction of reliability of wireless communication networks, in *UKSim-AMSS 15th International Conference on Computer Modeling and Simulation*, Cambridge, UK, 10–12 April 2013. IEEE Trans. 677–681 (2013)
2. W. Jeon, D. Jeong, Call admission control for CDMA mobile communications systems supporting multimedia services. IEEE Trans. Wireless Commun. **1**(4), 649–659 (2002)
3. J.I. Sanchez, F. Bercelo, J. Jordon, Inter-arrival time distribution for channel arrivals in cellular telephony, in *Proceedings of the 5th International Workshop on Mobile Multimedia Comm. MoMuc'98*, 12–13 October 1998, Berlin, Germany
4. F. Khan, *LTE for 4G Mobile Broadband: Air Interface Technologies and Performance* (Cambridge University Press, New York, 2009)
5. S. Parkvall, E. Dahlman, J. Sköld, P. Beming, *3G Evolution: HSPA and LTE for Mobile Broadband* (Elsevier Publishers, 2nd edn., 2008

6. J.C. Ikuno, M. Wrulich, M. Rupp, TCP performance and modeling of LTE H-ARQ, in *Proceedings of the International ITG Workshop on Smart Antennas (WSA 2009)*, Berlin, Germany (2009)
7. B. Lin, S. Mohan, A. Noerpel, Queueing priority channel assignment strategies for handoff and initial access for a PCS network. IEEE Trans. Veh. Technol. **43**(3), 704–712 (1994)
8. Y. Fang, I. Chlamtac, Teletraffic analysis and mobility modeling for PCS networks. IEEE Trans. Commun. **47**(7), 1062–1072 (1999)
9. R.E. Haskell, C.T. Case, Transient signal propagation in lossless isotropic plasmas (Report style). USAF Cambridge Research Laboratory, Cambridge, MA Report ARCRL-66-234 (II), vol. 2 (1994)
10. G. Boggia, P. Camarda, N. Di Fonzo, Teletraffic analysis of hierarchical cellular communication networks. IEEE Trans. Veh. Technol. **52**(4), 931–946 (2003)
11. S. Bhattacharya, H.M. Gupta, S. Kar, Traffic model and performance analysis of cellular mobile systems for general distributed handoff traffic and dynamic channel allocation. IEEE Trans. Veh. Technol. **57**(6), 3629–3640 (2008)
12. Y. Iraqi, R. Boutaba, Handoff and call dropping probabilities in wireless cellular networks, in *Proceedings of International Conference on Wireless Networks, Communications and Mobile Computing (WIRELESSCOM '05)*, Maui, Hawaii, USA, June 2005, pp. 209–213

Chapter 43
On the Numerical Solution of Multi-dimensional Diffusion Equation with Non Local Conditions

Ahmed Cheniguel

Abstract In this paper, we investigate the solution of multi-dimensional diffusion equation using decomposition method. We consider two cases: a two-dimensional equation with non local boundary conditions and a three-dimensional equation with an integral condition. The method is reliable and gives a solution in a series form with high accuracy. It also guarantees considerable saving of calculation volume and time as compared to traditional methods. The obtained results show that the decomposition method is efficient and yields a solution in a closed form.

Keywords Adomian decomposition method · Diffusion equation · Exact solution · Integral condition · Non local boundary conditions · Partial differential equations

1 Introduction

Over the last few years, various processes in science and engineering have led to the non classical parabolic initial/boundary value problems which involve non-local integral terms over the spatial domain [1–11, 13, 15]. These include chemical diffusion, heat conduction, population dynamics and control. Up to now partial differential equations with non local boundary conditions have been one of the fastest growing areas in various fields. In this paper we consider a two-dimensional and three dimensional diffusion equation. The two-dimensional case was solved by many authors using traditional numerical techniques such as finite difference

A. Cheniguel (✉)
Faculty of Sciences, Department of Mathematics and Computer Science,
Kasdi Merbah University, Ouargla, Algeria
e-mail: cheniguelahmed@yahoo.fr

H. K. Kim et al. (eds.), *Transactions on Engineering Technologies*,
DOI: 10.1007/978-94-017-9115-1_43,

method, finite elements method, spectral techniques, etc. for example Siddique [8] proposed a fourth-order finite difference Padé scheme and Cheniguel [3] has solved the same problem using new techniques the obtained results are all exact.

The aim of this work is to study and to implement the decomposition method for solving the two and three dimensional diffusion equation [12, 14–16]. The decomposition method can also be applied to a large class of system of partial differential equations with approximates that converges rapidly to accurate solutions. The implementation of the method has shown reliable results in that few terms are needed to obtain either exact solution or to find an approximate solution of a reasonable degree of accuracy in real physical models. Numerical examples are presented to illustrate the efficiency of the decomposition method, the obtained results are in good agreement with exact ones.

2 The Two Dimensional Equation with Non Local Boundary Conditions

2.1 Problem Definition

We consider the two-dimensional diffusion equation given by

$$u_t = u_{xx} + u_{yy}, 0 < x, y < 1, t > 0. \tag{1}$$

Initial conditions are assumed to be of the form

$$u(x, y, 0) = f(x, y), (x, y) \in \Omega \cup \partial\Omega.$$

And the Dirichelet time-dependent boundary conditions are:

$$\begin{aligned} u(0, y, t) &= \psi_0(y, t), 0 \le t \le T, 0 \le y \le 1 \\ u(1, y, t) &= \psi_1(y, t), 0 \le t \le T, 0 \le y \le 1 \\ u(x, 0, t) &= \varphi_0(x)\gamma(t), 0 \le t \le T, 0 \le x \le 1 \\ u(x, 1, t) &= \varphi_1(x), 0 \le t \le T, 0 \le x \le 1. \end{aligned} \tag{2}$$

And non local boundary condition:

$$\int_0^1 \int_0^1 u(x, y, t) dxdy = m(t), (x, y) \in \Omega \cup \partial\Omega \tag{3}$$

where $f, \psi_0, \psi_1, \varphi_0, \varphi_1$ and m are known functions and $\gamma(t)$ is to be determined.

2.2 Adomian Decomposition Method

In this section, we outline the steps to obtain a solution to the above problem using Adomian decomposition method, which was initiated by Adomian [12, 14, 16]. For this purpose we reformulate the problem in an operator form:

$$L_t(u) = L_{xx}(u) + L_{yy} + L_{zz} \tag{4}$$

where the differential operators $L_t(.) = \frac{\partial}{\partial t}(.)$ and $L_{xx} = \frac{\partial^2}{\partial x^2}, L_{yy} = \frac{\partial^2}{\partial Y^2}$, assuming that the inverse L_t^{-1} exists and is defined as:

$$L_t^{-1} = \int_0^t (.)dt. \tag{5}$$

2.3 Solution Procedure

Applying inverse operator on both the sides of (4) and using the initial condition, yields:

$$u(x,y,t) = L_t^{-1}\big(L_{xx}(u(x,y,t)) + L_{yy}(u(x,y,t))\big)$$

or

$$u(x,y,t) = u(x,y,0) + L_t^{-1}\big(L_{xx}(u(x,y,t)) + L_{yy}(u(x,y,t))\big). \tag{6}$$

Now, we decompose the unknown function $u(x, y, t)$ as a sum of components defined by the series:

$$u(x,y,t) = \sum_{k=0}^{\infty} u_k(x,y,t) \tag{7}$$

where $u_0(x, y, t)$ is identified as $u(x, y,0)$. Substituting Eq. (7) into Eq. (6) one obtains

$$\sum_{k=0}^{\infty} u_k(x,y,t) = f(x,y) + L_t^{-1}\left\{L_{xx}\left(\sum_{k=0}^{\infty} u_k(x,y,t)\right) + L_{yy}\left(\sum_{k=0}^{\infty} u_k(x,y,t)\right)\right\}. \tag{8}$$

The components $u_k(x, y, t)$ are obtained by the recursive formula:

$$u_0(x,y,t) = f(x,y) \tag{9}$$

$$u_{k+1}(x,t) = L_t^{-1}\big(L_{xx}(u_k(x,y,t)) + L_{yy}(u_k(x,y,t))\big), k \geq 0. \tag{10}$$

From Eqs. (9) and (10) we obtain the first few terms as:

$$\begin{aligned}
u_0 &= f(x,y)\\
u_1 &= L_t^{-1}\big(L_{xx}(u_0(x,y,t)) + L_{yy}(u_0(x,y,t))\big)\\
u_2 &= L_t^{-1}\big(L_{xx}(u_1(x,y,t)) + L_{yy}(u_1(x,y,t))\big)\\
u_3 &= L_t^{-1}\big(L_{xx}(u_2(x,y,t)) + L_{yy}(u_2(x,y,t))\big)
\end{aligned}$$

and so on. As a result, the components u_0, $u_1, u_2, \ldots$are identified and the series solution is thus entirely determined. However, in many cases the exact solution in a closed form may be obtained as we can see in our examples.

3 The Three Dimensional Equation with Integral Condition

3.1 Problem Definition

We consider the three-dimensional diffusion equation given by:

$$u_t = u_{xx} + u_{yy} + u_{zz}, 0 < x, y, z < 1, t > 0. \tag{11}$$

Initial condition is given by:

$$u(x,y,z,0) = f(x,y,z), (x,y,z) \in \Omega \cup \partial\Omega.$$

And the Dirichelet time-dependent boundary conditions are

$$\begin{aligned}
u(0,y,z,t) &= \psi_0(y,z,t), 0 \leq y,z \leq 1, 0 \leq t \leq T\\
u(1,y,z,t) &= \psi_1(y,z,t), 0 \leq y,z \leq 1, 0 \leq t \leq T
\end{aligned} \tag{12}$$

$$\begin{aligned}
u(x,0,z,t) &= \varphi_0(x,z) \times \gamma(t), 0 \leq x,z \leq 1, 0 \leq t \leq T\\
u(x,1,z,t) &= \varphi_1(x,z,t), 0 \leq x,z \leq 1, 0 \leq t \leq T\\
u(x,y,0,t) &= \emptyset_0(x,y,t), 0 \leq x,y \leq 1, 0 \leq t \leq T\\
u(x,y,1,t) &= \emptyset_1(x,y,t), 0 \leq x,y \leq 1, 0 \leq t \leq T.
\end{aligned}$$

And non local boundary condition

$$\int_0^1\int_0^1\int_0^1 u(x,y,z,t)dxdydz = m(t), (x,y,z) \in \Omega \cup \partial\Omega \tag{13}$$

where f, ψ_0, ψ_1, φ_0, φ_1, ϕ_0, ϕ_1 and m are known functions and γ(t) is to be determined.

3.2 Adomian Decomposition Method

In this section, we outline the steps to obtain a solution to the three dimension problem using Adomian decomposition method. As above, we reformulate the problem in an operator form:

$$L_t(u) = L_{xx}(u) + L_{yy}(u) + L_{zz}(u) \tag{14}$$

where the differential operators $L_t(.) = \frac{\partial}{\partial t}(.)$ and $L_{xx} = \frac{\partial^2}{\partial x^2}, L_{yy} = \frac{\partial^2}{\partial Y^2}, L_{zz} = \frac{\partial^2}{\partial z^2}$ assuming that the inverse L_t^{-1} exists and is defined as:

$$L_t^{-1} = \int_0^t (.)dt. \tag{15}$$

3.3 Solution Procedure

Applying the inverse operator on both the sides of Eq. (14) and using the initial condition yields:

$$u(x,y,z,t) = L_t^{-1}(L_{xx}(u(x,y,z,t) + L_{yy}(u(x,y,z,t) + L_{zz}(u(x,y,z,t)) \text{ Or}$$
$$u(x,y,z,t) = u(x,y,z,0) + L_t^{-1}(L_{xx}(u(x,y,z,t) + L_{yy}(u(x,y,z,t) + L_{zz}(u(x,y,z,t). \tag{16}$$

Now, we decompose the unkown function $u(x,y,z,t)$ as a sum of components defined by the series :

$$u(x,y,z,t) = \sum\nolimits_{k=0}^{\infty} u_k(x,y,z,t) \tag{17}$$

where u_0 is identified as $u(x,\ y,\ z,\ 0)$. Substituting Eq. (17) into Eq. (16) one obtains:

$$\sum_{k=0}^{\infty} u_k(x,y,z,t) = L_t^{-1}\left\{(L_{xx} + L_{yy} + L_{zz})\left(\sum_{k=0}^{\infty} u_k(x,y,z,t)\right)\right\}. \tag{18}$$

Or:

$$u_0(x,y,z,t) = f(x,y,z). \tag{19}$$

And:

$$u_{k+1}(x,y,z,t) = L_t^{-1}\left(L_{xx}(u_k(x,y,z,t)) + L_{yy}(u_k(x,y,z,t) + L_{zz}(u_k(x,y,z,t)))\right), k \geq 0. \tag{20}$$

The components are obtained by the recursive formula:

$$u_0(x,y,z) = f(x,y,z) \tag{21}$$

$$u_{k+1}(x,y,z,t) = L_t^{-1}\left(L_{xx}\left(u_k(x,y,z,t) + L_{yy}(u_k(x,y,z,t) + L_{zz}(u_k(x,y,z,t))\right)\right), k \geq 0. \tag{22}$$

From Eq. (19) and (20) we obtain the first few terms as:

$$\begin{aligned}
u_1(x,y,z,t) &= L_t^{-1}\left(L_{xx}(u_0(x,y,z,t)) + L_{yy}(u_0(x,y,z,t)) + L_{zz}(u_0(x,y,z,t))\right)\\
u_2(x,y,z,t) &= L_t^{-1}\left(L_{xx}(u_1(x,y,z,t)) + L_{yy}(u_1(x,y,z,t)) + L_{zz}(u_1(x,y,z,t))\right)\\
u_3(x,y,z,t) &= L_t^{-1}\left(L_{xx}(u_2(x,y,z,t)) + L_{yy}(u_2(x,y,z,t)) + L_{zz}(u_2(x,y,z,t))\right)
\end{aligned}$$

and so on. As a result, the components u_0, $u_1, u_2, \ldots$are identified and the series solution is thus entirely determined. However, in many cases the exact solution in a closed form may be obtained as we can see in our examples.

4 Examples

4.1 Example 1: Two Dimensional Problem

We consider the two-dimensional diffusion Eq. (1):

$$u_t = u_{xx} + u_{yy}.$$

In which $u = u(x, y, t)$.

The Dirichelet time-dependent boundary conditions on the boundary ∂Ω of the square Ω defined by the line $x = 0,\ y = 0,\ x = 1\ y = 1$ are given by:

$$\begin{aligned}
u(0,y,t) &= e^{(y+2t)} 0 \leq t \leq T\ 0 \leq y \leq 1\\
u(x,0,t) &= e^{(x+2t)} 0 \leq t \leq T\ 0 \leq x \leq 1\\
u(x,1,t) &= e^{(1+x+2t)} 0 \leq t \leq T\ 0 \leq x \leq 1.\\
u(1,y,t) &= e^{(1+y+2t)}
\end{aligned} \tag{23}$$

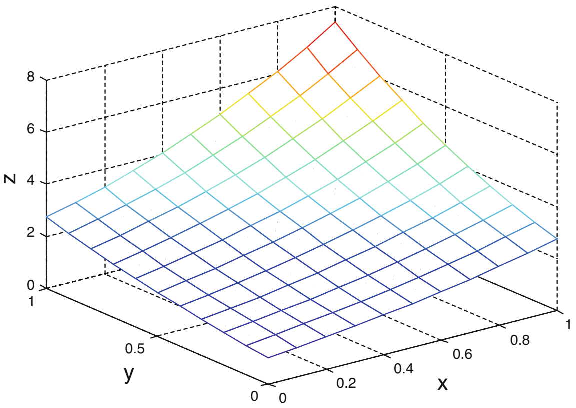


Fig. 1 *Example 1* Variation of the approximate solution for different values of x and y when t = 0.004

And non local boundary condition:

$$\int_0^1 \int_0^1 u(x, y, t)dxdy = (e - 1)^2 e^{2t} \tag{24}$$

with the initial conditions

$$u(x, y, 0) = e^{(x+y)}. \tag{25}$$

Theoretical solution is given by :

$$u(x, y, t) = e^{(x+y+2t)}. \tag{26}$$

Using the Adomians method, described above, Eq. (9) gives the first component

$$u_0(x, y, t) = f(x, y) = e^{(x+y)}. \tag{27}$$

And Eq. (10) gives the following components of the series:

$$u_1 = L_t^{-1}\left(L_{xx}(u_0) + L_{yy}(u_0)\right) = 2\int_0^t e^{x+y}dt = 2te^{x+y} \tag{28}$$

$$u_2 = L_t^{-1}\left(L_{xx}(u_1) + L_{yy}(u_1)\right) = 4\int_0^t te^{x+y}dt = 2t^2e^{x+y} \tag{29}$$

$$u_3 = L_t^{-1}\left(L_{xx}(u_2) + L_{yy}(u_2)\right) = 4\int_0^t t^2e^{x+y}dt = \frac{4}{3}t^3e^{x+y} \tag{30}$$

$$u_4 = L_t^{-1}\left(L_{xx}(u_3) + L_{yy}(u_3)\right) = \frac{8}{3}\int_0^t t^3e^{x+y}dt = \frac{2}{3}t^4e^{x+y} \tag{31}$$

$$u_5 = L_t^{-1}\big(L_{xx}(u_4) + L_{yy}(u_4)\big) = \frac{4}{3}\int\limits_0^t t^4 e^{x+y} dt = \frac{4}{3\times 5} t^5 e^{x+y} \tag{32}$$

$$u_6 = L_t^{-1}\big(L_{xx}(u_5) + L_{yy}(u_5)\big) = \frac{8}{3\times 5}\int\limits_0^t t^5 e^{x+y} dt = \frac{8}{3\times 5\times 6} t^6 e^{x+y} \tag{33}$$

$$u_7 = L_t^{-1}\big(L_{xx}(u_6) + L_{yy}(u_6)\big) = \frac{16}{3\times 5\times 6}\int\limits_0^t t^6 e^{x+y} dt = \frac{16}{3\times 5\times 6\times 7} t^7 e^{x+y}. \tag{34}$$

Substituting (27)–(34) into Eq. (7), we obtain the solution u(x,y,t) of (1) with (23), and (25) in series form as:

$$u(x,y,t) = e^{x+y}\left(1 + \frac{2}{1!}t + \frac{4}{2!}t^2 + \frac{4\times 2}{3!}t^3 + \frac{2\times 2\times 4}{4!}t^4 + \frac{2\times 4\times 4}{5!}t^5 + \frac{2\times 4\times 8}{6!}t^6 + \frac{2\times 4\times 16}{7!}t^7\right). \tag{35}$$

Which can be rewritten as :

$$u(x,y,t) = e^{x+y}\left(1 + \frac{2t}{1!} + \frac{2^2t^2}{2!} + \frac{2^3t^3}{3!} + \frac{2^4t^4}{4!} + \frac{2^5t^5}{5!} + \frac{2^6t^6}{6!} + \frac{2^7t^7}{7!}\right). \tag{36}$$

It can be easily observed that (36) is equivalent to the exact solution:

$$u(x,y,t) = e^{(x+y)}e^{2t} = e^{(x+y+2t)}. \tag{37}$$

Fig. 1 shows the plot of the solution.

4.2 Example 2: Two Dimensional Problem

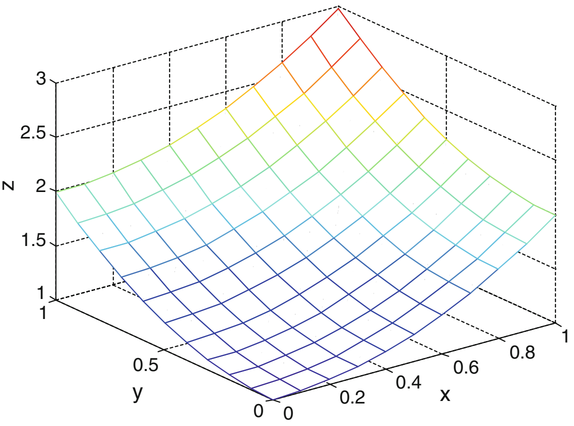
Consider the two-dimensional nonhomogeneous diffusion problem:

$$u_t = u_{xx} + u_{yy} - e^{-t}\big(x^2 + y^2 + 4\big), (x,y) \in \Omega, t > 0. \tag{38}$$

With Initial condition

$$u(x,y,0) = 1 + x^2 + y^2. \tag{39}$$

Fig. 2 *Example 2* Variation of the approximate solution for different values of x and y when t = 0.004



And the boundary conditions:

$$\begin{aligned} u(0,y,t) &= 1 + y^2 e^{-t} 0 \le t \le 1, 0 \le y \le 1 \\ u(1,y,t) &= 1 + (1+y^2)e^{-t} 0 \le t \le 1, 0 \le y \le 1 \\ u(x,0,t) &= 1 + x^2 e^{-t}, 0 \le t \le 1, 0 \le x \le 1 \\ u(x,1,t) &= 1 + (1+x^2)e^{-t}, 0 \le t \le 1, 0 \le x \le 1. \end{aligned} \tag{40}$$

And the non local boundary condition

$$\int_0^1 \int_0^1 u(x,y,t)dxdy = 1 + (2/3)e^{-t}. \tag{41}$$

The exact solution is:

$$u(x,y,t) = 1 + e^{-t}\left(x^2 + y^2\right). \tag{42}$$

Writing the problem in operator form and applying the inverse operator one obtains;

$$\begin{aligned} L_t^{-1}(L_t(u(x,y,t))) = L_t^{-1}(L_{xx}(u(x,y,t))) + L_t^{-1}\left(L_{yy}(u(x,y,t))\right) \\ + L_t^{-1}\left(-e^{-t}\left(x^2+y^2+4\right)\right) \end{aligned} \tag{43}$$

$$L_t^{-1}(L_t(u(x,y,0))) = u(x,y,0).$$

From which we obtain:

$$\begin{aligned} u(x,y,t) = u(x,y,0) + L_t^{-1}(L_{xx}u(x,y,t)) + L_t^{-1}\left(L_{yy}u(x,y,t)\right) \\ + L_t^{-1}\left(-e^{-t}\left(x^2+y^2+4\right)\right). \end{aligned} \tag{44}$$

Using Adomian decomposition, the zeroth component is given by:

$$u_0(x,y,t) = u(x,y,0) + L_t^{-1}\left(-e^{-t}\left(x^2+y^2+4\right)\right) \tag{45}$$

and

$$u_{k+1}(x,t) = L_t^{-1}\left(L_{xx}(u_k(x,y,t)) + L_{yy}(u_k(x,y,t))\right), k \geq 0. \tag{46}$$

Applying these formula, we obtain the components of the series as:

$$\begin{aligned} u_0(x,y,t) &= 1 + x^2 + y^2 - \left(x^2+y^2+4\right)\int_0^t e^{-t}dt \\ u_0(x,y,t) &= -3 + (4 + x^2 + y^2)e^{-t} \end{aligned} \tag{47}$$

$$\begin{aligned} u_1(x,y,t) &= L_t^{-1}\left(\left(L_{xx}+L_{yy}\right)(u_0(x,y,t))\right) = \int_0^t 4e^{-t}dt \\ u_1(x,y,t) &= -4e^{-t} + 4 \end{aligned} \tag{48}$$

$$\begin{aligned} u_2(x,y,t) &= L_t^{-1}\left(\left(L_{xx}+L_{yy}\right)(u_1(x,y,t))\right) = \int_0^t 0dt = 0 \\ u_k(x,y,t) &= 0, k \geq 2. \end{aligned} \tag{49}$$

Once the components are determined then, the series solution completely determined as follows:

$$\begin{aligned} u(x,y,t) &= u_0(x,y,t) + u_1(x,y,t) + \sum_{k=2}^{\infty} u_k(x,y,t) \\ u(x,y,t) &= 1 + (x^2+y^2)e^{-t}. \end{aligned} \tag{50}$$

This solution coincides with the exact one. Fig. 2 shows the plot of the solution.

4.3 Example 3: Three Dimensional Problem

We consider the three-dimensional diffusion equation:

$$u_t = u_{xx} + u_{yy} + u_{zz}. \tag{51}$$

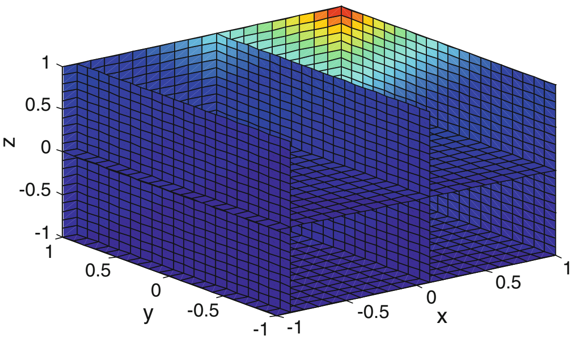


Fig. 3 *Example 3* Variation of the approximate solution for different values of x,y and z when t = 0.004

Table 1 Example 3, $h_x = h_y = h_z = 0.1, h_t = 0.004$

x_i	u_{ex}	u_{ad}	$\lvert u_{ex} - u_{ad} \rvert$
0.0	1.21	1.21	0.0
0.1	1.3662	1.3662	0.0
0.2	1.8441	1.8441	0.0
0.3	2.4893	2.4893	0.0
0.4	3.3602	3.3602	0.0
0.5	4.5358	4.5358	0.0
0.6	6.1227	6.1227	0.0
0.7	8.2648	8.2648	0.0
0.8	11.156	11.156	0.0
0.9	15.059	15.059	0.0
1.0	20.328	20.328	0.0

In which $u = u(x, y, z, t)$. The Dirichelet time-dependent boundary conditions on the boundary $\partial\Omega$ of the cube Ω defined by the lines

$$x = 0, y = 0, z = 0, x = 1, y = 1, z = 1.$$

Are given by:

$$\begin{aligned}
u(0, y, z, t) &= e^{y+z+3t}, 0 \le y, z \le 1, 0 \le t \le T \\
u(1, y, z, t) &= e^{1+y+z+3t}, 0 \le y, z \le 1, 0 \le t \le T \\
u(x, 0, z, t) &= e^{x+z+3t}, 0 \le x, z \le 1, 0 \le t \le T \\
u(x, 1, z, t) &= e^{1+x+z+3t}, 0 \le x, z \le 1, 0 \le t \le T \\
u(x, y, 0, t) &= e^{x+y+3t}, 0 \le x, y \le 1, 0 \le t \le T \\
u(x, y, 1, t) &= e^{1+x+y+3t}, 0 \le x, y \le 1, 0 \le t \le T.
\end{aligned} \tag{52}$$

And non local boundary condition

$$\int_0^1\int_0^1\int_0^1 u(x,y,z,t)dxdydz = (e-1)^3 e^{3t} \tag{53}$$

with the initial condition:

$$u(x,y,z,0) = e^{x+y+z}. \tag{54}$$

Analytic solution is given by:

$$u(x,y,z,t) = e^{x+y+z+3t}. \tag{55}$$

Using the decomposition method, described above, Eq. (21) gives the first component

$$u_0(x,y,z,t) = f(x,y,z) = e^{x+y+z}. \tag{56}$$

And Eq. (22) gives the following components of the series:

$$\begin{aligned} u_1(x,y,z,t) &= L_t^{-1}\left(L_{xx}(u_0(x,y,z,t)) + L_{yy}(u_0(x,y,z,t) + L_{zz}(u_0(x,y,z,t))\right) \\ &= \int_0^t 3e^{x+y+z}dt = 3te^{x+y+z} \end{aligned} \tag{57}$$

$$\begin{aligned} u_2 &= L_t^{-1}(L_{xx}(u_1(x,y,z,t)) + L_{yy}(u_1(x,y,z,t) + L_{zz}(u_1(x,y,z,t))) = \int_0^t 9te^{x+y+z}dt \\ &= ((3t)^2/2!)e^{x+y+z} \end{aligned} \tag{58}$$

$$\begin{aligned} u_3 &= L_t^{-1}(L_{xx}(u_2(x,y,z,t)) + L_{yy}(u_2(x,y,z,t)) + L_{zz}(u_2(x,y,z,t))) \\ &= \int_0^t (27/2)t^2 e^{x+y+z}dt = ((3t)^3/3!)e^{x+y+z}. \end{aligned} \tag{59}$$

…

And so on.

Then the solution in the series form is given by:

$$u(x,y,z,t) = \sum_{k=0}^{\infty} u_k(x,y,z,t). \tag{60}$$

With the above results:

$$u(x, y, z, t) = e^{x+y+z}\Big(1 + 3t/1! + (3t)^2/2! + (3t)^3/3! + \cdots\Big). \tag{61}$$

Which can be rewritten as:

$$u(x, y, z, t) = e^{x+y+z+3t}. \tag{62}$$

It can be easily observed that (62) is equivalent to the exact solution. Fig. 3 shows the plot of the solution and Table 1 gives the obtained values.

4.4 Example 4: Three Dimensional Problem

Consider the three-dimensional non homogeneous diffusion problem:

$$u_t = u_{xx} + u_{yy} + u_{zz} - e^{-t}\big(x^2 + y^2 + z^2 + 4\big), 0 < x, y, z < 1, t > 0 \tag{63}$$

with the initial condition

$$u(x, y, z, 0) = 1 + x^2 + y^2 + z^2. \tag{64}$$

And the boundary conditions

$$\begin{aligned}
&u(0, y, z, t) = 3 + \big(y^2 + z^2 - 2\big)e^{-t}, 0 \le y, z \le 1,\\
&0 \le t \le T\\
&u(1, y, z, t) = 3 + \big(-1 + y^2 + z^2\big)e^{-t}, 0 \le y, z \le 1, 0 \le t \le T\\
&u(x, 0, z, t) = 3 + \big(x^2 + z^2 - 2\big)e^{-t}, 0 \le x, z \le 1\ , 0 \le t \le T
\end{aligned}$$

$$u(x, 1, z, t) = 3 + \big(-1 + x^2 + z^2\big)e^{-t}, \quad 0 \le x, z \le 1, 0 \le t \le T, \tag{65}$$

$$\begin{aligned}
&u(x, y, 0, t) = 3 + \big(x^2 + y^2 - 2\big)e^{-t}, 0 \le x, y \le 1, \quad 0 \le t \le T\\
&u(x, y, 1, t) = 3 + \big(x^2 + y^2 - 1\big)e^{-t}, 0 \le x, y \le 1, \quad 0 \le t \le T.
\end{aligned}$$

And the non local boundary condition

$$\int_0^1\int_0^1\int_0^1 u(x, y, z, t)dxdydz = 3 - e^{-t}, 0 \le t \le T. \tag{66}$$

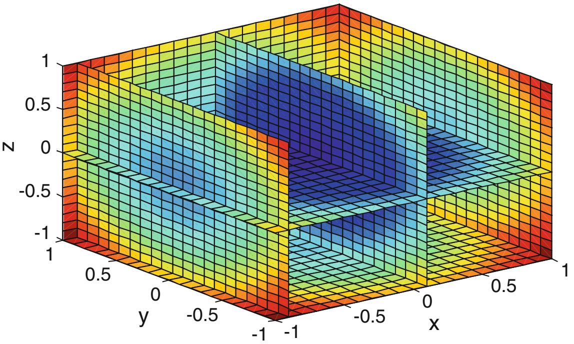


Fig. 4 *Example 4* Variation of the approximate solution for different values of x, y and z when t = 0.004

Theoretical solution is given by:

$$u(x,y,z,t) = 3 + \left(x^2 + y^2 + z^2 - 2\right)e^{-t}. \tag{67}$$

Writing the problem in operator form and applying the inverse operator one obtains:

$$\begin{aligned} L_t^{-1}(L_t(u(x,y,z,t))) = L_t^{-1}&(L_{xx}(u(x,y,z,t) + L_{yy}(u(x,y,z,t) + L_{zz}(u(x,y,z,t))) \\ &+ L_t^{-1}\left(-e^{-t}\left(x^2 + y^2 + z^2 + 4\right)\right) \end{aligned} \tag{68}$$

$$L_t^{-1}(L_t(u(x,y,z,0))) = u(x,y,z,0). \tag{69}$$

From which we obtain :

$$\begin{aligned} u(x,y,z,t) = u\left(x,y,z,0\right) &+ L_t^{-1}\left(L_{xx}(u(x,y,z,t)) + L_{yy}(u(x,y,z,t))\right. \\ &\left.+ L_{zz}(u(x,y,z,t)) + L_t^{-1}\left(-e^{-t}\left(x^2 + y^2 + z^2 + 4\right)\right)\right). \end{aligned} \tag{70}$$

Using Adomian decomposition, the zeroth component is given by:

$$u_0(x,y,z,t) = u(x,y,z,0) + L_t^{-1}\left(-e^{-t}\left(x^2 + y^2 + z^2 + +4\right)\right) \tag{71}$$

and

$$u_{k+1}(x,y,z,t) = L_t^{-t}\left(L_{xx}(u_k(x,y,z,t)) + L_{yy}(u_k(x,y,z,t)) + L_{zz}(u_k(x,y,z,t))\right) \tag{72}$$

Table 2 Example 4, $h_x = h_y = h_z = 0.1, h_t = 0.004$

x_i, y_j, z_k	u_{ex}	u_{ad}	$\lvert u_{ex} - u_{ad}\rvert$
0.0	1.0080	0.98403	0.02397
0.1	1.0976	1.0737	0.0239
0.2	1.3665	1.3426	0.0239
0.3	1.8148	1.7908	0.024
0.4	2.4422	2.4183	0.0239
0.5	3.249	3.225	0.024
0.6	4.235	4.2111	0.0239
0.7	5.4004	5.3764	0.024
0.8	6.7450	6.721	0.024
0.9	8.2689	8.2429	0.024
1.0	9.9721	9.9481	0.024

Applying these formulas, we obtain the components of the series solution as:

$$\begin{aligned} u_0(x,y,z,t) &= 1 + \mathrm{x}^2 + y^2 + z^2 + \int_0^t -e^{-t}\left(x^2 + y^2 + z^2 + 4\right)dt \\ &= -3 + (x^2 + y^2 + z^2 + 4)e^{-t} \end{aligned} \tag{73}$$

$$\begin{aligned} u_1(x,y,z,t) &= L_t^{-1}\left(L_{xx}(u_0(x,y,z,t)) + L_{yy}(u_0(x,y,z,t)) + L_{zz}(u_0(x,y,z,t))\right) \\ &= \int_0^t 6e^{-t}dt = 6 - 6e^{-t} \end{aligned} \tag{74}$$

$$\begin{aligned} u_2(x,y,z,t) &= L_t^{-1}\left(L_{xx}(u_1(x,y,z,t)) + L_{yy}(u_1(x,y,z,t)) + L_{zz}(u_1(x,y,z,t))\right) \\ &= \int_0^t 0\,dt = 0. \end{aligned} \tag{75}$$

Then:

$$u_k(x,y,z,t) = 0, k \geq 2. \tag{76}$$

Finally, we obtain the approximate solution:

$$u(x,y,z,t) = u_0(x,y,z,t) + u_1(x,y,z,t)$$

$$u(x,y,z,t) = -3 + \left(x^2 + y^2 + z^2 + 4\right)e^{-t} - 6e^{-t} \tag{77}$$

or:

$$u(x,y,z,t) = 3 + \left(x^2 + y^2 + z^2 - 2\right)e^{-t}. \tag{78}$$

And we can observe that the obtained result is exact. Fig. 4 shows the plot of the solution and Table 2 gives the obtained values.

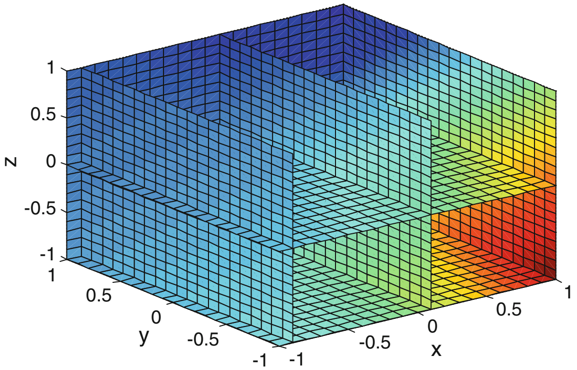


Fig. 5 *Example 5* Variation of the approximate solution for different values of x, y and z when t = 0.004

4.5 Example 5: Three Dimensional Problem

Consider the problem

$$u_t = u_{xx} + u_{yy} + u_{zz}, 0 < x, y, z < 1, t > 0. \tag{79}$$

Subject to the initial condition

$$u(x, y, z, 0) = (1 - y - z)e^x, 0 \le x, y, z \le 1. \tag{80}$$

And the boundary conditions

$$\begin{aligned} u(0, y, z, t) &= (1 - y - z)e^t, 0 \le y, z \le 1, 0 \le t \le 1 \\ u(1, y, z, t) &= (1 - y - z)e^{1+t}, 0 \le y, z \le 1, 0 \le t \le 1 \\ u(x, 0, z, t) &= (1 - z)e^{x+t}, 0 \le x, z \le 1, 0 \le t \le 1 \\ u(x, 1, z, t) &= -ze^{x+t}, 0 \le x, z \le 1, 0 \le t \le 1 \\ u(x, y, 0, t) &= (1 - y)e^{x+t}, 0 \le x, y \le 1, 0 \le t \le 1 \\ u(x, y, 1, t) &= -ye^{x+t}, 0 \le x, y \le 1, 0 \le t \le 1. \end{aligned} \tag{81}$$

And the local boundary condition

$$\int_0^1 \int_0^1 \int_0^{x(1-x)} u(x, y, z, t)dxdydz = 7.5(1 - e)e^t. \tag{82}$$

Consider the Eq. (79) in an operator form

$$L_t(u(x, y, z, t)) = L_{xx}(u(x, y, z, t)) + L_{yy}(u(x, y, z, t)) + L_{zz}(u(x, y, z, t)) \tag{83}$$

where, $L_t, L_{xx}, L_{yy}, L_{zz}, L_t^{-1}$ are defined as above. Assume that the inverse operator L_t^{-1} exists operating with L_t^{-1} on both sides of Eq. (83) we obtain

$$u(x,y,z,t) = L_t^{-1}(L_{xx}(u(x,y,z,t)) + L_{yy}(u(x,y,z,t)) + L_{zz}(u(x,y,z,t)). \tag{84}$$

Using the decomposition method, the zeroth component is given by

$$u_0(x,y,z,t) = u(x,y,z,0) \tag{85}$$

and

$$u_{k+1}(x,y,z,t) = L_t^{-1}(L_{xx}(u_k(x,y,z,t)) + L_{yy}(u_k(x,y,z,t)) + L_{zz}(u_k(x,y,z,t))). \tag{86}$$

Applying these formulas, we have

$$\begin{aligned}
u_0(x,y,z,t) &= (1-y-z)e^x \\
u_1(x,y,z,t) &= L_t^{-1}\big(L_{xx}(u_0(x,y,z,t)) + L_{yy}(u_0(x,y,z,t)) + L_{zz}(u_0(x,y,z,t))\big) \\
u_1(x,y,z,t) &= L_t^{-1}\big(L_{xx}(u_0(x,y,z,t)) + L_{yy}(u_0(x,y,z,t)) + L_{zz}(u_0(x,y,z,t))\big) \\
&= \int_0^t (1-y-z)e^x dt = (1-y-z)te^x \\
u_2(x,y,z,t) &= L_t^{-1}\big(L_{xx}(u_1(x,y,z,t)) + L_{yy}(u_1(x,y,z,t)) + L_{zz}(u_1(x,y,z,t))\big) \\
&= \int_0^t (1-y-z)e^x tdt = (1-y-z)e^x t^2/2! \\
u_3(x,y,z,t) &= L_t^{-1}(L_{xx}(u_2(x,y,z,t) + L_{yy}(u_2(x,y,z,t) + L_{zz}(u_2(x,y,z,t)) \\
u_3(x,y,z,t) &= (1-y-z)t^3/3! \\
&\dots \\
u_k(x,y,z,t) &= L_t^{-1}\big(L_{xx}(u_{k-1}(x,y,z,t)) + L_{yy}(u_{k-1}(x,y,z,t)) + L_{zz}(u_{k-1}(x,y,z,t))\big) \\
&= \int_0^t (1-y-z)e^x(t)^{k-1}/(k-1)!dt = (1-y-z)e^x t^k/k!
\end{aligned} \tag{87}$$

And so on, once the components are determined then, the series solution is given by:

$$u(x,y,z,t) = \sum_{k=0}^{\infty} u_k(x,y,z,t) = (1-y-z)e^x\left(\sum_{k=0}^{\infty} t^k/k!\right) \tag{88}$$

or equivalently:

$$u(x,y,z,t) = (1-y-z)e^{x+t}. \tag{89}$$

This result is in good agreement with the exact one. Fig. 5 shows the plot of the solution and Table 3 gives the obtained values.

Table 3 Example 5, $h_x = h_y = h_z = 0.1, h_t = 0.004$

x_i, y_j, z_k	u_{ex}	u_{ad}	$\lvert u_{ex} - u_{ad} \rvert$
0.0	1.004	1.004	0.0
0.1	0.88768	0.88767	0.1
0.2	0.73578	0.73577	0.1
0.3	0.54211	0.54210	0.1
0.4	0.29956	0.29956	0.0
0.5	0.0	0.0	0.0
0.6	0.36588	0.36588	0.0
0.7	0.80873	0.80872	0.1
0.8	1.3407	1.3407	0.0
0.9	1.9756	1.9756	0.0
1.0	2.7292	2.7292	0.0

5 Conclusion

In this work, we have detailed the study of the Adomian decomposition method ADM and used it for finding the solution of the multi-dimensional diffusion equation. We consider two cases: a two dimensional equation with non local boundary conditions and a three dimensional equation with an integral condition This method is employed without using linearization, discretization, transformation, or restrictive assumptions. It is very much compatible with the diversified and versatile nature of physical problems, the results obtained are all in good agreement with the exact solutions of the problems under study. Moreover this method is efficient, reliable, accurate, easier to implement as compared to the traditional techniques.

References

1. A. Cheniguel, Numerical method for solving wave equation with non local boundary conditions, in *Proceeding of the International Multi-Conference and Computer Scientists 2013*, vol. II, pp. 1190–1193, IMECS 2013, , Hong Kong, 13–15 March 2013
2. A. Cheniguel, On the numerical solution of three-dimensional diffusion equation with an integral condition, in *Lecture Notes in Engineering and Computer Science 2013*, WCECS 2013, pp. 1017–1021, San Francisco, 23–25 Oct 2013
3. A. Cheniguel, Numerical simulation of two-dimensional diffusion equation with non local boundary conditions. Int. Math. Forum **7**(50) 2457–2463 (2012)
4. A. Cheniguel, Numerical method for solving heat equation with derivative boundary conditions, in *Proceedings of the World Congress on Engineering and Computer Science 2011*, vol. II, pp. 983–985, WCECS 2011, San Francisco, 19–21 Oct 2011
5. A. Cheniguel, A. Ayadi, Solving heat equation by the adomian decomposition method, in *Proceeding of the World Congress on Engineering 2011*, vol. I, pp. 288–290, WCE 2011, London, 6–8 July 2011

6. A. Cheniguel, A. Ayadi, Solving non homogeneous heat equation by the adomian decomposition method. Int. J. Numer. Methods Appl. **4**(2), 89–97 (2010)
7. A. Cheniguel, Numerical method for non local problem, Int. Math. Forum **6**(14),659–666 (2011)
8. M. Siddique, Numerical computation of two-dimensional diffusion equation with non local boundary conditions, IAENG Int. J. Appl. Math. **40**(1) IJAM_401_04, 91–99 (2010)
9. M. Akram, A parallel algorithm for the heat equation with derivative boundary conditions. Int. Math. forum **2**(12) 565–574 (2007)
10. A. Akram, M.A. Pasha, Numerical method for the heat equation with a non local boundary condition. Int. J. Inf. Syst. Sci. **1**(2), 162–171 (2005)
11. A.B. Gumel, W.T. Ang, F.H. Twizell, Efficient parallel algorithm for the two-dimensional dffusion equation subject to specification of mass. Inter. J. Comput. Math. **64,** 153–163 (1997)
12. G. Adomian, *Solving frontier problems of physics : the decomposition method* (Kluver Academic Publishers, Dordrecht, 1994)
13. B.J. Noye, K.J. Hayman, Explicit two level finite difference methods for two-dimensional diffusion equation. Inter. J. Comput. Math. **42** 223–236 (1992)
14. G. Adomian, R. Rach, Noise terms in decomposition solution series. Comput. Math. Appl. **24**(11), 61–64 (1992)
15. G. Ekolin, Finite difference methods for a non local boundary value problem for the heat equation. BIT **31**, 245–261 (1991)
16. G. Adomian, A review of the decomposition method in applied mathematics. J. Math. Anal. Appl. **135**, 501–544 (1988)

Chapter 44
Parameter Identification for Population Equations Modeling Erythropoiesis

Doris H. Fuertinger and F. Kappel

Abstract Physiological models explaining the anemia of chronic kidney disease have become more complicated over the last years. Identification of model parameters poses difficulties as measurements are very limited. A model for erythropoiesis, consisting of coupled partial differential equations, is adapted to individual patients. The numerical approximations make use of evolution operators and are based on the theory of abstract Cauchy problems. The abstract Cauchy problems corresponding to the model equations are approximated by Cauchy problems on finite-dimensional subspaces of the state space of the original problem. A low approximation dimension suffices to obtain accurate numerical solutions and estimates for the parameters. An example of (locally) well identifiable parameters expressing numerical convergence for increasing dimensions of the finite dimensional approximating system is discussed. Moreover, it is demonstrated that a clever choice of cost-functionals can reduce the observation time needed for parameter identification from 150 to 90 days.

Keywords Abstract cauchy problems · Anemia · Chronic kidney disease · Erythropoiesis · Parameter estimation · Structured population equations · Weighted cost functionals

D. H. Fuertinger (✉)
Renal Research Institute, 315 East 62nd Street, New York, NY 10065, USA
e-mail: Doris.Fuertinger@rriny.com

F. Kappel
Institute for Mathematics and Scientific Computing, University of Graz, Heinrichstrasse 36, 8010 Graz, Austria
e-mail: franz.kappel@uni-graz.at

H. K. Kim et al. (eds.), *Transactions on Engineering Technologies*,
DOI: 10.1007/978-94-017-9115-1_44,

1 Introduction

Anemia affects almost all patients suffering from chronic kidney disease (CKD) and is mainly caused by a failure of renal excretory and endocrine function which results in an insufficient production of erythropoietin (EPO) in CKD patients. Already partial correction of anemia in CKD patients by administration of erythropoiesis stimulating agents (ESA) reduces cardiac-related morbidity and mortality, which is the most common cause of death among these patients (see for instance [3, 12, 22]). ESAs stimulate the bone marrow to produce red blood cells, exerting a similar effect than the endogenous hormone erythropoietin (EPO).

Erythropoiesis—the production of new red blood cells—is a very complex process. Stem cells in the bone marrow commit to the erythroid lineage and start to develop into red blood cells (RBC), so called erythrocytes. This process takes about 2 weeks. During this time the cells divide, differentiate and some of them eventually die. The development from stem cells into erythrocytes involves a number of different cell stages which exhibit diverse characteristic patterns with regard to the rate of proliferation, the rate of apoptosis, the maturation velocity and the need for EPO and other substances (for details see e.g. [16]). Therefore a model for erythropoiesis has to include submodels for various cell populations. A rather comprehensive model for erythropoiesis was developed in [10] (see also [11]). The model presented in [10] consists of five structured population equations with cell age being the structuring attribute, two ordinary differential equations describing the development of endogenous and administered EPO over time plus a number of auxiliary equations describing the influence of EPO on maturation and mortality rates and the control of EPO secretion by the kidneys based on the oxygen carrying capacity of blood which is proportional to the erythrocyte population.

In comparison to healthy persons, CKD patients have a very high inter-individual variability in red blood cell lifespan, bone marrow response to EPO, endogenous EPO production and half-life of the administered EPO compound. Routine measurements of these quantities are not practicable in a clinical environment or are simply impossible. Therefore prediction of the individual response to EPO administration schemes is extremely difficult and is further aggravated by the fact that there is a long delay in reaction of the RBC population to EPO levels. This results in severe limitations of ESA treatment regimens so that hemoglobin levels in CKD patients tend to fluctuate widely and cycling phenomena are frequently observed [4, 7].

In the next sections we describe numerical schemes for solving the erythropoiesis model and for conducting parameter identification. The construction makes use of the theory of evolution operators and is based on approximation of population densities by Legendre polynomials. A standard nonlinear least-squares formulation for the cost functional is used and minimized using a simplex method, thus avoiding to compute derivatives with respect to parameters for the relatively complicated PDE-ODE model. The numerical approximation scheme presented in Sect. 2 is used for parameter identification. We show that a rather low approximation dimension

N suffices to get accurate estimates for the parameter values. The results indicate that the "optimal" parameters of the approximations converge as $N \to \infty$. In addition, we demonstrate that an appropriate choice of cost-functionals allows to reduce the observation time from 150 to 90 days.

2 Numerical Approximation

The core of the model developed in [10] consists of a coupled system of five structured population equations which are of the following type:

$$\begin{aligned} \frac{\partial}{\partial t}u(t,x) + v(E(t))\frac{\partial}{\partial x}u(t,x) &= (\beta - \alpha(E(t),x))u(t,x), \quad t \geq 0,\ x \in \Omega, \\ u(0,x) &= \phi(x), \quad x \in \Omega, \\ v(E(t))u(t,x_{\min}) &= f(t), \quad t \geq 0, \end{aligned} \tag{1}$$

where $u(t,\ x)$ is the population density at time $t \geq 0$ and cell age $x \in \Omega = [x_{\max}, x_{\min}]$. Furthermore, $v(E(t))$ denotes the maturation velocity of cells according to the erythropoietin level $E(t)$ at time t, β is the proliferation rate, $\alpha(E(t),x)$ the rate of apoptosis depending on the erythropoietin level $E(t)$ at time t and the cell age x. The initial population density is denoted by $\phi(x)$ and $f(t)$ is the influx of cells from the precedent population class at time t. A function $u(t,\ x)$ is a solution of (1) if it is obtained by the method of characteristics.

Our approach to solve the model equations numerically is based on semigroup theory, respectively on the theory of abstract Cauchy problems (see e.g. [5, 6, 13, 20]). The abstract Cauchy problems corresponding to our model equations are approximated by Cauchy problems on finite-dimensional subspaces of the state space of the original Cauchy problem. Although the natural state space is $L^1(\Omega, \mathbb{R})$ (because the L^1-norm of positive density functions is the corresponding total population), in this paper we concentrate on finding solutions to problem (1) in the Hilbert space $X = L^2(\Omega, \mathbb{R})$. The idea is to formulate problem (1) as an abstract Cauchy problem in the state space $X = L^2(\Omega, \mathbb{R})$,

$$\begin{aligned} \dot{y}(t) &= \mathscr{A}(t)y(t) + \tilde{f}(t), \qquad t \geq 0, \\ y(0) &= \phi, \end{aligned} \tag{2}$$

where the operator $\mathscr{A}(t)$ is given by $(\kappa(t,x) = \beta - \alpha(E(t),x))$,

$$\begin{aligned} \operatorname{dom}\mathscr{A}(t) = \{\phi \in X | \phi \text{ absolutely continuous on } \Omega, \\ \phi(x_{\min}) = 0,\ v(E(t))\phi' - \kappa(t,\cdot)\phi \in X\}, \\ \mathscr{A}(t)\phi = -v(E(t))\phi' + \kappa(t,\cdot)\phi, \quad \phi \in \operatorname{dom}\mathscr{A}(t). \end{aligned}$$

The term $\tilde{f}(t)$ represents the boundary condition in (1) and is given by $\delta_0 f(t)$, where δ_0 is the delta impulse defined by $\langle \delta_0, \phi \rangle_X = \phi(0)$ for continuous ϕ. For a mild solution $y(t)$ of (2) the function $u(t,x) = y(t)(x)$ is a solution of (1).

Since the range for the attribute is different for the different cell populations, it is convenient to transform the abstract Cauchy problems for each population equation to an abstract Cauchy problem on the weighted L^2-space $X_\rho = L^2_\rho(0,1;\mathbb{R})$ with the constant weight $\rho = x_{\max} - x_{\min}$. Hence, we normalize the attribute x and consider all population classes on the unit interval [0, 1]. The transformation is given by $\phi \to \phi \circ h, \phi \in X$, where $h(\tau) = x_{\min} + \rho\tau, 0 \le \tau \le 1$. The weight ρ is introduced in order to make X and X_ρ isometric under this transformation, i.e., $\|\phi\|_X = \|\phi \circ h\|_{X_\rho}$. The solutions of the transformed equation on the state space X_ρ are given by $\tilde{y}(t) = y(t) \circ h$, where $y(t)$ is a mild solution of Eq. (2).

In order to obtain approximations to the solutions of (1) we choose finite dimensional subspaces $X^N = \operatorname{span}(e_0, \dots, e_N) \subset X_\rho, N = 1,2,\dots,$ where $e_j(\tau) = \rho^{-1/2} L_j(-1+2\tau), 0 \le \tau \le 1, j = 1,2,\dots$. Here, L_j denotes the j-th Legendre polynomial. For details on Legendre polynomials see, for instance [1]. From orthogonality of the sequence of Legendre polynomial in $L^2(-1,1;\mathbb{R})$ we immediately get that e_k, $k = 0,1,2,\dots,$ is an orthogonal sequence in X_ρ.

Following the approach taken in [15] (see also [8]) we define the approximating operators $\mathscr{A}^N(t)$ on X^N as $\mathscr{A}^N(t)\phi = \Pi^N \mathscr{A}_\rho(t)\phi, \phi \in X^N$, where Π^N is the orthogonal projection $X_\rho \to X^N$ and $\mathscr{A}_\rho(t)$ is the transformed evolution operator on X_ρ. $\mathscr{A}_\rho(t)\phi$ for $\phi \in X^N$ is obtained by formally applying $\mathscr{A}_\rho(t)$ to ϕ. In general, a $\phi \in X^N$ is not in $\operatorname{dom}\mathscr{A}_\rho(t)$. It certainly has the necessary smoothness properties, but $\phi(0) \neq 0$ in general. Formally applying $\mathscr{A}_\rho(t)$ to ϕ means $\mathscr{A}_\rho(t)\phi = -\rho^{-1} v(E(t))\phi' + (\kappa \circ h)\phi + \delta_0 \phi(0)$ (see [10, Sect. 4.1]). Then $\mathscr{A}^N(t)$ is given by

$$\mathscr{A}^N(t)\phi = -\rho^{-1} v(E(t)) \Pi^N \phi' + \Pi^N(\kappa \circ h)\phi + \delta_0^N \phi(0), \qquad \phi \in X^N.$$

The 'approximating' delta impulse $\delta_0^N \in X^N$ is defined as

$$\langle \delta^N, \phi \rangle_{X_\rho} = \phi(0), \quad \text{for } \phi \in X^N. \tag{3}$$

The approximation $\tilde{y}^N(t), N = 0,1,\dots,$ for $\tilde{y}(t)$ are obtained as the solutions of the following Cauchy problems on the finite dimensional subspaces X^N:

$$\begin{aligned} \frac{d}{dt}\tilde{y}^N(t) &= \mathscr{A}^N(t)\tilde{y}^N(t) + \delta_0^N f(t), \\ \tilde{y}^N(0) &= \Pi^N(\phi \circ h). \end{aligned}$$

The approximations $u^N(t,x)$ for $u(t,x)$ are given by

$$u^N(t,x) = \big(\tilde{y}^N \circ H\big)(x), \quad t \geq 0,\ x_{\min} \leq x \leq x_{\max},$$

where $H(x) = (x - x_{\min})/\rho$ is the inverse mapping of h.

2.1 Coordinate Representation of the Approximating Systems

Finally, one has to compute matrix representations of the operators $\mathscr{A}^N(t)$, coordinate vectors of the approximating delta impulse δ_0^N and coordinate vectors of the projections $\Pi^N(\phi \circ h)$. We introduce the "basis matrix"

$$E^N = (e_0, \ldots, e_N), \quad N = 1, 2, \ldots,$$

where e_k, $k = 0, 1, \ldots$, are the weighted Legendre polynomials defined previously.

Let $z^N(t)$ be the coordinate vector of $u^N(t)$ with respect to the basis E^N, i.e., $u^N(t) = E^N z^N(t)$. Then $z^N(t)$ solves

$$\begin{aligned} \dot{z}^N(t) &= A^N(t) z^N(t) + d^N f(t), \\ z^N(0) &= \alpha^N(\tilde{\phi}), \end{aligned} \tag{4}$$

where A^N is the matrix representation of $\mathscr{A}^N(t)$ with respect to E^N and d^N is the coordinate vector of the delta impulse δ_0^N. We denote with $\alpha^N(\tilde{\phi})$ the coordinate vector of $\Pi^N \tilde{\phi}$ with respect to the basis E^N, i.e.,

$$\Pi^N \tilde{\phi} = E^N \alpha^N(\tilde{\phi}), \quad \tilde{\phi} \in X_\rho.$$

Rearranging the preceding equation and use of standard computations yields

$$\begin{aligned} \alpha^N(\tilde{\phi}) &= \langle E^N, E^N \rangle_{X_\rho}^{-1} \Big\langle E^N, \tilde{\phi} \Big\rangle_{X_\rho} \\ &= \Big(\Big\langle e_0, \tilde{\phi} \Big\rangle_{X_\rho}, 3\Big\langle e_1, \tilde{\phi} \Big\rangle_{X_\rho},\ \ldots, (2N+1)\Big\langle e_N, \tilde{\phi} \Big\rangle_{X_\rho} \Big)^T. \end{aligned} \tag{5}$$

From (3) and $\delta_0^N = E^N d^N$ we get

$$d^N = \rho^{-1/2}\big(1, -3,\ \ldots, (-1)^N(2N+1)\big)^T. \tag{6}$$

In order to compute the columns of the matrix representation $A^N(t)$ of $\mathscr{A}^N(t)$ with respect to the basis E^N we have to compute the coordinate vectors of $\mathscr{A}^N(t)e_k$,

$$\mathscr{A}^N(t)e_k = -\rho^{-1}v(E(t))\Pi^N e_k' + \Pi^N((\kappa \circ h)e_k) - \delta_0^N e_k(0), \quad k = 0, \ldots, N.$$

We obtain

$$A^N(t) = -D^N + B^N - F^N,$$

where the columns of the matrices $D^N, B^N, F^N \in \mathbb{R}^{(N+1)\times(N+1)}$ are the coordinate vectors of $\rho^{-1}v(E(\overline{t}))\Pi^N e_k', \Pi^N((\kappa \circ h)e_k)$ and $\delta_0^N e_k(0), k = 0, \ldots, N$. After some lengthy but standard computations using also properties of Legendre polynomials (see [15]) we get

$$D^N = 2\rho^{-1}v(E(t))\mathrm{diag}(1, 3, \ldots, 2N+1) \times \mathrm{II}^N,$$

where

$$\mathrm{II}^N = \begin{pmatrix} 0 & 2 & 0 & 2 & \cdots & \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & \ddots & 2 \\ \vdots & & & \ddots & \ddots & 0 \\ \vdots & & & & \ddots & 2 \\ 0 & \cdots & \cdots & \cdots & \cdots & 0 \end{pmatrix}.$$

The matrix $B^N = (B^N_{j,k})_{j,k=0,\ldots,N}$ is given by

$$B^N_{j,k} = \frac{2k+1}{2}\int_{-1}^{1} \kappa((x_{\min} + x_{\max} + \rho\tau)/2)L_j(\tau)L_k(\tau)\,d\tau, \quad j,k = 0, \ldots, N.$$

In case $\kappa(x) \equiv \kappa_0$ on Ω this simplifies to $B^N = \kappa_0 I_{N+1}$, where I_{N+1} denotes the $(N+1) \times (N+1)$ identity matrix.

Using (6) and $e_k(0) = \rho^{-12}(-1)^k, k = 0, \ldots, N$, we get immediately

$$F^N = \rho^{-1/2}(d^N, -d^N, \ldots (-1)^N d^N) \in \mathbb{R}^{(N+1)\times(N+1)}.$$

Finally the coordinate vector of $\delta_0^N f(t)$ is given by $d^N f(t)$ where d^N is given by Eq. (6).

2.2 Approximation of the Total Population

The total population corresponding to a density $u(t,x) \in L^2(\Omega,\mathbb{R}) \cap L^1(\Omega,\mathbb{R})$ is given by

$$P(t) = \int_{x_{\min}}^{x_{\max}} u(t,x)\,dx, \quad t \geq 0. \tag{7}$$

The approximated total population is

$$P^N(t) = \int_{x_{\min}}^{x_{\max}} u^N(t,s)\,ds = \sum_{k=0}^{N} z_k^N(t) \int_{x_{\min}}^{x_{\max}} e_k(H(s))\,ds,$$

where $z^N(t)$ is the solution of (4). Using also orthogonality of the Legendre polynomials and $L_0(\tau) \equiv 1$ we get

$$P^N(t) = \frac{\rho^{1/2}}{2} \sum_{k=0}^{N} z_k^N(t) \int_{-1}^{1} L_k(\tau)\,d\tau = \rho^{1/2} z_0^N(t). \tag{8}$$

2.3 General Comments

A very positive feature of the scheme presented in this section is that we get already a very precise approximation of the solutions and in particular for the total population already for rather low dimensions of the approximating subspaces. In many cases $N = 10$ is sufficient, i.e., we have to solve 11 ODEs which amounts to relatively low computational costs. This plays a critical role when one does parameter identification, a process where, in general, the model has to be solved several hundred times. In this context it is important to notice that our approach gives immediately an approximation for the total population as defined in (7) simply by (8) which just involves the first component of the approximating solution.

Another advantage of approximation schemes of the type presented here is that the finite dimensional approximations (when appropriately designed) preserve qualitative properties of the dynamics of the original system. This plays an increasingly important role when one is dealing with parameter identification. We expect that the parameters we identify for the approximating system are convergent (as $N \to \infty$) to the optimal parameters of the original system.

3 Parameter Identification

We are using the least-squares formulation for the parameter identification, i.e., in order to get the parameter estimates we minimize the cost functional

$$J(p) = \sum_{j=1}^{N} w(t_j)\big(y_j - g\big(t_j, p\big)\big)^2, \tag{9}$$

where p is the parameter vector to be estimated. Furthermore, w is a weighting function, y_j are post-dialytic measurements of the hemoglobin concentration and $g(t_j, p)$ is the corresponding 'model output'. The sampling times $t_j \in T := [0, t_{\max}]$ are the days at which the dialysis treatment takes place (usually on Monday, Wednesday and Friday every week).

3.1 Different Cost Functionals

The weighting function $w : T \to \mathbb{R}^+, T = [0, t_{\max}]$, can be any positive function suitable for the specific problem. The weight $w(t_j)$ put on the j-th term in the cost functional reflects the importance we assign to the measurement y_j. The choice of the weighting function usually is based on information we have on the parameters in the system and/or on information on the quality of the measurements.

We are interested in identifying parameters for individual patients that play an important role for erythropoiesis and have high inter-individual variability. In the context of chronic kidney disease one has to consider that patients might not be "stable" over a longer period of time, i.e., parameters in those patients are likely to change over time (reasons for this might be inflammation, gastro-intestinal bleeding, iron-deficiency etc.). There exist methods that are able to deal with situations were parameters vary over time, i.e. $p(t) \neq \text{const.}, t \in T$. As an example we mention the unscented Kalman filtering (see [2] and the literature quoted there). In this paper we focus on the less complex situation where the parameters do not change over time. Hence, we assume that the parameters are constant over time and thus, the parameters that we identify over the given time period T may present "average" values of the parameter function.

In [9] we used a time period of 150 days to identify parameters. A reduction of this time frame would be desirable as in a clinical setting it might be difficult to gather frequent data over such a long time. If we shorten the time interval on which data are taken for patient adapted parameter estimation and still want to get similar results, it can be advantageous to consider weighting of the data. The idea is to put more weight on more recent data points than on older ones, because if the patient's parameters have changed, we would like to "catch" the current status, i.e., the "true" parameter value close to $t = t_{\max}$.

3.2 Data

The data y_j is obtained using a Crit-Line Monitor. The Crit-Line device provides readings of the hemoglobin concentration and the oxygen saturation during hemodialysis. It is a non-invasive method based on an optical sensor technique. Note, that the function $g(t, p)$, which represents the hemoglobin concentration predicted by the model, is not a state variable of the model but an observable output. We need to express $g(t, p)$ in terms of the state variables of the system such that it corresponds indeed to the hemoglobin concentration in the patients blood. Our output model is (see [9])

$$g(t,p) = \frac{M(t,p) \cdot MCH}{TBV} \cdot c,$$

where $M(t,p)$ is the total RBC population, MCH is the mean corpuscular hemoglobin in pg (which we take to be 29 pg), TBV is the post-dialytic blood volume in ml (estimated by an empirical formula, see [17]) and $c = 10^6$ is a factor which is needed in order to get the hemoglobin concentration in g/dl.

3.3 Method

We use an implementation of the Nelder-Mead algorithm (see [18, 19]), to find minima of the cost functional (9). We choose a direct search method, because it only uses function values and requires no derivative information. The fact, that we do not need to provide derivatives is of distinct advantage, in particular in case of the complicated PDE-ODE model describing erythropoiesis.

The numerical approximation scheme described in Sect. 2 was implemented in Python, see [21]. We use the Nelder-Mead simplex method implemented in the SciPy package for Python (see [14]). The model was adapted to individual CKD patients by adjusting the following parameters: the total blood volume, the number of stem cells committing to the erythroid lineage per day, the bone marrow response to EPO, the half-life of the administered EPO compound in the patient, the RBC lifespan and the endogenous EPO level. Some of those parameters were estimated using empirical formulae. The remaining parameters were determined as described in this section using the weighting function w_{const} with $t_{\max} = 150$. For the investigation presented in Sect. 4 these parameters are taken as the 'optimal' respectively nominal parameters.

4 Results

For the sake of simplicity and to avoid selecting a subset of parameters which is difficult to estimate, we focus in this paper on identifying two parameters simultaneously, using the Hgb measurements from the Crit-Line device: the RBC lifespan p_1 and the endogenous EPO level p_2.

4.1 Numerical Convergence of the Estimation Scheme

For the patient whose data are shown in Fig. 1 the optimal parameter vector is $p_{\text{opt}} = (p_1, p_2) = (63.05,\ 21.08)$. We consider two different starting vectors for the parameter estimation, $p_{\text{init},1} = (56,\ 18.75)$ and $p_{\text{init},2} = (70,\ 23.125)$, which are obtained by perturbing the coordinates of p_{opt} by ±10 %. Furthermore, we conduct the parameter identification runs for $N = 8, 10, 14, 20, 30, 40, 50, 60$. In Table 1 the results of the parameter identification runs are shown. It can be observed that the estimated parameters for $p_{\text{init},1}$ and $p_{\text{init},2}$ are close to each other for all N. The fact that a perturbation of the optimal parameter values of ±10 % provides the same estimates when minimizing the cost-functional $J(p_1, p_2)$ implies that the parameter vector chosen is well identifiable. Moreover, from Table 1 one can see that, the parameter estimates for increasing N show a convergent behavior. Whereas for $N = 8$ the estimates for p_1 and p_2 differ distinctly from p_{opt}, for $N = 10$ the estimated parameter vector is already very close. Increasing N results only in minor changes of the estimated parameter values. In Fig. 1 we compare the model outputs obtained for $N = 10$ (solid line) and $N = 60$ (dashed line) with the corresponding estimated parameter values given in Table 1. The difference between the two simulations is barely recognizable.

4.2 Parameter Identification on Reduced Sampling Intervals

In this section we want to investigate, whether it is possible to identify the parameters p_1 (RBC lifespan) and p_2 (endogenous EPO level) when shortening the observational time and still retain a similar quality of the parameter identification. We reduce the observational time $t_{\max} = 150$ days to 120, 90 and 60 days. All parameter identifications are started by perturbing the nominal parameter values (determined using $t_{\max} = 150$ days) by +10 %. Further, for each time interval considered, we explore if using weighting functions improves the results. In the following all computations are done with $N = 14$.

We use four different types of weighting functions $w(t)$ in the cost functional $J(p)$ (see Eq. (9)): a constant function, a step function, a sigmoid function and an exponential function.

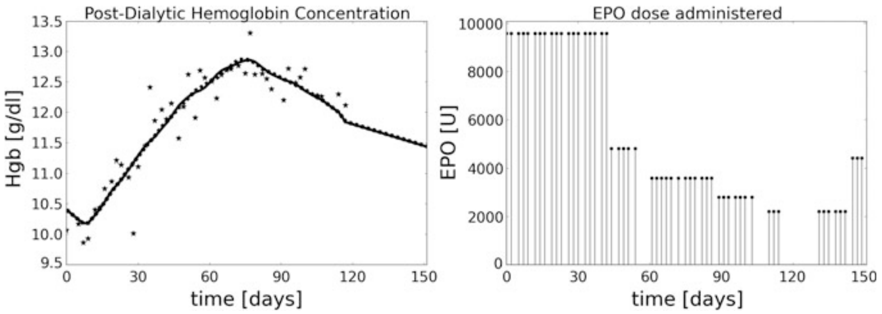


Fig. 1 Model output for N = 10 (*solid line*) and N = 60 (*dashed line*). Stars represent measured data. The *right panel* shows the administered EPO dose

Table 1 Parameter estimates obtained using the starting parameter vectors $p_{\mathrm{init},1}$ and $p_{\mathrm{init},2}$

N	$p_{\mathrm{init},1}$		$p_{\mathrm{init},2}$	
	p_1	p_2	p_1	p_2
28	71.3589	15.0967	71.3599	15.0956
10	63.1753	20.9977	63.1755	20.9969
14	63.0866	21.0543	63.0867	21.0544
20	63.0237	21.1054	63.0238	21.1054
30	63.0314	21.0979	63.0313	21.0979
40	63.0263	21.0995	63.0263	21.0994
50	63.0112	21.1092	63.0112	21.1092
60	63.0532	21.0817	63.0531	21.0817

$$w_{\mathrm{const}}(t) \equiv 1, \quad t \in T, \tag{10}$$

$$w_{\mathrm{step}}(t) = \begin{cases} 1 & \text{for}\, 0 \le t < t_{\max}/2, \\ w_{\max} & \text{for } t_{\max}/2 \le t_{\max}. \end{cases} \tag{11}$$

$$w_{\mathrm{sigm}}(t) = \frac{w_{\max} - 1}{1 +^{-a(t-c)}} + 1, \quad t \in T, \tag{12}$$

$$w_{\exp}(t) = e^{bt}, \quad t \in T. \tag{13}$$

The weighting function w_{const} corresponds to the situation where all measurements have the same weight. For the weighting functions w_{step}, w_{sigm} and $w_{\exp}$ we consider a 'minor' and a 'moderate' version. Let w denote one of these weighting functions. Then we always have $w(0) = 1$. For the minor versions we have $w(t_{\max}) = 3$ and for the moderate versions $w(t_{\max}) = 6$. For w_{step} and w_{sigm} this means that for the minor versions we have $w_{\max} = 3$ and for the moderate version $w_{\max} = 6$. For w_{sigm} we set $c = t_{\max}/2$ which implies that the point of inflection of w_{sigm} is at $t = t_{\max}/2$. The constant a is chosen such that $w_{\mathrm{sigm}}(0) \approx 1$ and

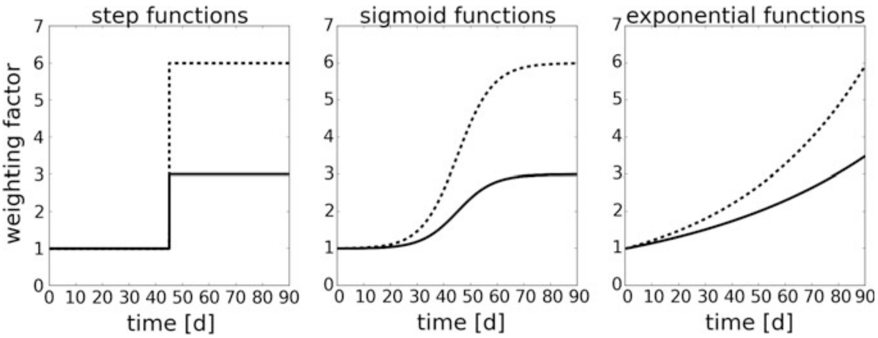


Fig. 2 Minor (*solid line*) and moderate (*dashed line*) weight functions used

$w_{\mathrm{sigm}}(t_{\max}) \approx w_{\max}$. For instance, in case $t_{\max} = 90$ we set $a = 0.15$ which gives $w_{\mathrm{sigm}}(0) = 1.0058$ and $w_{\mathrm{sigm}}(t_{\max}) = 5.994$. In case of $w_{\exp}$ we have $b = \ln 3/t_{\max}$ for the minor version and $b = \ln 6/t_{\max}$ for the moderate version. The different weighting functions in case of $t_{\max} = 90$ days are depicted in Fig. 2. We can see that w_{step} and w_{sigm} put more emphasis on the information content of the data in the second half of the observation period. Using $w_{\exp}$ as weighting function implies that one is especially interested in the very recent history.

Using data from 36 patients we do parameter estimation for $t_{\max} = 60,\ 90,\ 120$ and 150 days and different weighting functions. For comparison of the results we use the mean absolute error (MAE) between model output for different estimates and the data. MAEs, standard deviations, minimal and maximal errors were always computed on the interval [0, 150].

4.3 Comparison of Different Observational Times and Weighting Functions

For $t_{\max} = 120$ the results are close to what we observe for an observational time of 150 days. The mean absolute errors are similar for the different cost functionals. Nevertheless, constant or minor weighting functions give the best results. In Table 2 the MAE over 150 days are listed for 10 patients. The approaches depicted are cost functionals using constant weights and a minor step, sigmoid and exponential weight function. The mean and the standard deviation of the errors over the 36 patients for $t_{\max} = 120$ are comparable to the results for $t_{\max} = 150$. Moreover, the difference of the parameter values identified, is nominal for cost functionals using constant or minor weighting.

An observational time of 60 days, even though working well for some patients, seems to be not applicable, as the performance of none of the approaches tried, is reliable and consistent throughout patients. In Fig. 3 we give two examples for the parameter identification using $t_{\max} = 60$ days compared to $t_{\max} = 150$ days.

Table 2 MAE for $t_{\max} = 150$ d and $t_{\max} = 120$ d with constant and minor versions of the non-constant weighting functions

pat.	150 d	120 d			
	w_{const}	w_{const}	w_{step}	w_{sigm}	w_{exp}
001	0.2483	0.2582	0.2835	0.2847	0.2867
002	0.2466	0.2501	0.2572	0.2552	0.2793
003	0.4471	0.4203	0.426	0.4312	0.4329
004	0.2779	0.2767	0.2838	0.2787	0.276
005	0.3717	0.3677	0.4087	0.3835	0.3827
006	0.4047	0.3943	0.3948	0.3968	0.3935
007	0.3341	0.333	0.333	0.3321	0.3321
008	0.4589	0.4609	0.4745	0.464	0.4577
009	0.3254	0.3071	0.3008	0.3012	0.3055
010	0.3322	0.374	0.3675	0.3606	0.3519

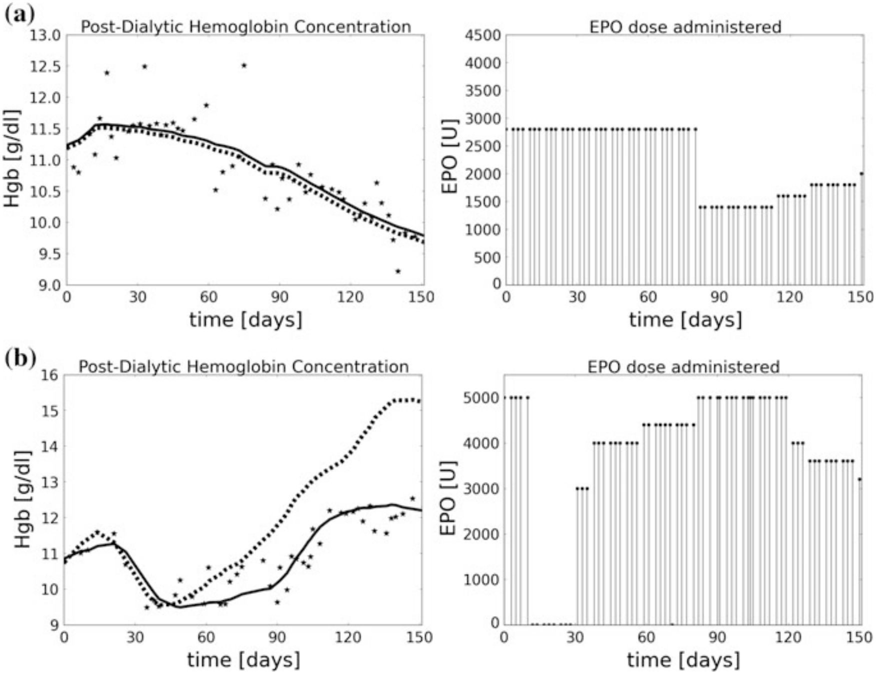


Fig. 3 Model adaption for an observation time of 60 days compared to 150 days. **a** Estimates for $t_{\max} = 60$ (*dashed line*): 69.02 d, 19.82 U/l; $t_{\max} = 150$ (*solid line*): 69.15 d, 20.33 U/l, **b** Estimates for $t_{\max} = 60$ (*dashed line*): 94.53 d, 6.51 U/l; $t_{\max} = 150$ (*solid line*): 65.38 d, 10.63 U/l

Subplot (a) shows an example where $t_{\max} = 150$ and $t_{\max} = 60$ give similarly good results. The weighting function used in case $t_{\max} = 60$ days is the moderate exponential function. Subplot (b) depicts the results for a situation where the

Table 3 Results for 36 patient using criterion (14) in case of $t_{\max} = 90$ d and w_{const} in case of $t_{\max} = 150$ d

$t_{\max}$	150 d	90 d
Average MAE	0.3621	0.3757
Standard deviation	0.0788	0.0805
Median	0.3751	0.3754
Minimum error	0.2077	0.2106
Maximum error	0.5114	0.6114

information contained in the interval [0, 60] is not sufficient for identifying the parameters properly. Again the moderate exponential weighting function was used.

The case $t_{\max} = 90$ is the most interesting one. The reduction of the observational time by 2 months seems to give acceptable, but more diverse results than those obtained for $t_{\max} = 120$. Different cost functionals show different behaviors. In general, w_{const} works well. Nevertheless, for some patients the weighting functions w_{sigm}, w_{step} and $w_{\exp}$ show improved behavior. This is probably the case whenever the parameters, we try to identify, change over the period of time considered. Each of the proposed weighting functions is superior in some situations, i.e. for some patients. Therefore, we focus on consistency in the performance of the different approaches.

Throughout the 36 patients, we observe that the weighting functions given by Eqs. (10)–(12) work well. Overall, exponential functions show poorer performance. Further, in general, minor weighting functions work better than moderate ones. In order to state a criterion for deciding a priori which weighting function is best to be used for a specific patient we consider the change in post-dialytic hemoglobin measurements. The reasoning behind this is, that large variations in the hemoglobin data might be an indicator that the patient's status and consequently his parameters have changed. However, such fluctuations might also (entirely) be caused by treatment. Our computations showed that in case of $t_{\max} = 90$ d the following criterion gave the best results:

$$w = \begin{cases} w_{\text{const}} & \text{if } \Delta\text{Hgb} \le 2.5\,\text{g/dl}, \\ \text{minor } w_{\text{sigm}} & \text{if } \Delta\text{Hgb} > 2.5\,\text{g/dl}, \end{cases} \tag{14}$$

where $\Delta\text{Hgb} = \max(y(t_j)) - \min(y(t_j)), t_j \in [0, 90]$. The parameter identification with w_{const} gives results very close to the nominal parameter values, whereas identification with minor w_{sigm} performs poorly. The parameter values are far away from the nominal values and the model fit is bad. For patients with $\Delta\text{Hgb} > 2.5\,\text{g/dl}$ we do not observe any similar exceptions. On the contrary, in this situations minor w_{sigm} gives almost always slightly better results than w_{const}. Altogether, applying the criterion (14) we are able to achieve almost the same quality for our model adaptations for $t_{\max} = 90$ as for $t_{\max} = 150$. The criterion $\Delta\text{Hgb} \le 2.5 g/dl$ was met by 15 patients and $\Delta\text{Hgb} > 2.5 g/dl$ by 21 patients. Table 3 shows the average mean absolute error for the 36 patients and the standard deviation. Further, it lists the median and minimum and maximum of the errors for $t \in [0,\ 150]$. The average error, its standard deviation, the median and the minimum error are very close in both approaches. Only the maximum error observed is slightly higher, when $t_{\max}$ is reduced to 90 days.

5 Conclusions

The parameters *RBC lifespan* and *endogenous EPO level* are simultaneously well identifiable using post-dialytic measurements for the hemoglobin concentration on sufficient long observation periods. Our findings show that the originally used period of 150 days can be shortened to 90 days when using appropriate weighting functions in the cost functional for the least-squares parameter estimation according to a criterion which considers the variations in the hemoglobin concentration of the patient. The parameters identified for the approximating systems converge to the parameters of the original system as the approximation dimension $N \to \infty$. In order to get sufficiently accurate estimates for the parameters it is sufficient to choose $N = 10$.

References

1. M. Abramowits, I. Stegun (eds.), *Handbook of Mathematical Functions*, 10th edn. National Bureau of Standards—Applied Mathematics Series (1972)
2. A. Attarian, J. Batzel, B. Matzuka, H.T. Tran, in *Application of the unscented Kalman filtering to parameter estimation*, ed. by J.J. Batzel, M. Bachar, F. Kappel. Mathematical Modeling and Validation in Physiology: Application to the Cardiovascular and Respiratory Systems, Lecture Notes in Mathematics (Mathematical Biosciences Subseries), vol 2064 (Springer, New York, 2013) pp. 75–88
3. A. Besarab, W. Bolton, J. Browne, J. Egrie, A. Nissenson, D. Okamoto, S. Schwab, D. Goodkin, The effects of normal as compared with low hematocrit values in patients with cardiac disease who are receiving hemodialysis and epoetin. N. Engl. J. Med. **339**, 584–590 (1998)
4. A. Collins, R. Brenner, J. Ofman, E. Chi, N. Stuccio-White, M. Krishnan, C. Solid, N. Ofsthun, J. Lazarus, Epoetin alfa use in patients with ESRD: an analysis of recent US prescribing patterns and hemoglobin outcomes. Am. J. Kidney Dis. **46**, 481–488 (2005)
5. E. Davies, *One-Parameter Semigroups*. (Academic Press, London, 1980)
6. K.J. Engel, R. Nagel, *One Parameter-Semigroups for Linear Evolution Equations*. (Springer, Berlin, 2000)
7. S. Fishbane, J. Berns, Hemoglobin cycling in hemodialysis patients treated with recombinant human erythropoietin. Kidney Int. **68**, 1337–1343 (2005)
8. D. Fuertinger, F. Kappel, A numerical method for structured population equations modeling control of erythropoiesis. in *1st IFAC Workshop on the Control of Systems Modeled by Partial Differential Equations (CPDE 2013)*, September 25–27, Paris (France) (2013)
9. D. Fuertinger, F. Kappel, A parameter identification technique for structured population equations modeling erythropoiesis in dialysis patients, Lecture Notes in Engineering and

Computer Science. in *Proceedings of the World Congress on Engineering and Computer Science 2013*, WCECS 2013, 23–25 October, 2013, San Francisco, USA, pp. 940–944

10. D. Fuertinger, F. Kappel, S. Thijssen, N. Levin, P. Kotanko, A model of erythropoiesis in adults with sufficient iron availability. J. Math. Biol. **66**(6), 1209–1240 (2013)
11. D.H. Fuertinger, A model for erythropoiesis. Ph.D. thesis, University of Graz, Austria, 2012
12. A. Go, G. Chertow, D. Fan, C. McCulloch, C. Hsu, Chronic kidney disease and the risks of death, cardiovascular events, and hospitalization. N. Engl. J. Med. **351**, 1296–1305 (2004)
13. K. Ito, F. Kappel, *Evolution Equations and Approximations* (World Scientific, Singapore, 2002)
14. E. Jones, T. Oliphant, P. Peterson et al., SciPy: Open source scientific tools for Python. http://www.scipy.org/ (2001)
15. F. Kappel, K. Zhang, Approximation of linear age-structured population model using Legendre polynomials. J. Math. Anal. Appl. **180**, 518–549 (1993)
16. M. Lichtman, E. Beutler, T. Kipps, U. Seligsohn, K. Kaushansky, J. Prchal (eds.), *Williams Hematology*, 7th edn. (McGraw-Hill, New York, 2005)
17. S. Nadler, J. Hidalgo, T. Bloch, Prediction of blood volume in normal human adults. Surgery **51**, 224–232 (1962)
18. J. Nelder, R. Mead, A simplex method for function minimization. Comput. J. **7**, 308–313 (1965)
19. J. Nocedal, S. Wright, *Numerical Approximation*, 2nd edn. Operations Research and Financial Engineering. (Springer, NewYork, 2006)
20. A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations*. (Springer, Berlin, 1983)
21. G. van Rossum, F.L. Drake (eds.), Python Reference Manual. Python Software Foundation, http://docs.python.org/ref/ref.html (2012)
22. G. Strippoli, J. Craig, C. Manno, F. Schena, Hemoglobin targets for the anemia of chronic kidney disease: a meta-analysis of randomized, controlled trials. JASN **15**, 3154–3165 (2004)

Chapter 45
Design Support System of Fishing Vessel Through Simulation Approach

Stefano Filippi, Piero Giribone, Roberto Revetria, Alessandro Testa, Guido Guizzi and Elpidio Romano

Abstract The objective of this work is to create a module for a ship maneuvering simulator, which will allow the training of the crews of vessels in deep water. In this work we developed a system of "artificial intelligence" to model marine biological entities: this system can simulate the movement of a school of fish in a realistic manner. In addition, we developed an analysis of the main instruments on board and problems relating to the virtual simulation.

Keywords Boids · Echo-sounder · Fish · Modeling · Ship · Simulation · Training

S. Filippi (✉) · P. Giribone · R. Revetria · A. Testa
DIME, Univesity of Genoa, Via dell'Opera Pia 15, Genoa, Italy
e-mail: stefanofil@gmail.com

P. Giribone
e-mail: piero@itim.unige.it

R. Revetria
e-mail: roberto.revetria@unige.it

A. Testa
e-mail: atesta@consorzioabaco.it

G. Guizzi · E. Romano
DICMAPI, Univesity of Naples "Federico II", P.le Tecchio 80, 80125 Naples, Italy
e-mail: g.guizzi@unina.it

E. Romano
e-mail: elromano@unina.it

H. K. Kim et al. (eds.), *Transactions on Engineering Technologies*,
DOI: 10.1007/978-94-017-9115-1_45,

1 Introduction

Fishing and aquaculture supplied about 147 million tons of fish in year 2010. Of these, 128 million were used as food for the world's population, with a mean annual quota of 19 kg per person, which is the highest ever. The fishing sector is a source of income for millions of people all around the globe. The number of employees in fleets and aquaculture increased significantly during the three last decades, with an annual growth rate of 3.6 % since 1980. It is estimated that in year 2008 almost 45 million persons earned their living from work that is directly linked to the sector and that 12 % were women. The number of fishing vessels is estimated at about 4.3 million, a value that has remained constant in the past few years, of which 51 % are motorised; the remaining 47 % are located mainly in Asia (77 %) and Africa (20 %). Currently, fishing depends significantly on fossil fuels, which are essential for the movement of the vessels and the functioning of all auxiliary systems onboard (processing and storage). This strong link is counter-productive for current environmental concerns and the consequent limitations on emissions that all governments are now launching and, in particular, links the cost of fishing to the cost of the fuel. As a consequence, commercial risks increase.

2 The Fishing Extension Project

The objective of this work is to create a new module for the Naval++ manoeuvring simulator, capable of recreating the main operations of commercial fishing on the high seas [1]. To that end, it is necessary to develop an instrument allowing optimal interfacing with the existing architecture (exploiting all the characteristics associated with navigation and the environment), and being to create and manage new elements:

- Fishing boats;
- Fishing gears;
- Additions electronic instrumentation;
- School of fish.

This is the heart of the new system, from its similarity with the real one depends on the correct functioning of the entire module.

The purpose of the simulator is to recreate the most common situations that the crew of a motorised fishing vessel may encounter: a series of possible behaviours of the school of fish is created on the basis of the choices activated by the user of the system [2]. The functioning of the fishing simulator can be illustrated by the following diagram (Fig. 1).

As shown in the figure, the manoeuvring simulator and the fishing extension should become two perfectly complementary units: the instructor should interact with both systems, so as to be able to describe the behaviour of the schools of fish

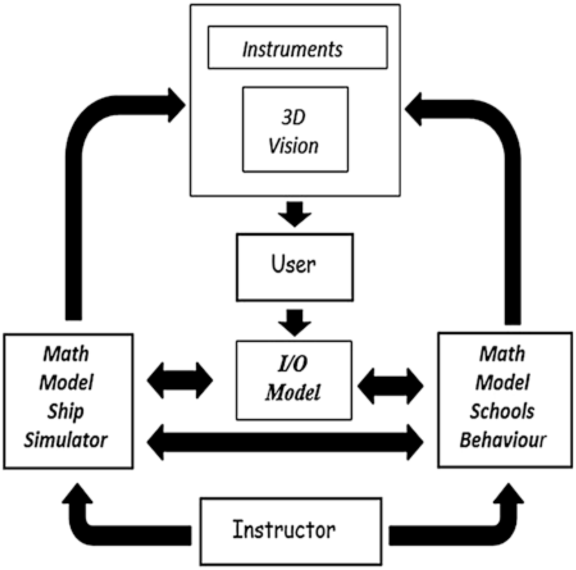


Fig. 1 General naval++ model integrated with fish school simulator

and the environment appropriately. The heart of the new system is in "Mathematical Model School Behaviour", which essentially includes two fields [3–5]:

- Reciprocal behavioural dynamics of the elements within the school;
- Behavioural dynamics of the school in relation to the operations of the user and the surrounding environmental conditions.

This is then used to interact with the mathematical simulation model of the vessels and with the onboard instruments (I/O Model), both navigation and fishing instruments [6].

To achieve a realistic simulation of the processes of fishing is necessary to implement a series of instruments, such as:

- Sonar: it is the main tool, it is used to perceive all the elements that are located below the water surface with a coverage of 360°.
- Echo sounder: It's similar to sonar, but allows you to analyze with greater detail what is under the boat, you can work with great depth, but it has a fixed direction;
- Net sounder: it is also an ultrasound system which has the peculiarity of having a transducer placed on the upper part of the network and it provides information about its location, depth level and elements within the network.

These should be integrated with all the other instruments on board, such as ARPA RADAR, ECDIS, Binocular, etc.

The interaction between the simulator and the user has be made by:

- Control lever for driving electric winches;

- Indicators for the cable length, speed and stress;
- Interface for the control of equipment and alarms.

Finding a school of fish is linked to an acoustic model and the type of machinery. In order to have a good mathematical model, it is necessary to take into account the following characteristics:

- Losses related to the propagation of waves due to absorption of water;
- Refractions due to the variation of the propagation velocity in the water and the consequent presence of shadow areas;
- Losses due to reflection from the seabed and surface area;
- Doppler effect.

Furthermore, the feedback signal comprises:

- Reverberation due to the seabed;
- Reverberation due to the surface;
- Noise due to the boat;
- Noise due to the sea;
- Eco by schools of fish and individual elements;
- Eco due to fishing gear, Echoes from the seabed.

The creation of a code for describing the natural phenomena required quite a lot of precious time: after an initial bibliographic research and analysis phase and literature review, it was necessary to decide on how to develop the project, considering the time-frame and cost of the initiative.

3 The Behavioural Simulation Code

The theory encompassing the fundamental concepts involved in the schematization of the movement of groups is called “Flocking behaviour”: the first scholar of this subject, capable of creating a model resembling a calculator, was Craig Reynolds in 1986, with his “Boids” programme [7].

The model was constructed in an extremely simple way, exploiting three basic types of behaviour:

- Repulsion: the subject moves away from its neighbours, if under a defined distance;
- Alignment: the subject moves with its neighbours, if at a suitable distance;
- Cohesion: the subject approaches the other elements if too distant.

As can be seen, the heart of the system is the relation with neighbours: each element wishes to be as close as possible to the center of the group but since, as happens in nature, the perception of the surrounding environment is limited, the elements are “content” with being in the center of their neighbours.

More complex is the situation of those who are on the edges: these will always have a speed component that leads them to go to the body of the group. Using these three simple rules the element moves in an extremely realistic, creating a complex motion interactions with otherwise very difficult to recreate.

Other scholars have also reworked the algorithm in order to obtain different behaviors, even assuming the search for food or the presence of a predator.

The schools of fish, unlike other groups of animals, haven't a leader, but they move based on a group consciousness: the main aim of each component is to remain in a state of parallelism than their neighbors, keeping the distances as constant as possible; in this way, the group is able to move very fast and also very large distances [8, 9]. For the modeling of the first level of motion were made the following assumptions:

- Each member of the model moves in agreement with the other elements, according to the same laws;
- The movement of the group is independent of external inputs;
- The movement of each fish is influenced only by the immediate neighbors.

This simple algorithm has thus been adapted to our purposes [10]: the first step of the code developed by us was the definition of parameters, including the dimension of the space in which the school of fish moves, the number of elements participating in the simulation, and a whole series of parameters related to the speed of displacement, achievement and generation of the objectives; after all the basic settings are supplied, the code first generates the fish that are to form the various schools, considering them all as belonging to the same class, and then the objectives.

At this point, the "guidance" dynamics of the elements start functioning, respecting the laws of "group motion" [11, 12].

After the defined time has lapsed (or the objective has been achieved) a new goal is created automatically, consisting in recalculation of all the parameters relating to the movement. The cycle continues in this loop until the operator intervenes.

At this point, the "guidance" dynamics of the elements start functioning, respecting the laws of "group motion". After the defined time has lapsed (or the objective has been achieved) a new goal is created automatically, consisting in recalculation of all the parameters relating to the movement. The cycle continues in this loop until the operator intervenes.

3.1 School Generation

Considering the details of functioning, after the programme's launch data have been defined, the code can initiate its own work cycle.

The schools of fish are generated: these have been established in advance in terms of dimension and number; the n distinct elements, the fish, which compose the schools, are created in space and attracted to a common gathering point, a different one for each school of fish. The individual elements are objects that have the same characteristics; the principle characteristics are:

- Current position;
- Previous instant position;
- Current speed;
- Previous instant speed;
- Direction.

These parameters are fundamental for calculating the interactions between individual fish [13].

3.2 Generation of Objectives

The purpose was to create a motion generator allowing the schools of fish to move within all the work area, in such a way as to generate totally random swimming paths, different for each element at each code launch.

The method functions in the following way:

- goals are generated in the work space;
- a single fish moves in the direction of the goal;
- while moving in the direction of the goal, the fish is expected to interact with the other members of the school.

In particular, in order to generate random goals, three equations are used (one for each spatial coordinate, X Y Z) in the form:

$$Goal[0] = (lenght_{x_{axis}}) \cdot \left(\frac{rand}{rand\_max}\right) - (lenght\_\mathrm{x\_axis}/2) \tag{1}$$

The term *rand / ((float) RAND_MAX)* generates a value between 0 and 1 that is multiplied by the length of the space along the x axis and is reduced by half of the dimension. This allows you to get any value in the interval [−Δx /2, +Δx /2], pointing out that by default the origin of the system is placed at the midpoint of the three dimensions.

However, in order to avoid that at each launch of the simulation the generation of the same random numbers is generated, a code line is written that is able to link the time of the system (the time of the computer) to the generation function: In this way, different launch instants correspond to different generations of objectives (during every cycle, the “srand” function provides a new “seed” for the random generation of “rand”).

3.3 Calculation of Group Paths and Dynamics

In order to calculate the position of each element, at each step, in relation to its objective and the other components of the school, a function was created [14]. The first step is to update the variables of the previous cycle, in particular the position and speed; then, to calculate the distance between an element and the objective, using the following simple equation:

$$line\_to\_goal = global_{goal} - oldpos \tag{2}$$

After which the progress of the element is evaluated in relation to the previous instant through:

$$pos = oldpos + oldvel \cdot dt \tag{3}$$

And its speed:

$$vel = oldvel + t\arg etdir \cdot dt \tag{4}$$

With a maximum value, however, beyond which the value is truncated. In the case where the new position differs from the previous one, a new direction vector is created:

$$direction = oldpos - pos \tag{5}$$

Moreover, the inclination angles of the fish are updated during the “swimming” phase (the inclination of the element various at each moment in time, or rather, the cone representing the element, in such a way as to simulate the swimming phase in the best possible way): being a simple simulation, these vary cyclically between two values, +angle and –angle.

Going back to the previous formulae, it is necessary to evaluate parameter “targetdir”, as this parameter represents the real direction of the fish’s movement (and the movement vector): in fact, as the fish is moving in an extremely crowded environment, it cannot move simply in the “line to goal” direction but must change direction so as to follow the rules of the school’s movement or rather that of the Boids.

For each step of the simulation, the ideal path linking each component to the objective is calculated, after which a cycle that calculates the neighbours of each fish commences: these neighbours are the elements that are at a distance of less than a set limit and have a path that crosses the path of the element under examination.

After all the neighbours have been noted, the code controls which ones are under a set distance value (thus entering the influencing area) and calculates an average distancing direction. Thus, an ideal escape direction and an “escape priority” is defined, i.e. the importance of increasing the distance from the group compared to movement towards the objective.

The smaller the distance the higher the “priority” value. Thus, the value obtained is subtracted from the level of maximum priority, generally set at 1. Having noted this value, if priority is still available, the code calculates an approach “priority” for the elements which, however, is less important compared to the previous one, and also calculates an ideal approach direction (both the priority and the direction are based on the orientation of the average speed vector of the neighbours compared to the element under examination).

A similar procedure is followed by evaluating the distance using the centre of gravity of the positions of the neighbours (this guarantees group compactness). Finally, the final direction is calculated as the sum of the results of priorities and ideal directions, considering the residual priority as the priority that pushes towards the objective:

$$\begin{aligned} & t\arg etdir \\ & = (borders \cdot weight) + (\text{min\_dist} \cdot \text{weight}) + (averag\text{e\_velocity} \cdot weight) \\ & + (line\_to\_goal \cdot weight) \end{aligned} \tag{6}$$

4 From Schools to Ellipsoids

The code described has shown optimum and stable functioning: at a qualitative level, it optimally replicates a generic school of fish, even a large one [15] (Fig. 2).

In order to fulfill the requests of the customer, a mathematical structure was implemented that provides the same advantages as regards the emulation of group behaviour, but a lower calculation weight. Therefore, the school ellipsoid envelope was introduced. This is the minimum volume figure that is able to contain all the elements of each school, notably reducing the number of data elements memorized by the computer. In particular, for each n (optional fixed value, generally equal to 10) time steps, the software calculates this geometry that always assumes different dimensions, thus reproducing the dynamics of the school.

4.1 Rotation Ellipsoid

An ellipsoid is a closed quadratic surface, which represents the three-dimensional equivalent of an ellipse. Its initial equation, in Cartesian coordinates and with the axes aligned with the main ones, is as follows:

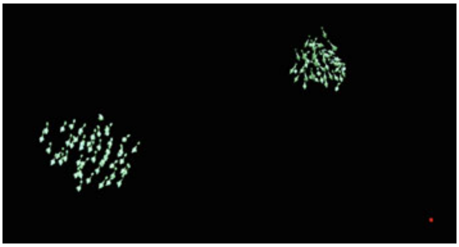


Fig. 2 3D virtual reality GUI for schools behaviour validation

$$(x^2/a^2) + (y^2/b^2) + (z^2/c^2) = 1 \tag{7}$$

This is a typical choice when it is necessary to envelope a cloud of points. Matlab was used to distinguish, at each step, the minimum volume, as it contains a special instrument that designs these objects rapidly. A minimization function was drafted, centred around the "fmincon" command, the functioning of which foresees the setting of an equation to minimise, while varying specific unknown components and under established conditions.

A vector x was defined containing the six unknowns:

- x(1) x(2) x(3) are the three ellipsoid radii
- x(4) x(5) x(6) are the three centre coordinates

Thus, knowing that the volume of the ellipsoid is obtained from the following formula:

$$volume = 4/3 \cdot \pi \cdot a \cdot b \cdot c \tag{8}$$

where a, b, c, are the three radii, this minimization function was chosen, excluding the multiplication constants.

$$f = x(1) \cdot x(2) \cdot x(3) \tag{9}$$

It was thus necessary to set the constraint equations: these were all characterised by non-linear inequality, in quantities equal to the number of elements constituting the school. In order to establish these, reference was made to formulation within the Cartesian area of the ellipsoid:

$$\left\lfloor (x - x_0)^2/a^2 \right\rfloor + \left\lfloor (y - y_0)^2/b^2 \right\rfloor + \left\lfloor (z - z_0)^2/c^2 \right\rfloor \tag{10}$$

After all the positions of the fish were set in matrix Q [*nxm*], with *n* equal to the number of fish, their generic form was written:

$$C = \left\lfloor (Q_{i1} - x(4))^2/x(1)^2 \right\rfloor + \left\lfloor (Q_{i2} - x(5))^2/x(2)^2 \right\rfloor + \left\lfloor (Q_{i3} - x(6))^2/x(3)^2 \right\rfloor \tag{11}$$

As the spatial distribution of fish varies, in order to have an idea of the school's compactness, a part of code was inserted in order to define the density of the ellipsoid's population. This is calculated as follows:

$$Ro = n/volume \tag{12}$$

To conclude, the calculation process, compared to an input of a matrix QQ [nxm], where n = number of fish, restores 7 values, six for the construction of the ellipsoid and the seventh for the definition of density.

5 On-Board Instruments: Echo Sounder

The objective was to create a dynamic image able to replicate, as faithfully as possible, the images on the monitor of a real instrument for the purpose of analysing the seabed.

The point of departure of the work was the analysis of the onboard instruments of a vessel, while focusing on the main characteristics of echo sounders: having analysed the problem in a simplified way, it can be said that echo sounders emit a "cone" of acoustic waves that are reflected after reaching a body with a higher density than water; a suitable microphone collects such sounds and a microprocessor calculates the time between the emission and the return, thus calculating the depth [16, 17].

5.1 Dynamic Representation of the Vessel

In order to represent the area of interest of the sonar in the best possible way, it was decided to schematize everything using a cone; the required parameters are as follows: basic radius, height and vertex angle [18–20].

The code foresees generation via the introduction of a "grid" created with the points of area a at defined steps of the rotation height. This, as can be noted by reading the code is parameterised in such a way that its origin is at point B, i.e. the vector containing the position coordinates of the vessel, and its height is equal to the depth [21–23].

5.2 Study of the Sonar Cone/Schools of Fish Interference Problem

This is the heart of this part of the code, as it includes writing the necessary mathematical laws that would allow a correct intersection between the zone of

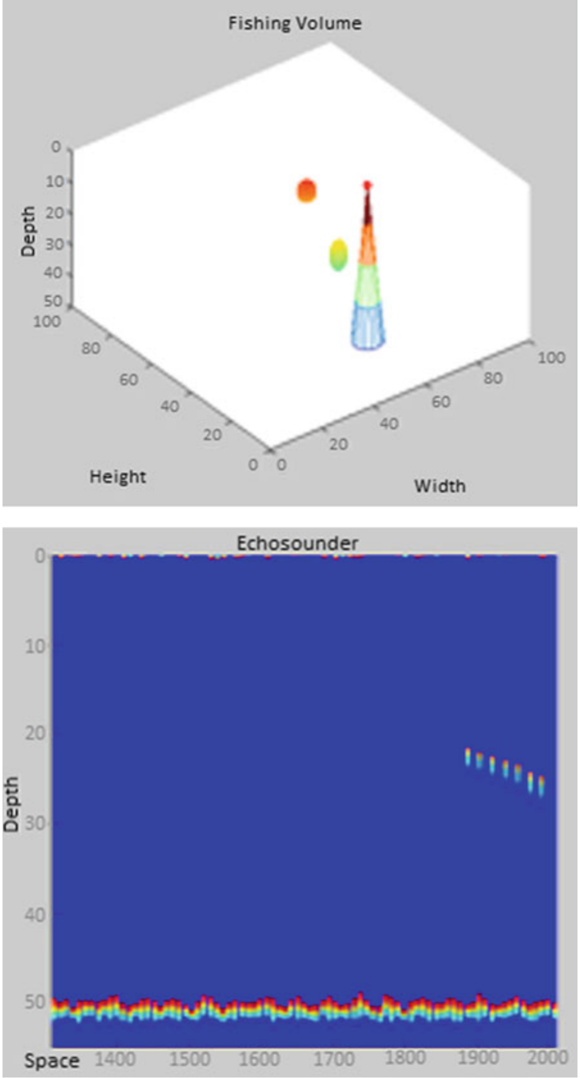


Fig. 3 Echo sounder and echo cone GUI for validation purpose

interest of the sonar and the moving school of fish. Its correct structuring has required countless hours of work, including the development of various approach methods [24–26]. It is necessary to recapitulate the information about the system in order to understand the complexity of the problem as much as possible:

- Area;
- Position of the vessel and geometrical characteristics of the cone, including the cone's mesh value;
- Ellipsoid construction parameters, centre and mesh values.

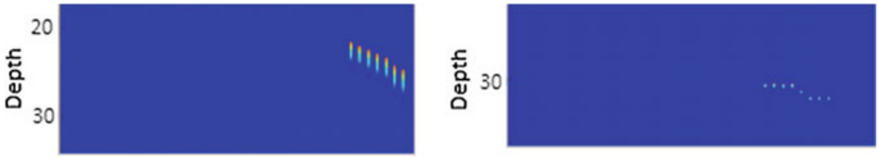


Fig. 4 Echo sounder and echo GUI for validation purpose (details)

Avoiding an excessively heavy code was a major problem as the system studied has to function in real time and thus with very brief reaction times. In order to obtain such results, at each step, the system considers that there is interference in cases where the distance between the axis of the cone and the centre of the ellipsoid is smaller than the radius of the cone at that height, and larger than one of the three radii of the ellipsoid. After the mathematical laws regarding the intersection between the ellipsoids and the cone were written, we tried to reconstruct an echo sounder return signal graphically.

The work was carried out in two steps:

- Schematic visualisation of the signal;
- Graphic visualisation of the signal.

The first was fundamental for starting to understand the problem and assess the potential solutions; for each step, the cycle assessed whether there were any intersections between the sonar and the cone: in the affirmative case a dot (blue) was reproduced, corresponding to the depth of the centre of the ellipsoid, inside a “run area” and “depth” graphic, while in the negative case a red point appeared, corresponding to the seabed [27, 28].

The second step consisted in providing a graphic that would correspond as far as possible to real echo sounders (introducing bands of various colours) and, moreover, a logic was introduced that, based on the density of the fish school, varied the dimensions of the image showing the position of the ellipsoid.

The result obtained showed optimum functioning and a degree of similarity closely resembling the real situation, as shown in the figure below (Fig. 3):

It is interesting to note how the density parameter influences the representation of the school of fish in the simulated echo sounder (Fig. 4):

The two images show a comparison between how a signal produced by the intersection of the sonar cone with a very dense school of fish is described (visible on the left) and, therefore, represented by a small ellipsoid, and a low density school (visible on the right), represented by a large ellipsoid, with the number of school components being equal.

6 Conclusion

The codes created in this activity have shown optimum functioning characteristics. Among these, the capacity to work in real time is of fundamental importance. During the entire programming phase efforts were made to maintain the simplest possible model, by eliminating or minimizing large calculation cycles: the final result has shown that it is possible to make the various components interact in a very quick and efficient way, without running into unpleasant problems or requiring high capacity calculators. After the Fishing and Fish Finder software systems were developed, it was noticed that their use would not only be valid within Naval++, but also as an autonomous system: it was observed that the school of fish, after being subjected to suitable biological validations, could be used to analyze the impact of the vessels on marine fauna. Therefore, this could allow its use not only in simulation but also in design, thus creating new and interesting scenarios for the future. The current state of the work evidences the achievement of an excellent result as regards the interaction of the school's elements, but it is not yet able to react to external stimuli that could compromise integrity and safety. Nevertheless, development has already been pursued with the aim of developing these concerns: in fact, movement is assigned to a system for which the school is brought to move in a certain direction, currently set in a random way but, if external elements were to be introduced, able to compromise the safety of the school, it would suffice to develop a logic for the creation of objectives that would return the school to safety conditions. To conclude, it can be seen how the bases used for the creation of this simulation module were created: a very flexible model was written, capable of emulating the motion of general biological marine entities, but already predisposed for the specific definition, and the issues relevant to the simulation of the instruments on board, which are fundamental for reproducing the right fishing conditions, were studied in depth.

References

1. M.V. Abrahams, P.W. Colgan, Risk of predation, hydrodynamic efficiency and their influence on school structure, Environ. Biol. Fishes **13**, 195–202 (1985)
2. A. Alvarez, Z. Ye, Effects of fish school structures on acoustic scattering. CES J. Mar. Sci. **56**(3), 361–369 (1999)
3. A.G. Bruzzone, P. Giribone, R. Revetria, Operative Requirements and Advances for the New Generation Simulators. in *Multimodal Container Terminals, Winter Simulation Conference Proceedings*, 2, pp. 1243–1252 (1999)
4. D. Chiocca, G. Guizzi, T. Murino, R. Revetria, E. Romano, A methodology for supporting lean healthcare. Stud. Comput. Intell. **431**, 93–99 (2012)
5. R. Di Micco, D.R. Montella, G. Naviglio, E. Romano, Design of experiments in a single stage multi product kanban system. Frontiers Artif. Intell. Appl. **246**, 518–537 (2012)
6. M. Furusawa, K. Amakasu, Exact Simulation of Fish School Echoes and its Applications, OCEANS 2008—MTS/IEEE Kobe Techno-Ocean, no., pp.1, 6 (2008)

7. M. Furusawa, Prolate spheroidal models for predicting general trends of fish target strength. J. Acoust. Soc. Jpn. (E) **9**, 13–24 (1988)
8. D.A. Demera, M. Barangeb, A.J. Boydb, Measurements of three-dimensional fish school velocities with an acoustic Doppler current profiler. Fish. Res. **47**, 201–214 (2000)
9. C. Dai, D. Pi, Z. Fang, H. Peng, Wavelet transform-based residual modified GM(1,1) for hemispherical resonator gyroscope lifetime prediction. IAENG Int. J. Comput. Sci. **40(**4), 250–256 (2013) ISSN 1819-656X
10. S. Filippi, P Giribone, R Revetria, A Testa, An Integrated Model for Supporting Better Fishering Vessel Design by Modeling Fish Schools Dynamics Ready for HLA, Lecture Notes in Engineering and Computer Science. in *Proceedings of The World Congress on Engineering and Computer Science*, WCECS 2013, 23-25 October, 2013, San Francisco, USA, pp. 1009-1016 (2013)
11. A. Barbaro, B. Einarsson, B. Birnir, S. Sigurðsson, H. Valdimarsson, O.K. Pálsson, S. Sveinbjörnsson, P. Sigurðsson, Modelling and simulations of the migration of pelagic fish. ICES J Mar. Sci. **66**(5), 826–838 (2009)
12. K.J. Benoit-Bird, W.L. Whitlow, Acoustic backscattering by Hawaiian lutjanid snappers. Acoust. Soc. Am. **114**(5), 2757–2766 (2003)
13. J.W. Hannon, Image based computational fluid dynamics modeling to simulate fluid flow around a moving fish, University of Iowa Iowa Research Online (2011)
14. A.J. Holmin, Simulations of multi-beam sonar echos from schooling individual fish in a quiet environment. Acoust. Soc. Am. **132**(6), 3720–3734 (2012)
15. J.K. Horne, J.M. Jech, Multi-frequency estimates of fish abundance: constraints of rather high frequencies. ICES J Mar. Sci. 56(2), 184–199 (1999)
16. G. Guizzi, D. Miele, L.C. Santillo, E. Romano, New formalism for production systems modeling, 25th European Modeling and Simulation Symposium. EMSS **2013**, 571–576 (2013)
17. P. Holimchayachotikul, R. Derrouiche, K. Leksakul, G. Guizzi, B2B supply chain performance enhancement road map using data mining techniques, in *International conference on System Science and Simulation in Engineering—Proceedings*, pp. 336–341 (2010)
18. X.H. Li, R.G. Jianli, W. Hongru, *An Adaptive Meta-cognitive Artificial Fish School Algorithm, International Forum on Information Technology and Applications*, vol. 01 (IEEE Computer Society, Washington DC, 2009). ISBN 978-0-7695-3600-2
19. H.M. Manik, Underwater acoustic detection and signal processing near the seabed. Sonar Syst. **68**(9): 1973–1985 (2011) ISBN: 978-953-307-345-3, chapter 12
20. A.D. Rijnsdorp, W. Dol, M. Hoyer, M.A. Pastoors, Effects of fishing power and competitive interactions among vessels on the effort allocation on the trip level of the Dutch beam trawl fleet, ICES J. Mar. Sci. **57**(4), 927–937 2000
21. R. Samborski, M. Ziolko, Filter-based model of multimicrophone array in an adverse acoustic environment, Eng. Lett. **20**(4), 336–338 (2012) ISSN 1816-093X
22. D.J.T. Sumpter, J. Krause, R. James, Consensus decision making by fish. Curr. Biol. **18**(22), 1773–1777 (2008)
23. R.H. Towler, J.M. Jech, J.K. Horne, Visualizing fish movement, behavior, and acoustic backscatter. Aquat. Living Resour. **16(**3), 277–282 (2003) ISSN 0990-7440
24. Y. Tang, Y. Nishimori, M. Furusawa, The average three-dimensional target strength of fish by spheroid model for sonar surveys. ICES J. Mar. Sci. **66**, 1176–1183 (2009)
25. M. Soria, P. Fréon, F. Gerlotto, Analysis of vessel influence on spatial behaviour of fish schools using a multi-beam sonar and consequences for biomass estimates by echo-sounder. ICES J. Mar. Sci. **53**, 453–458 (1996)
26. R. Revetria, A. Catania, L. Cassettari, G. Guizzi, E. Romano, T. Murino, G. Improta, H. Fujita, Improving healthcare using cognitive computing based software: an application in emergency situation, Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 7345 LNAI, pp. 477–490 (2012)

27. T.C. Weber, H. Pena, J.M. Jech, Consecutive acoustic observations of an Atlantic herring school in the Northwest Atlantic. ICES J. Mar. Sci. **66**, 1270–1277 (2009)
28. E. Briano, C. Caballini, M. Mosca, R. Revetria, A system dynamics decision cockpit for a container terminal: The case of voltri terminal europe. Int. J. Math. Comput. Simul. **3**(2), 55–64 (2009)

Chapter 46
Fuzzy Controller Design for an FCC Unit

Hossein Tootoonchy and Hassan H. Hashemi

Abstract This paper examines the procedure for a nonlinear modeling and fuzzy controller design of a Fluidized Catalytic Cracking Unit, also known as FCCU. The case study is an FCCU plant in Abadan Refinery, Iran. FCCU is one of the most important sections in the Petrochemical industry. In 2006 alone, FCCUs refined one-third of the Crude Oil worldwide. FCCUs convert heavy distillates, such as Gasoil (feed) and Crude Oil, to Gasoline, Olefinic gases and other more usable products. Factors including but not limited to FCCU's high efficiency, and daily price fluctuations in Gas, Oil and Petrochemical products, make the optimization of such units the center of focus for both engineers and investors. Unlike the conventional controllers, Fuzzy Logic is the perfect choice for stochastic, dynamic and nonlinear processes where the mathematical model of the plant cannot be produced, or if realizable, a great deal of approximation is involved. The heuristic approach in Fuzzy Logic controllers is the closest form to the human language, and this virtue will make it a perfect candidate for a wide range of industrial applications. The investigations in this paper are simulated and proven by MATLAB Fuzzy Logic Toolbox R2013a. In this paper, the applicability and promising features of Fuzzy Logic controllers for such a complex and demanding plant will be investigated.

Keywords FCCU · Fuzzy controllers · Fuzzy logic · Fuzzy surface · Membership functions · Nonlinear modeling · Petrochemical plant

H. Tootoonchy (✉) · H. H. Hashemi
Department of Electrical Engineer, California State University, Fullerton, CA 92834, USA
e-mail: Tootoonchy@csu.fullerton.edu

H. H. Hashemi
e-mail: hhashemi@fullerton.edu

H. K. Kim et al. (eds.), *Transactions on Engineering Technologies*,
DOI: 10.1007/978-94-017-9115-1_46,

1 Introduction

The Fluidized Catalytic Cracking Units, also known as FCCUs, are amongst the most important and valuable facilities in Petrochemical plants. FCCUs convert the heavy weight Oil feed, like Gasoil, into lighter hydrocarbons, which are more valuable and applicable in industry. The overall economics of the refinery, largely depends on the economic operation of FCCU [1]. The unit consists of two separate, yet interconnected sections; the Riser reactor (Separator) and the Regenerator reactor. The riser reactor is where the cracking process occurs, and the Coke covers the Catalyst, and reduces its activity. The process is perfect for seasonal, and diverse range of products, and thus can be employed for various projects throughout the year. Figure 1 depicts a schematic diagram of a FCCU with its important instruments and sections [2].

Fuzzy Logic is a systematic mathematical formulation for investigating and characterizing processes. Although Fuzzy Logic is applicable to both linear and nonlinear systems, the true power appears when dealing with nonlinear, stochastic systems. It is the best choice when the mathematical model for the process does not exist, or exists but is too complex to be formulated. In these situations, difficulties arise in using conventional control methods [2, 3]. The FCCU popularity is mainly market driven, and this fact should also be considered in the controller design. Because most of the time, one or two products are in demand more seasonally, the adoptability of the plant is crucial in the long run [2, 3]. Since FCCU processes are notorious for being nonlinear, time variant and full of uncertainties, they are very difficult to model, simulate and control. For such processes, the conventional controllers (PID) become inefficient, since they require a good mathematical model of the plant. The inapplicability of conventional controllers will give rise to a series of advanced algorithms like Fuzzy Logic, which is fully able to handle the nonlinear and stochastic processes. Therefore, new methods and approaches are necessary, and inevitable [4].

Moreover, the FCCU continue to play a key role in any refinery as the primary conversion unit, and also, they are the key to profitability. The successful operation of the unit, determines whether or not the refinery can remain competitive in today's market. FCCUs utilize a micro-spheroid Catalyst that fluidizes when properly aerated. The main purpose of the unit is to convert high-boiling petroleum fractions called Gasoil to high-value, high-octane gasoline, and heating Oil. Gasoil is the portion of the Crude Oil that boils in the range of 650–10508+ °F (330–5508 °C) and contains a diversified mixture of paraffin, naphthenes, aromatics and olefins [5]. In the FCCU, feed Oil is contacted with re-circulating Catalyst and reacted in a riser tube. The feed Oil vaporizes and is cracked as it flows up the riser, thus forming lighter hydrocarbons (the gasoline fraction). Large amounts of Coke are formed as a by-product. The Coke deposits on the Catalyst and reduces its activity. The lighter hydrocarbon products are separated from the spent Catalyst in a Reactor. Steam is supplied to strip the volatile hydrocarbons from the Catalyst. The Catalyst is then returned to the regenerator where the Coke

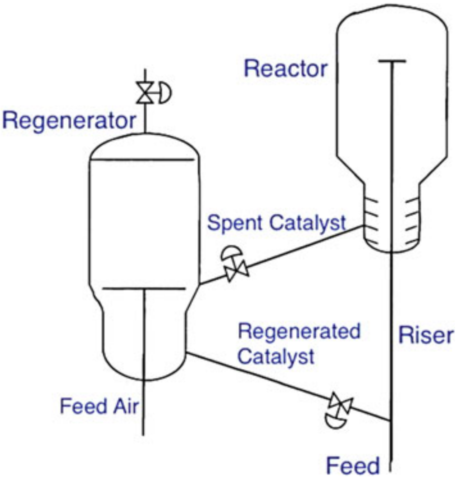


Fig. 1 Schematic diagram of an FCCU

is burnt off in contact with air. This is usually done by partial or complete combustion. The regenerated Catalyst is then re-circulated back to be mixed with the inlet feed Oil from the crude unit [6].

The selection of variables also plays a crucial role in the performance of Petrochemical plants like FCCUs. There are many discussions on proper selection of FCCU variables in Fuzzy optimization projects [7, 8]. However, the focus of this research is the key variables with which the process can be manipulated to achieve desirable results. These variables can either be categorized as Input-Output or Dependent-Independent ones. The input variables are Feed Rate, Specific Gravity, Catalyst Recirculation Rate, Air Flow Rate, Cumulative Feed Rate and Regenerator Temperature. The output variables are Riser Temperature, CO_2/CO ratio, Coke deposited on the Catalyst, Feed (Gasoil) conversion Rate, Coke and LPG. Selection of proper variables can be tricky and may lead to quite different results. A detailed review on variable selection and its consequent outcomes in FCCUs has been investigated [9].

2 Evolution of Fuzzy Logic in FCCU

Before the introduction of Fuzzy Logic, the investigations of scientists and researchers were limited to mathematical models, which had been exclusively developed for FCCU plants [10, 11]. These models had different levels of precision. Others focused their research on the comparison of different models and their advantages and disadvantages over each other [12]. Due to the importance of FCCUs in the industry and market, many scientists have approached this topic from different angles; stability, optimization, mathematical modeling and simulation are among the topics of interest. A complete literature on FCCU controllers and their continuous progress over the years have been explored [3, 9]. Also available are some significant works on the analysis and implementation of FCCUs with the focus on safe operation [13, 14]. In addition, there is plenty of research on the optimization and stabilization of FCCU plants [15–18].

An earlier work showed that the Fuzzy model has better accuracy compared to statistical methods for the process identification [19]. Many researchers and scientists have already tried to implement the linear regression techniques and complex Kalman filtering approaches to enhance the accuracy [20, 21]. Nevertheless, all of the aforementioned methods suffer from the inability to model and control a real plant, which has a great deal of inherent nonlinearity, impreciseness, and uncertainty [22, 23]. Thus, other new approaches, such as Neurofuzzy and Genetic Algorithm, started to emerge [19, 24–29]. Modern control techniques, such as parameter estimation, stochastic and optimal control, are used in model identification. However, some industrial processes are too complicated to be modeled or controlled by math-based algorithms because they are highly nonlinear and significantly uncertain with unknown or imprecise parameters [30]. Fuzzy Logic is an ideal tool for dealing with dynamic, nonlinear and imprecise models. It employs the linguistic rules to deal with mathematically vague processes and plants. These kinds of processes are widely present in industrial units, such as Petrochemical and nuclear plants and water treatment facilities.

For processes, which are known microscopically, the hard control is clearly the preferable methodology. However, conventional control techniques have generally failed to solve industrial problems with poorly developed mathematical models. Fuzzy Logic, and artificial neural networks are two examples of soft computing, which have migrated into the realm of industrial control over the last two decades. Chronologically, Fuzzy control was the first and its application in the process industry have led to significant improvements in product quality, productivity and energy consumption. Currently, Fuzzy control is firmly established as one of the leading advanced control techniques in practice. Today, the scientific trend is toward Fuzzy Logic and Neural networks [31]. The intelligent control becomes the center of interest when the system parameters can be manipulated to derive the results using familiar linguistic rules. The goal of this study is then to find the nonlinear relationship between Input-Output variables and define a solid optimization scheme to increase the efficiency by reducing the deposited Coke on the Catalyst and increasing the Gasoil conversion and LPG production.

3 Fuzzy Control: A Conceptual Review

Fuzzy Logic is a system that emulates the human expert decisions. Therefore, it is intuitively easy for humans to comprehend and use it in engineering and non-engineering applications. Fuzzy Logic results require no further elaboration or explanation because the results are often times are described in terms like cold,

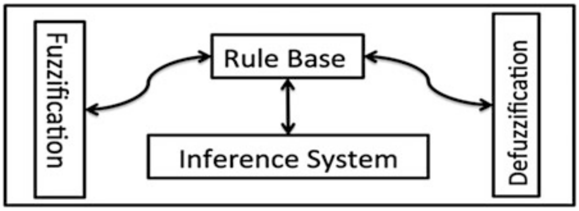


Fig. 2 Typical components of fuzzy logic controller

hot, small, big, fast, slow, which are easy for everyone to understand, and comprehend. In order to implement Fuzzy Logic, the knowledge, and experience of an expert are necessary. The experience is written in a rule-based format, which is used for making database as well as Fuzzy rules. The more accurate the rules are, the more applicable the results will be. It is noteworthy to mention that these rules are approximate, just like a human's decisions [3]. The human expert can be substituted with a combination of Fuzzy rule-based system (FRBS), and a block called defuzzifier. The sensory crisp data is then fed into the system where the physical values are represented or compressed into heuristic variables based on the appropriate membership functions. These linguistic variables then will be used in antecedent (IF-Then) part of statements, and will be changed, and revised to a crisp (numerical) output that represents an approximation to the actual output y(t) during the defuzzification process. The key point of Fuzzy Logic is that it does not require deep knowledge of the plant itself, or how the processes are interconnected. This useful characteristic is not feasible with conventional controllers like PIDs [31, 32]. Figure 2 depicts the main parts of a Fuzzy Logic controller. Note that (1) the "rule-based system" holds the knowledge in the form of a set of rules showing how to best control the system; (2) the inference mechanism evaluates which control rules are relevant to the current time and decides what the inputs to the plant should be accordingly; (3) the Fuzzification interface modifies the inputs so that they can be interpreted and compared to the rules in the rule-base; and (4) the defuzzification interface converts the conclusions reached by the inference mechanism into the inputs to the plant [32].

4 Fuzzy Modeling of FCCU

As mentioned earlier, due to the nonlinearity and time variance, developing a transfer function for the FCCU plant is not feasible. There are other solutions to approach the transfer function, method including the use of predefined and linearized models, which of course are accurate only to a certain degree. However, Fuzzy Logic modeling offers another solution, which is an excellent compromise in terms of accuracy, parameter manipulation, and applicability. The Fuzzy Logic

Table 1 Process variables—fuzzy logic inputs and outputs

Inputs	Outputs
Regenerator temperature (RET)	Coke as bypass product (Coke)
Specific gravity factor (SG)	Liquefied petroleum (LPG)
Airflow to regenerator (ATR)	Deposited coke on catalyst (DCC)
Gasoil (Feed)	CO_2/CO
Cumulative feed rate (CFR)	Riser temperature (RIT)
Airflow rate	Recycled feed rate (RFR)
Recycled catalyst rate	Regenerator gas temperature

modeling contains different sections, including but not limited to, correct diagnosis and selection of process Inputs and Outputs, Fuzzy Logic algorithm, defining the Fuzzy Logic rules, and Fuzzification-Defuzzification processes.

4.1 Variable Selection: Input-Output Parameters

The data in this study is gathered from the operation manual and other technical documentations of the Abadan FCCU refinery in 2004. Due to the lack of the mathematical model, rule-based Fuzzy approach is employed. In order to have the plant modeled, the FCCU operating variables have been identified as input and output ones, which will correspond to independent and dependent variables, respectively. Table 1 shows 16 major variables in an FCCU process, including the manipulative and measured variables. The Fuzzy controller determines the behaviors of the variables and their relationships with each other via generating dynamic nonlinear graphs, known as surface graphs. Among all of the factors affecting FCCU, and also based on their importance and level of consequence in the process, six input and six output variables are selected. In order to optimize the plant, measured, and manipulative variables are also identified. The Riser and Regenerator temperatures will be monitored all the time, and the Catalyst feed rate and airflow rate, as the manipulative variables, will be altered to adjust the parameters and achieve the desired results. The list of the input and output variables employed in this study is as follows: The input variables are Feed Rate, Specific Gravity, Catalyst Recirculation Rate, Air Flow Rate, Cumulative Feed Rate and Regenerator Temperature. The output variables are Riser Temperature, CO_2/CO ratio, Coke deposited on the Catalyst, Feed (Gasoil) conversion Rate, Coke and LPG. Selection of proper variables can be tricky and may lead to quite different results. A detailed review on variable selection and its consequent outcomes in FCCUs has been already investigated [9].

Table 2 Variable clustering and range allocation

Clustering groups	Equivalent action
Low	Small impact
Medium	Steady state
High	High impact

4.2 Fuzzy Logic Controller Design for FCCU

In order to process the Input-Output nonlinear relationship, six steps are considered in the creation of the rule-based Fuzzy system [34]. These steps are shown in Table (2):

- Identify the inputs and their ranges, then Fuzzify the inputs
- Identify the outputs and their ranges
- Create Fuzzy membership function, and degrees of truth
- Create the Rule-base required for controller design
- Assign weight of the rules, and their interactions
- Combine the rules, and defuzzify the output.

The data clustering for membership functions is also shown in Table 3. For generating knowledge base and rules, the knowledge of an experienced Process engineer and a Senior Instrumentation engineer, who were working on the plant for a long time, was used. The rules were later refined using the operation manuals, and other technical documents issued by the licensor.

4.3 Fuzzy Logic Rules

The list of the rules connecting the input to output variables are shown below. These rules were implemented in Matlab Fuzzy rule editor to generate the inference and nonlinear surface model [2].

1. If (SG is H) then (LPG is M)(GOCR is H)
2. If (SG is H) then (Coke is H) (CO_2/C0 is H)
3. If (SG is H) then (DCC is M) (RIT is L)
4. If (SG is L) then (CO_2/C0 is L)
5. If (SG is L) then (RIT is H)
6. If (ATR is H) then (Coke is H)
7. If (ATR is H) then (RIT is M)(CO_2/C0 is M)
8. If (ATR is M) then (CO_2/C0 is M)
9. If (ATR is M) then (DCC is L)
10. If (ATR is M) then (Coke is M)
11. If (ATR is L) then (CO_2/C0 is L)

Table 3 Clustering ranges for input variables

Input variables	Low (L)	Medium (M)	High (H)
RET (°C)	0–610	575–645	630–670
SG (-)	0–0.660	0.452–0.796	0.668–0.878
ATR (m^3/h)	0–29,873	27,001–47,209	39451–60167
Gasoil (m^3/d)	0–1,980	1,770–2,151	1,967–2,250
CFR (m^3/d)	0–2,260	2,011–2,450	2,289–2,650
CRR (t/min)	0–15.2	11.2–16.1	14.9–16.9

Table 4 Clustering ranges for output variables

Output variables	Low (L)	Medium (M)	High (H)
Coke (wt%)	0–4.2	3.4–7.5	5.2–9.1
LPG (wt%)	0–18.5	14.3–21.8	19.5– 30.9
DCC (-)	0–0.791	0.397–0.865	0.753–0.980
GOCR (wt%)	0–76.8	44.9–93.8	79.3–98.16
CO_2/CO (mol/mol)	0–1.8	0.9–3.9	2.2–6.2
RIT (°C)	50–479	404–520	505–528

12. 1f (ATR is L) then (DCC is M)
13. If (RET is H) then (RIT is M)(CO_2/C0 is L)
14. If (RET is H) then (DCC is H)(LPG is M)(GOCR is L)
15. If (RET is H) then (Coke is M)(DCC is H)
16. If (RET is H) then (RIT is H)
17. 1f (RET is M) then (Coke is M)(LPG is M)(GOCR is H)
18. If (RET is M) then (CO_2/C0 is M)
19. 1f (RET is M) then (Rh T is H)
20. If (RET is M) then (Coke is M)
21. If (RET is L) then (RIT is M)
22. 1f (RET is L) then (DCC is L)
23. If (RET is L) then (CO_2/C0 is L)
24. 1f (RET is L) then (Coke is L)
25. If (RET is L) then (LPG is M)(RIT is L)(GOCR is L)
26. If (CFR is H) then (RIT is M)(GOCR is M)
27. If (CFR is H) then (DCC is L)(LPG is H)(RIT is M)(GOCR is H)
28. If (CFR is M) then (DCC is M)(LPG is M)(RIT is M)
29. If (CFR is L) then (DCC is M)(LPG is L)(GOCR is H)
30. If (CRR is H) then (Coke is M) (RIT is H) (GOCR is L)(CO_2/C0 is H)
31. If (CRR is M) then (Coke is M)(GOCR is M)
32. If (CRR is L) then (Coke is L)(GOCR is M)
33. If (Gasoil is H) then (RIT is M)(GOCR is L)(CO_2/C0 is L)
34. If (Gasoil is M) then (GOCR is M)
35. If (Gasoil is L) then (GOCR is H)

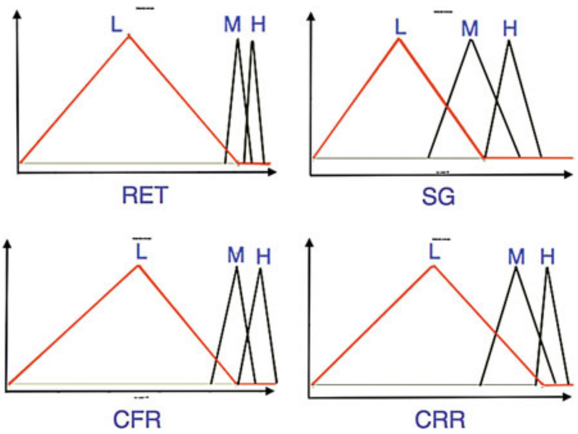


Fig. 3 Input-output membership functions-for accurate results refer to Tables 3, 4

In Fuzzy control, there is an emphasis on using rules, while in conventional control, this level of emphasis is on ordinary differential equations (ODEs). Using linguistic rules rather than the math-based system is more natural to human cognition. In Fuzzy rule, the rules are always true, but to different levels, ranging from zero to one. The inference system first checks if the premises of the rules are valid for the current case. If the premises are satisfied, those rules are selected, and consequent actions ensue. This step is also known as "Matching." The inference system makes the decisions afterwards.

Another advantage of Fuzzy modeling is that once the whole system is modeled, selection of different variables and thus different modes can be easily evaluated without further manipulation of the plant or the controller. Many possibilities can be checked and verified seamlessly only by selecting variables from dropdown menus.

Figure 3 depicts four of the membership functions that were used in this study. It is noteworthy to mention that the user is free to choose different forms for membership functions, including the Gaussian, triangular, and even unconventional ones. However, unless the characteristics of the actual process, such as covariance of the parameters involved, or the correlation of the process and noise, are known, selection of different forms is not going to affect the outcome very much, and often surface figures produced are very similar.

4.4 *Fuzzification-Defuzzification*

The Fuzzification is preparing the input value, which is extracted by sensors and most of the time analog, and then finding a value corresponding to it in membership functions. Defuzzification is the last step Fuzzy controller takes to produce

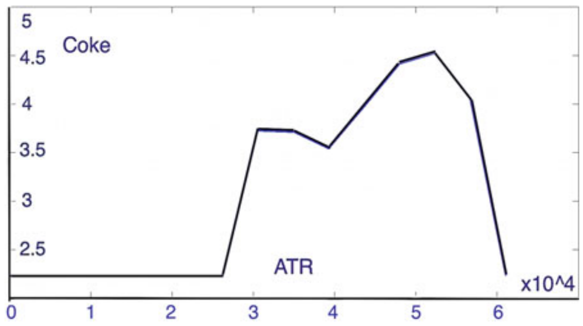


Fig. 4 Coke production based on ATR only

the control signal, which will be fed into the plant via manipulating variables. The inference mechanism selects the most certain situation, and produces the output accordingly. Defuzzification aims to produce a nonfuzzified control action that best represents the possibility distribution of the inferred Fuzzy decision [4].

Fuzzy surface provides very valuable information about the plant, including the correlation of the Input-Output variables; speed of the system reacts to the changes in the input and the direction of changes. These types of information enable engineers to analyze the plant in a completely new way, which is not feasible by conventional control methods.

5 Results

After finalizing the rules in the Fuzzy rule editor section, the results are produced by generating the surface, which can be chosen to be 2D or 3D, readily. Conventional control has successfully provided the industry with satisfactory results with which many math-based problems can be addressed accurately. However, the inability of this type of control, along with its dependence on approximating the nonlinear and highly dynamic plants, have made the Fuzzy Logic a superior choice for control engineers dealing with nonlinear and dynamic cases. In Fuzzy control, ODEs are replaced with the skill of an expert in the field. In recent years, many scientists have focused on other aspects of Fuzzy Logic, such as learning through experience. The Neuro Fuzzy approach is now well established in the industry and has a very promising future.

The data in this paper was extracted from operational manuals, and also experience of the user to generate the necessary information for Fuzzy Logic modeling and controller design. This information was implemented in the Matlab Fuzzy Logic Toolbox 2013a, and the results were demonstrated in Figs. 4, 5 and 6. The Fuzzy Logic is exempted from the heavy mathematical formulations to produce the output. However, the expertise and knowledge of an experienced operator are essential. The plant examined in this paper used to be controlled and monitored

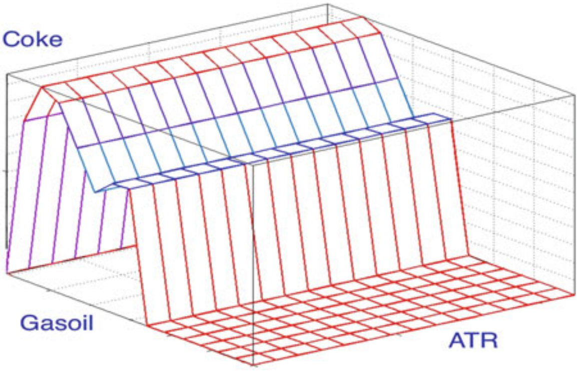


Fig. 5 Coke production based on Gasoil and ATR

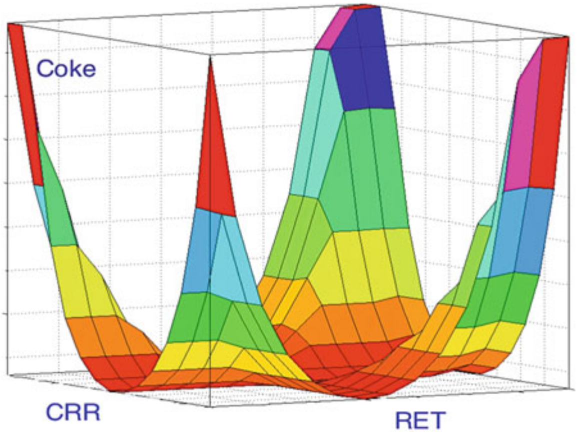


Fig. 6 Coke Production according to CRR and RET

via a Distributed Control System (DCS) and the rules had been implemented by customized programming, which required parameter manipulation and manual intervention when modes were switched or at the time of maintenance and overhauls. The data generated in this paper had an acceptable precision with those obtained in plant operation. The Fuzzy controller was successfully modeled and could produce decent results. Some of the results found in this research were in good compliance with actual ones in operation; others like CRR and CFR required more tuning to be acceptable.

Once the Fuzzy modeling of the plant is complete, many useful insights and conclusions can be made. For instance, in Fig. 4, the Coke production pattern according to ATR is demonstrated. As depicted, the Coke production follows a linear trend from 26,000 to 30,000 (wt%) according to the input variable ATR, and consequently, it can be modeled via a linear mathematical formula.

In addition, upon the increment of ATR variable from 30,000 to 40,000 (m^3/h), a linear decrease in Coke production is observed. The maximum of Coke production occurs at ATR value of 54,000. This pattern recognition also enables engineers to implement the numerical optimization techniques to enhance the plant productivity. Another noteworthy advantage of the Fuzzy Logic over conventional control is observed in the 3D result of Fig. 6. As shown, the 3D graph near the origin is almost similar to that generated via a PID controller. Therefore, a Fuzzy Controller can produce the results of a PID controller while the conventional controllers are not the proper means to address the nonlinear and uncertain industrial models like FCCUs.

6 Conclusion and Future Work

Fuzzy Logic approach proved to be capable of generating satisfactory results while facing nonlinear and dynamic situations. It was also the prefect tool to address the nonlinearity and uncertainty inherent in FCCUs and Petrochemical plants in general. A Fuzzy controller was designed to address the nonlinearity and uncertainty of the FCC plant with acceptable performance. Generating more accurate rules, increasing the number of rules and using numerical optimization techniques can further refine the results. In addition, modern control techniques, such as Neuro Fuzzy and Artificial Intelligence, can be employed to tailor the outcome further.

References

1. F.Z. Tatrai, P.A. Lant, P.L. Lee, C. Ian T, R.B. Newell, Control relevant model reduction: a reduced order model for model IV fluid catalytic cracking units. J. Process Control **4**(1), 3–14 (1994)
2. H. Tootoonchy and H. Hashemi, in *Fuzzy Logic Modeling and Controller Design for a Fluidized Catalytic Cracking Unit*. Proceedings of The World Congress on Engineering and Computer Science 2013, WCECS. Lecture Notes in Engineering and Computer Science (San Francisco, USA, 2013), pp 982–987, 23–25 Oct 2013
3. H. Taşkin, C. Kubat, Ö. Uygun, S. Arslankaya, FUZZYFCC: Fuzzy logic control of a fluid catalytic cracking unit (FCCU) to improve dynamic performance. Comput. Chem. Eng. **30**(5), 850–863 (2006)
4. Chun Chien Lee, Fuzzy logic in control systems: fuzzy logic controller. Part II IEEE Trans. Syst. Man Cybern. **20**(2), 419–435 (1990)
5. R. Sadeghbeigi *Fluid Catalytic Cracking Handbook: an Expert Guide to the Practical Operation, Design, and Optimization of FCC Units* (Elsevier, Oxford, 2012)
6. A. Arbel, Z. Huang, I.H. Rinard, R. Shinnar, No Title. Ind. Eng. Chem. Res. **34**, 1228 (1995)
7. M.F. Azeem, N. Ahmad, M. Hanmandlu, Fuzzy modeling of fluidized catalytic cracking unit. Appl. Soft Comput. **7**(1), 298–324 (2007)
8. A. Arbel, Z. Huang, I.H.I. Rinard, R. Shinnar, A.V. Sapre, Dynamic and control of fluidized catalytic crackers 1 modeling of the current generation of FCC's. Ind. Eng. Chem. Res. **34**(4), 1228–1243 (1995)

9. V.W. Weekman Jr, Model of catalytic cracking conversion in fixed, moving, and fluid-bed reactors. Ind. Eng. Chem. Process Des. Dev. **7**(1), 90–95 (1968)
10. I. Pitault, D. Nevicato, M. Forissier, J.-R. Bernard, Kinetic model based on a molecular description for catalytic cracking of vacuum gas oil. Chem. Eng. Sci. **49**(24), 4249–4262 (1994)
11. R. Maya-Yescas, F. López-Isunza, Comparison of two dynamic models for FCC units. Catal. Today **38**(1), 137–147 (1997)
12. J. Alvarez-Ramirez, R. Aguilar, F. López-Isunza, Robust regulation of temperature in reactor-regenerator fluid catalytic cracking units. Ind. Eng. Chem. Res. **35**(5), 1652–1659 (1996)
13. M.I. Wan-lin, The application of TRICON ESD in FCCU. Shenyang Chem. Ind. **4**, 22 (2005)
14. Y.-Z. Lu, M. He, C.-W. Xu, Fuzzy modeling and expert optimization control for industrial processes. Control Syst. Technol. IEEE Trans. **5**(1), 2–12 (1996)
15. M. Delgado, M.A. Vila, J. Kaprzyk, J.L. Verdegay, *Fuzzy optimization: recent advances* (Springer, New York, 1994)
16. J.M. Cadenas, J.L. Verdegay, Towards a new strategy for solving fuzzy optimization problems. Fuzzy Optim. Decis. Mak. **8**(3), 231–244 (2009)
17. E.E. Ali, S.S.E.H. Elnashaie, Nonlinear model predictive control of industrial type IV fluid catalytic cracking (FCC) units for maximum gasoline yield. Ind. Eng. Chem. Res. **36**(2), 389–398 (1997)
18. T. Takagi, M. Sugeno, Fuzzy identification of systems and its applications to modeling and control. Syst. Man Cybern. IEEE Trans. **1**, 116–132 (1985)
19. E.T. Van Donkelaar, P.S.C. Heuberger, P.M.J. Van den Hof, Identification of a fluidized catalytic cracking unit: an orthonormal basis function approach, in *Proceedings of the American Control Conference*, vol 3 (1998), pp. 1914–1917
20. Y. Huang, S. Dash, G.V Reklaitis, V. Venkatasubramanian, EKF based estimator for FDI in the model IV FCCU, in *Proceedings of SAFEPROCESS* (2000), pp. 14–16
21. L.A. Zadeh, Outline of a new approach to the analysis of complex systems and decision processes. Syst. Man Cybern. IEEE Trans. **1**, 28–44 (1973)
22. E.H. Mamdani, Application of fuzzy algorithms for control of simple dynamic plant. Electr. Eng. Proc. Inst. **121**(12), 1585–1588 (1974)
23. C.-W. Xu, Y.-Z. Lu, Fuzzy model identification and self-learning for dynamic systems. Syst. Man Cybern. IEEE Trans. **17**(4), 683–689 (1987)
24. M. Sugeno, T. Yasukawa, A fuzzy-logic-based approach to qualitative modeling. IEEE Trans. Fuzzy Syst. **1**(1), 7–31 (1993)
25. Y. Nakamori, M. Ryoke, Identification of fuzzy prediction models through hyperellipsoidal clustering. Syst. Man Cybern. IEEE Trans. **24**(8), 1153–1173 (1994)
26. A. Kandel, G. Langholz, *Fuzzy Control Systems*. (CRC press, Boca Raton, 1994)
27. A. Lotfi, A.C. Tsoi, Learning fuzzy inference systems using an adaptive membership function scheme. IEEE Trans. Syst Man Cybern. Part B Cybern. **26**(2), 326–331 (1996)
28. K. Nozaki, H. Ishibuchi, H. Tanaka, Adaptive fuzzy rule-based classification systems. Fuzzy Syst. IEEE Trans. **4**(3), 238–250 (1996)
29. Y.-Z. Lu, *Industrial Intelligent Control: Fundamentals and Applications*. (John Wiley & Sons, Chichester, 1996)
30. S.A. Manesis, D.J. Sapidis, R.E. King, Intelligent control of wastewater treatment plants. Artif. Intell. Eng. **12**(3), 275–281 (1998)
31. S. Passino, K.M. Yurkovich, K.M. Passino, S. Yurkovich, Fuzzy control. Citeseer **42**, 498 (1998)
32. S. Sumathi, S. Panneerselvam, *Computational Intelligence Paradigms: Theory and Applications Using MATLAB*. (CRC Press, New York, 2010)
33. P.B. Venuto, E.T. Habib Jr, *Fluid Catalytic Cracking with Zeolite Catalysts* (Marcel Dekker, New York, 1979)

Chapter 47
A Cost-Criticality Based (Max, +) Optimization Model for Operations Scheduling

Karla Quintero, Eric Niel, José Aguilar and Laurent Piétrac

Abstract The following work proposes a (max, +) optimization model for scheduling batch transfer operations in a flow network by integrating a cost/criticality criterion to prioritize conflicting operations in terms of resource allocation. The case study is a seaport for oil export where real industrial data has been gathered. The work is extendable to flow networks in general and aims at proposing a general, intuitive algebraic modeling framework through which flow transfer operations can be scheduled based on a criterion that integrates the potential costs due to late client service and critical device reliability in order to satisfy a given set of requests through a set of disjoint alignments in a pipeline network. The research exploits results from previous work and it is suitable for systems handling different client priorities and in which device reliability has an important short-term impact on operations.

Keywords Algebraic modeling · Flow networks · Oil pipeline networks · (max, +) theory · Schedule optimization · System reliability

K. Quintero (✉)
Thales Group, 78141 Vélizy, France
e-mail: karla.quintero@thalesgroup.com

E. Niel · L. Piétrac
INSA, 69621 Lyon, France
e-mail: eric.niel@insa-lyon.fr

L. Piétrac
e-mail: laurent.pietrac@insa-lyon.fr

J. Aguilar
ULA, Mérida 5101, Venezuela
e-mail: aguilar@ula.ve

H. K. Kim et al. (eds.), *Transactions on Engineering Technologies*,
DOI: 10.1007/978-94-017-9115-1_47,

1 Introduction

The following work continues the developments in [1, 2] in which a (max, +) optimization model for scheduling transfer operations on a flow network was proposed. The case study continues to be a seaport for oil export in which oil batches must be transported between 2 different points through a path in an intricate pipeline network.

Given that different oil batches must not mix, conflict in resource allocation arises in order to process more requests than the network is able to handle simultaneously. Some approaches to manage resource allocation conflicts are Petri Nets, specifically event graphs, where conflicts are previously solved through a routing policy. This routing policies are criteria allowing to choose a transition among a group of conflicting ones demanding to be fired. Naturally, the routing policy must be coherent with the special needs of each system to be modeled; see [3–5] for an overview on common routing policies. Some heuristic approaches can also be considered; for instance, [6] implements an ant colony optimization algorithm in which conflict is modeled as a probabilistic choice rule depending on the pheromone trail and a heuristic function.

In [1, 2], conflict resolution has been intuitively modeled for a flow network through (max, +) algebra. In these developments, industrial data indicates that in case of delayed service the seaport incurs into a different monetary penalty depending on the client. In those results, the objective was to find a schedule minimizing the Total Cost due to Penalties (TCP) incurred by the seaport. In [1], fixed preventive maintenance tasks on valves were considered as constraints in the model in order to schedule oil transfer operations optimally. In [2], maintenance relaxation was explored in order to obtain better global schedules through a tradeoff between satisfying maintenance operations and satisfying a given set of clients implying some potential costs in the case of delays. Also, in that work, conflict resolution in resource allocation was explored based on the Total Potential Penalty (TPP) of each client, i.e. a product (in conventional algebra) between the request's processing time and the penalty per time unit.

In this paper, we explore the integration of failure risk into the already established penalty-based framework. More specifically, here an analogy is done with an approach, proposed in [7], used to prioritize maintenance tasks on devices and it is modified in order to prioritize conflicting oil transfer operations. Namely, operations are proposed to be prioritized according to an index reflecting the failure probability of the underlying alignment in the network by the monetary consequence which is associated to potential penalties. In order to do so, the failure probability of an alignment must be established, for which we rely on some previous work on alignment search techniques for the case study.

Some approaches, other than (max, +) based, formulating similar optimization problems are: [8], where an optimization model for flow-shop scheduling with setup times is formulated as sets of recursive constraints expressing the dependency between completion times for jobs on machines, and [9, 10], with classic resource conflict constraints where decision variables impose a precedence between machine operations. The fundamental ideas of these approaches are similar to the proposed model but with the algebraic structure provided by the (max, +) approach constraint formulations can be intuitively built and additional and more intricate phenomena can be easily integrated.

To our knowledge, no similar work has been developed for this type of system, other than the foundations in [1, 2] which constitute the base of this work. The results are extendable to applications to flow networks of different nature. The developments in this work are part of a larger research scope aiming at optimizing operations in a more complex framework with industrial application in the oil sector.

Firstly, we present the case study in Sect. 2. Section 3 shows some basic notions on (max, +) algebra. Section 4 presents the proposed (max, +) optimization model and the proposed criterion for prioritizing conflicting operations based on penalty costs and alignment failure probability, which is presented in Sect. 5. Results are shown in Sects. 6 and 7 presents concluding remarks.

2 Case Study

The case study is a seaport for oil export, but the work is be extendable to flow networks of different nature. Oil batches requested by clients must be transported from a set of tanks to a set of loading arms placed at the docks of the seaport through an intricate pipeline network. It is considered that oil flows by gravity through the pipeline network as it is the case of some real seaports.

2.1 Oil Transfer Aspects

An oil transfer operation represents the transfer of a requested oil batch (of a specific type and quantity) from a tank to a specific dock. In reality, a dock may be equipped with one or several loading arms which load the oil batch into the tanker that requests it. Here, it is considered only one loading arm per dock. Each tanker has a loading deadline to be respected which, if exceeded, implies a monetary penalty incurred by the seaport. This penalty is related to the time delay and also to the client's priority. Each of these requests is fulfilled through the selection of an alignment (i.e. a path) in the oil pipeline network, which implies opening the valves included in this alignment and closing all adjacent valves, in order to isolate it from the rest of the network since two types of oil must not mix.[1] From industrial

[1] a specific scenario is the mixture of two identical oil types. However, oil mixture is not allowed in any scenario since sharing an alignment's section by two transfer operations could result in lower product flow rate and several aspects such as pumping power and pipeline dimensions would have to be considered and are not the focus of this work.

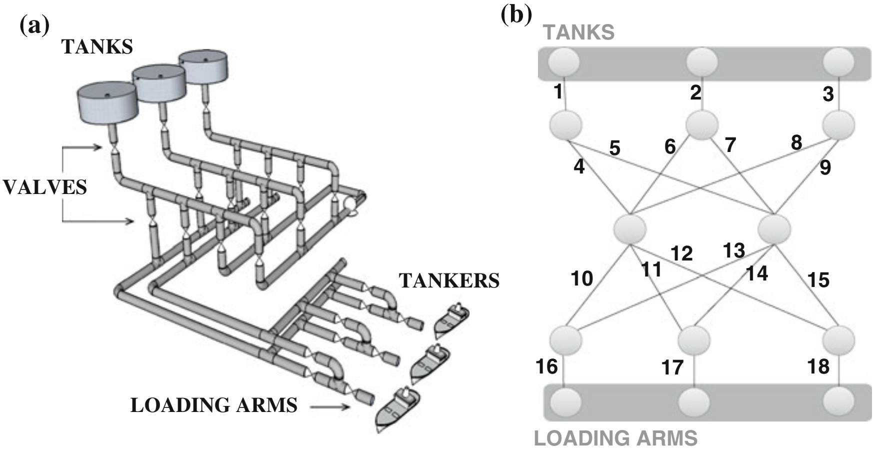


Fig. 1 Oil seaport example

data it is known that oil transfer operations take hours, whereas valve commutations are assumed to take seconds. In this work, it is considered that the alignment is previously established for each transfer operation.

Considerable effort has been devoted to optimizing other features for the case study, most of the results being adaptable for flow networks in general. [11] can be consulted for generic alignment selection maximizing operative capacity (i.e. simultaneous disjoint alignments) in the network and [12] for generic alignment selection maximizing operative capacity while minimizing failure risk on valves. For illustration purposes on the system configuration, Fig. 1a shows an example of a simplified oil seaport and Fig. 1b shows the network model as an undirected graph where arcs represent the valves and the nodes represent the linked pipeline segments.

2.2 *Conflicts in Resource Allocation*

Simultaneous alignments for two or more requests must be disjoint since different oil batches must not mix. The work in [1] yields the following definition.

Definition 1 Two or more alignments (for oil transfer) are in conflict if they share at least one valve and if either the valve requires different states for different alignments or if it requires being open for more than one alignment.

Figure 2a (from [1]) shows two disjoint alignments to satisfy requests R_1 and R_3. Solid lines illustrate the valves to open and dotted lines (of the same tone) the valves to close in order to isolate the alignment; e.g.: to enable the alignment for R_1 valves 1, 4, 10, and 16 must open and valves 5, 6, 8, 12, 11, and 13 must close. In Fig. 2a, no conflict arises for any valve since the common resources (valves 5,

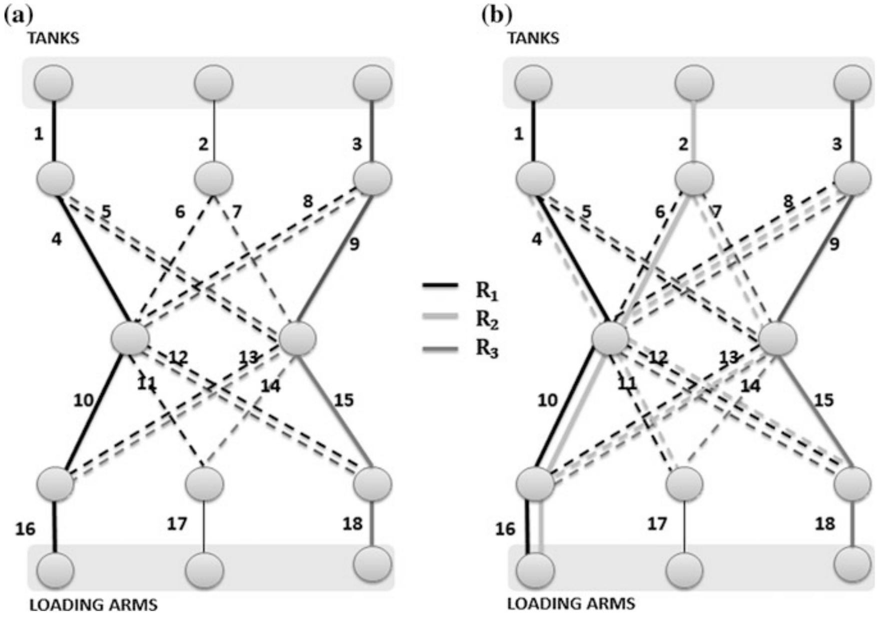


Fig. 2 Conflict representation in a undirected graph

8, 12, and 13) are all valves to be closed, therefore they can enable both transfer operations simultaneously.

On Fig. 2b, another request (R_2) is added and conflicts arise for valves 10 and 16, since they should open for 2 transfer operations (therefore, mixing 2 oil types), and for valves 4 and 6, since the required commutations are different for both transfer operations (which is physically impossible); therefore, R_1 and R_2 cannot be processed simultaneously.

2.3 Scheduling Oil Transfer Operations on a Seaport: Penalty-Related Aspects

In this work, resources of interest are valves and their availability is determined by their allocation by different alignments aiming at satisfying oil transfer operations for several clients. Client requirements include deadlines for tanker loading, which in case of violation by the seaport imply monetary penalties.

For each client, a negotiation occurs with the seaport. In this phase, the client imposes (within certain conditions not relevant to this work) for a specific tanker, the penalty to be paid by the seaport in case of delay (in thousands of dollars per hour) caused by the seaport. At the same time, the seaport imposes a time window of three days within which the tanker can arrive and be immediately docked and served. From the moment of arrival within this time window, the maximum service time for every tanker is 36 h for loading and 4 h for paperwork. Since the focus of this paper is on seaport transfer operations, the paperwork interval is discarded and the focus is on the maximum loading interval of 36 h as the deadline for each tanker. From that point on, every extra hour invested in the service of the tanker will result in a penalty for the seaport, if the delay has been indeed caused by the seaport. Conversely, if the service of a tanker surpasses the 36 h due to tanker's technical difficulties, then the client pays the seaport a penalty for dock over-occupation. Operations management on the seaport contributes to the general objective of profit maximization but client-incurred-penalties do not represent in any way an optimization objective, i.e. they are unexpected events which the seaport does not aim at maximizing through operations' scheduling. If the tanker arrives after its time window, the seaport does not incur into any penalties for the waiting time for the tanker to be served. No further information has been granted concerning other arrival scenarios and possible consequences in the service.

3 Preliminaries on (max, +) Algebra

This section provides a (max, +) theory overview allowing to understand the basis of this mathematical modeling technique with application to the scheduling problem approached in the research.

(max, +) algebra is defined as a mathematical structure denoted as R_{max}, constituted by the set $R \bigcup \{-\infty\}$ and two binary operations $\oplus$ and $\otimes$, which correspond to maximization and addition, respectively. This algebraic structure is called an idempotent commutative semifield. As [13] states, a semifield $\mathcal{K}$ is a set endowed with two generic operations $\oplus$ and $\otimes$ complying with certain classic algebraic properties. The zero element is $\varepsilon = -\infty$, and the identity element is $e = 0$. The main properties of this algebraic structure (similar to the ones defined in conventional algebra) are:

Operation $\oplus$:

- is associative (e.g. $a \oplus (b \oplus c) = (a \oplus b) \oplus c$),
- is commutative (e.g. $a \oplus b = b \oplus a$),
- has a zero element ε (e.g. $a \oplus \varepsilon = a$),
- is idempotent (i.e. $a \oplus a = a$; $\forall a \in \mathcal{K}$).

Operation $\otimes$:

- is distributive with respect to $\oplus$ (e.g. $a \otimes (b \oplus c) = (a \otimes b) \oplus (a \otimes c)$),
- is invertible. For example, in (max, +) algebra: if $2 \otimes 3 = 5$ then $2 = 5 ø 3$ or in conventional notation: if $2 + 3 = 5$ then $2 = 5 - 3$ (here, operator ø denotes the inverse of the $\oplus$ operation),
- has an identity element e which satisfies $\varepsilon \otimes e = e \otimes \varepsilon = \varepsilon$.

Some equivalent (max, +) and conventional algebra expressions are the following:

$$\begin{array}{cc} a \oplus b \Leftrightarrow max(a,b) & a \otimes b \Leftrightarrow a + b \\ a \oplus \varepsilon \Leftrightarrow max(a,\varepsilon) = a & a \otimes \varepsilon \Leftrightarrow a + \varepsilon = \varepsilon \\ a \oplus e \Leftrightarrow max(a,e) = a & a \otimes e \Leftrightarrow a + e = a \end{array}$$

Based on the aforementioned basic operations, more intricate ones are defined in the context of the algebraic structure such as matrix product for example. Synchronization phenomena can be modeled in a very straightforward fashion through (max, +) algebra which has led to a very wide application to transportation systems. However, research in this field continues to explore further possibilities.

For the purposes of this research, the interest is on the application of the modeling technique to a system in which resource allocation conflicts constitute the main feature.

The application of this theory to discrete event systems exhibiting synchronization phenomena leads to the formulation of very intuitive (max, +)-linear models formed by equations such as $x_3 = x_1 \otimes \tau_1 \oplus x_2 \otimes \tau_2$. In this equation, x_i is the start date of an event i, and τ_i is its duration. x_i is usually denoted as '*dater*' in the (max, +) context. In this example, the dater of event 3 is given by the maximum of the completion times of events 1 and 2; which can be interpreted as the synchronization of 2 tasks or 2 task sequences (e.g. a train that only departs when 2 other trains arrive at the station with connecting passengers).

With the principle shown in the former equation, a (max, +)-linear system model describing the interactions among all relevant tasks or processes can be obtained in the form of $X = AX$, where X is the variables vector (i.e. $X = [x_1 \; x_2 \; \ldots \; x_n]^T$) and A is the matrix containing all time relations between the variables. Analogies with classic linear system theory would be applicable to this simple model by considering maximization and addition as basic operations, as well as all the aforementioned properties in the algebraic structure.

(Max, +) theory is a research field that has caught the attention of the scientific community for its intuitive modeling potential of discrete event systems' phenomena that would usually involve more intricate mathematical models. For further information on (max, +) algebra for production chains and transportation networks [14] can be consulted. [13] can be consulted for (max, +)-linear system theory, [15] for (max, +) theory applied to traffic control, [16] for an application to production scheduling in manufacturing systems, and [17] for maintenance modeling for a helicopter. Moreover, considerable effort has been dedicated to exploiting the potential of (max, +) algebra combined with automata theory, leading to the study of (max, +) automata which can also be applicable to schedule optimization problems; see [18–20] for developments in this field.

4 (max, +) Optimization Model

The basis for the mathematical optimization model used in this paper have been defined in [1]. In this work, maintenance aspects are not considered and the main objective is to prioritize operations according to a relation between the potential penalty due to late service and the reliability of the network alignment. In [1] the purpose was to minimize the *TCP*, which was defined as the total cost in which the seaport incurs due to late service of a set of clients for a time horizon. One of the main set of constraints proposed is the one modeling conflicts in resource allocation, presented in (1) in conventional algebra and in (2) in (max, +) algebra,[2] where $\mathcal{J}$ is the set of all possible requests.

$$x_i = max(t_0; u_i; max_{i'}(x_{i'} + p_{i'} + zp_{i'} + zc_{i'} + V_{i,i'})) \quad \forall i, i' \in \mathcal{J} | i \neq i' \tag{1}$$

$$x_i = t_0 \oplus u_i \oplus (\oplus_{i'}(x_{i'} \otimes p_{i'} \otimes zp_{i'} \otimes zc_{i'} \otimes V_{i,i'})) \quad \forall i, i' \in \mathcal{J} | i \neq i' \tag{2}$$

The aforementioned equivalent constraints determine the start date (x_i), also called '*dater*' in the (max, +) context, to satisfy a request i. Variables include:

- x_i: dater for an oil transfer operation, also called request i,
- $x_{i'}$: dater for a conflicting transfer operation i' requesting common resources to operation i,
- $V_{i,i'}$: decision variable that defines the precedence between two conflicting oil transfer operations i and i',
- u_i: arrival date for the tanker for request i,
- $zp_{i'}$, and $zc_{i'}$ represent, respectively, the possible unexpected delays in client service due to technical difficulties in the seaport and due to technical difficulties within the tanker.
- Parameters include: t_0, and $p_{i'}$ which respectively correspond to the start date of the scheduling time horizon, and the nominal duration of the oil transfer operation.

Equation (2) states that the start date for an oil transfer operation will depend on the start date of the time horizon for scheduling, the arrival date of the tanker in the seaport, and the completion time of all conflicting oil transfer operations which are to be executed before request i. Notice that for all conflicting operations interruption variables model the possible delays that could arise in the execution of operations. All decision variables are binary, taking the values $e = 0$ or $\varepsilon = -\infty$, as the identity and zero elements defined in (max, +) theory. For instantiation purposes, values are *zero* or B, so that B is a very large negative real number.

Moreover, each decision variable has a complementary one (e.g. if $V_{i,i'} = e$, then $V_{i',i} = B$ or vice versa). For example, in (2), when $V_{i,i'} = B$ the entire third

[2] taking into consideration that maintenance is not approached in this work.

term of the global maximization is negligible, which implies that the completion time of operation $x_{i'}$ is not relevant to calculate x_i, indicating that request i will be executed before request $i^{'}$. This value assignment would automatically generate the value assignment of the complementary decision variable (i.e. $V_{i',i} = e$) which means that in the constraint to determine $x_{i'}$ the completion time of operation i would indeed be taken into account.

$$V_{i,i'} \otimes V_{i',i} = B \quad \forall\, i, i' \in \mathcal{J} \tag{3}$$

$$V_{i,i'} \oplus V_{i',i} = e \quad \forall\, i, i' \in \mathcal{J} \tag{4}$$

Equations (3) and (4) restrict the values of the decision variables to be either *zero* or B for potential conflicts between two transfer operations.

$$D_i = \begin{cases} u_i \otimes 36 & \forall\, i \in \mathcal{J} | u_i \in tw_i \\ x_i \otimes 36 & \forall\, i \in \mathcal{J} | u_i > utw_i \\ ltw_i \otimes 36 & \forall\, i \in \mathcal{J} | u_i < ltw_i \end{cases} \tag{5}$$

$$dpr_i = (x_i \otimes p_i \otimes zp_i \otimes zc_i ø D_i) \oplus e \qquad \forall i \in \mathcal{J} \tag{6}$$

In (5) the deadline D_i for a request i is modeled where $tw_i = [ltw_i, utw_i]$ is the authorized time window of three days for the tanker's arrival. Since no further information has been gathered on deadlines given early arrival of the tanker, it has been considered that the 36 h for loading start at the beginning of the authorized time window.

The *delay per request* (*dpr*) is determined in (6) which is the difference between a request's completion time (including the possible delays caused by the seaport and/or the client) and its deadline. In [1], Hypothesis 1 was proposed to deal with combined delays between the seaport and the client. Within this context, (7) modeled the *penalized delay for the seaport* (*pds*) per request; i.e. the time interval (hours) for which the seaport will actually incur into penalties.

Hypothesis 1. the dock over-occupation penalty per hour per client (paid by each client) is considered equal to the penalty per hour for that same client paid by the seaport in the case of delay caused by the seaport.

$$pds_i = \begin{cases} ø[(øzu_i øzp_i \otimes zc_i) \oplus (ødpr_i)] & \forall (zu_i \otimes zp_i) > zc_i \\ e & \text{otherwise} \end{cases} \tag{7}$$

Notice that (5–7) allow to determine the *TCP* as presented in (8), which has already proven to be a crucial metric in operations management. Equation (8) is the (max, +) algebra representation for the sum of the products of each penalized delay (in hours) and its corresponding penalty (in \$/hour).

$$Min\ TCP = \bigotimes_{i} \left(\bigotimes_{n=1}^{pds_i} c_i \right) \ \forall i \in \mathcal{J} \tag{8}$$

In [1], the optimal schedule for oil transfer operations was obtained while considering a fixed preventive maintenance program to be respected. In [2], some (max, +)-linear representations of the model were obtained through value assignment of decision variables based on a routing (or conflict resolution) policy which consisted on prioritizing operations with the greatest *TPP*, defined as the product of the nominal duration and the penalty per time unit for the tanker. In this work, the focus lies on exploring a routing policy that integrates reliability data of the network section of interest with related potential costs. Namely, failure probability on each alignment is considered in order to estimate consequent costs (measured as potential penalties for late service due to failure of the predefined alignment) and hence prioritize operations according to a failure/cost relation.

The basic premise is that this approach is useful in systems where device reliability is a very influential metric when it comes to managing operations. This could be related with 'forced production' situations, in which maintenance is forced to be delayed due to operational requirements and therefore device reliability is crucial in carrying out operations in the network. Given such scenario, in which device conditions are susceptible to influence operational performance, alignment failure consequently implies potential penalties for the request that is being processed with such alignment. This approach is the result of an analogy applied with a similar approach used to prioritize maintenance activities on devices as it is stated in [7]. In the developments therein, an index called CBC (Cost-Based Criticality) is used to rank maintenance tasks based on the device's failure probability and its consequence. The index is computed as the product between the device's failure probability and the consequent monetary costs that arise due to failure (in which production losses, environmental impact, quality loss, are considered among other costs).

In this work, a similar index is used as a routing policy to solve conflicts between oil transfer operations with alignments sharing common resources. Alignments are considered to be previously defined for each request. Alignment's failure probability is computed based on previous work on alignment search for the case study and the monetary consequence of failure is considered as the TPP for each specific client. The aforementioned index is denoted in this work as Penalty-Criticality Index (PCI) and is defined in (9), where $TPP_i = p_i \times c_i$ (as defined in [2]) which is the product between the processing time for operation i and the penalty per hour for such client, and $P_{f(i)}$ is the failure probability of the alignment assigned to process such request.

$$PCI_i = TPP_i \times P_{f(i)} \tag{9}$$

5 Failure Probability for an Alignment

In [12], an approach was proposed to find the greatest set of independent simultaneous alignments (also called maximum operative capacity) in a pipeline network while minimizing failure risk for the same case study. The approach was based on a minimum flow cost algorithm in which costs were related to devices' reliability and flow was considered to be either existent or nonexistent on pipeline segments (i.e. no flow rate was managed).

For an alignment to function properly in order to carry out an oil transfer operation, all valves in the alignment should be able to commute to the 'open' state properly and all adjacent valves should commute to the 'closed' state properly, in order to isolate the alignment. Considering that proper commutation behavior on each valve is independent from all others and that it can be described as a random variable, an alignment's estimation of a well-functioning probability can be described as the product of the well-functioning probabilities for each and every one of the valves involved. This metric is defined in (10), where $v \in A$ stands for all valves involved in alignment A.

For example, in the case of Fig. 2, for the alignment for request R_1, the well-functioning probability ($P_{w(i)}$) would be $P_{w(i)} = P_1 \times P_4 \times P_{10} \times P_{16} \times P_5 \times P_6 \times P_8 \times P_{12} \times P_{11} \times P_{13}$. Consequently, the alignment's failure probability is defined as $P_{f(i)} = 1 - P_{w(i)}$. It is fundamental to understand that this metric is used exclusively for differentiation purposes among alignments and their condition in order to execute a set of given requests.

$$P_{w(i)} = \prod_{(v \in A)} P_{f(v)} \tag{10}$$

This approach is in no way restrictive, and the well-functioning probability could be determined otherwise for a different system, and could be fed with the proper probability estimations in each flow network according to condition monitoring results on devices in the best case scenario. Moreover, a different criticality level according to the needs of each particular system could be considered in order to prioritize transfer operations.

6 Results

Figure 3 shows an instance with 7 requests to be executed through the depicted alignments (only open valves are depicted for better comprehension). In this figure, 3 zones are identified as A, B, and C in order to define 3 different probability values for well-functioning behavior on valves (for illustration purposes). Most valves clearly fall into a specific zone, and those that do not are considered as follows: valves 5 and 13 $\in$ Zone A, and 8 and 12 $\in$ Zone C. In Fig. 3, conflicts

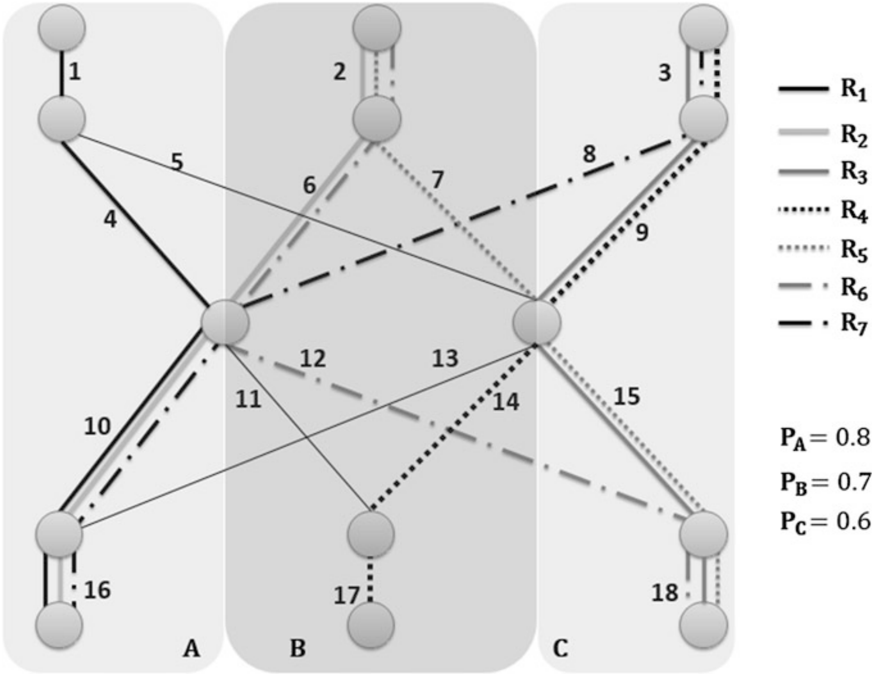


Fig. 3 Alignments for oil transfer operations

among alignments can be easily identified, and according to the structure proposed in (2), the set of conflict constraints is obtained in (11–17).

$$x_1 = u_1 \oplus x_2 p_2 V_{1,2} \oplus x_6 p_6 V_{1,6} \oplus x_7 p_7 V_{1,7} \tag{11}$$

$$x_2 = u_2 \oplus x_1 p_1 V_{2,1} \oplus x_5 p_5 V_{2,5} \oplus x_6 p_6 V_{2,6} \oplus x_7 p_7 V_{2,7} \tag{12}$$

$$x_3 = u_3 \oplus x_4 p_4 V_{3,4} \oplus x_5 p_5 V_{3,5} \oplus x_6 p_6 V_{3,6} \oplus x_7 p_7 V_{3,7} \tag{13}$$

$$x_4 = u_4 \oplus x_3 p_3 V_{4,3} \oplus x_5 p_5 V_{4,5} \oplus x_7 p_7 V_{4,7} \tag{14}$$

$$x_5 = u_5 \oplus x_2 p_2 V_{5,2} \oplus x_3 p_3 V_{5,3} \oplus x_4 p_4 V_{5,4} \oplus x_6 p_6 V_{5,6} \tag{15}$$

$$x_6 = u_6 \oplus x_1 p_1 V_{6,1} \oplus x_2 p_2 V_{6,2} \oplus x_3 p_3 V_{6,3} \oplus x_5 p_5 V_{6,5} \oplus x_7 p_7 V_{6,7} \tag{16}$$

$$x_7 = u_7 \oplus x_1 p_1 V_{7,1} \oplus x_2 p_2 V_{7,2} \oplus x_3 p_3 V_{7,3} \oplus x_4 p_4 V_{7,4} \oplus x_6 p_6 V_{7,6} \tag{17}$$

Table 1, presents input data in columns 2, 3 and 4 as it is known in real operational conditions (i.e. the operation's processing time p_i, as well as the penalty per time unit for that specific tanker c_i, and finally the obtained *TPP*). The failure probability of each alignment is computed with the reliability probabilities

Table 1 Input data for linear model validation

Request	p_i (hours)	c_i (\$/hour)	TPP_i	$P_{f(i)}$
R_1	20	3000	60000	0.954
R_2	20	3000	60000	0.975
R_3	20	3000	60000	0.985
R_4	20	2000	40000	0.980
R_5	20	2000	40000	0.980
R_6	20	4000	80000	0.980
R_7	20	4000	80000	0.974

proposed in Fig. 3 for each valve depending on the zone, and through the approach proposed in Sect. 5, yielding the results on column 5 in Table 1.

In this table, operations' processing times are assumed to be equal to allow a better manual comprehension of the prioritization of conflicting tasks. This is in no way restrictive and it is only assumed for result illustration purposes. The result from the proposed data is the vector of *PCI* indices as $PCI = [57240, 58500, 59100, 39200, 39200, 78400, 77920]^T$.

Assuming the worst case scenario, in which all tankers for all requests arrive at the same time (which is unlikely but holds for illustration purposes), and all within their authorized time windows, then all potential conflicts (due to resource sharing) become actual conflicts that must be dealt with by assignment of decision variable values according to the proposed *PCI* criterion.

Given the obtained *PCI* vector, R_6 is the most pressing operation. Consequently, this operation does not depend on the completion time of other conflicting operations and therefore decision variable assignment must be such that for a conflicting operation i, $V_{6,i} = \varepsilon$.

In (16), this translates into the value assignments: $V_{6,1} = V_{6,2} = V_{6,3} = V_{6,5} = V_{6,7} = \varepsilon$, which automatically yields the assignments of all complementary variables in all other equations, i.e. $V_{1,6} = V_{2,6} = V_{3,6} = V_{5,6} = V_{7,6} = e$. Analogously, all remaining decision variables values are assigned according to the *PCI* prioritization criterion, e.g. in (11) $V_{1,2} = e$ since operation R_1 depends on the completion time of R_2, because $PCI_2 > PCI_1$, which yields $V_{2,1} = \varepsilon$ in (12), and so forth. Hence, from 11 to 17, the following (max, +)-linear system is obtained:

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \\ x_7 \end{pmatrix} = \begin{pmatrix} \cdot & p_2 & \cdot & \cdot & \cdot & p_6 & p_7 \\ \cdot & \cdot & \cdot & \cdot & \cdot & p_6 & p_7 \\ \cdot & \cdot & \cdot & \cdot & \cdot & p_6 & p_7 \\ \cdot & \cdot & p_3 & \cdot & \cdot & \cdot & p_7 \\ \cdot & p_2 & p_3 & p_4 & \cdot & p_6 & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & p_6 & \cdot \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \\ x_7 \end{pmatrix} \oplus \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \\ u_5 \\ u_6 \\ u_7 \end{pmatrix}$$

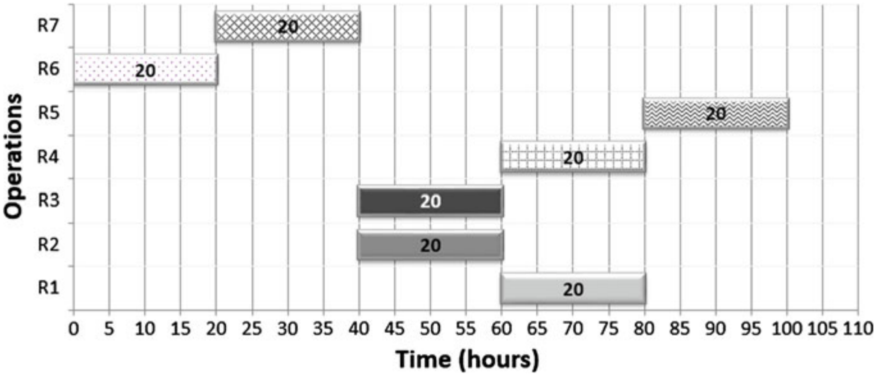


Fig. 4 Prioritization

Since the system is (max, +)-linear, the model's structure is quite simple and intuitive. This linearity property can be explored eventually as it is done for classic linear systems in conventional algebra. This is however not the focus of this work.

For simplicity, let the arrival dates be: $u = [e, e, e, e, e, e, e]^T$, i.e. all tankers arrive at $t_0 = 0$. The obtained schedule is as shown in Fig. 4. Through this schedule, resource allocation is done according to exploitation conditions of the network and the underlying costs that could be generated due to device malfunctioning. Notice that in other scenarios other than the one shown in Fig. 4 some operations could be executed simultaneously, therefore reducing the makespan. However, knowing that through the *PCI* vector operations are already ranked, this forced simultaneous execution would actually reduce device reliability and increase failure risk for following tasks with higher priority. For example, R_4 could be executed from $t_0 = 0$ simultaneously with R_6 but this would decrease resource reliability for R_7 which would be executed later and has higher priority.

7 Concluding Remarks

The proposed approach exploits a (max, +) optimization model in order to schedule operations in a way that conflict resolution is managed through prioritization of operations according to a cost-reliability relation. The proposal approaches the case where device reliability can vary in a short-term, therefore affecting operative capacity with consequent costs related to penalties due to late service. Some other approaches have been explored focused on minimizing the *TCP* (see [1]). The approach proposed in this paper does not aim at replacing these previous results but is rather complementary, enriching the information that can be provided to supervision operators in order to improve decision making in a given operational situation.

Acknowledgments This research has been supported by Thales Group France, and by the PCP (Post-graduate Cooperation Program) between Venezuela and France involving the collaboration between the academic institutions: ULA (in Spanish: *Universidad de Los Andes*)—research laboratory: CEMISID in Mérida, Venezuela and the INSA (in French: *Université de Lyon, INSA Lyon, Ampère* (UMR5005)) in Lyon, France; and the industrial partners Thales Group France and PDVSA (in Spanish: *Petróleos de Venezuela Sociedad Anónima*), the Venezuelan oil company. Industrial data for model validation has been granted by PDVSA.

References

1. K. Quintero, E. Niel, J. Aguilar et al. (Max, +) Optimization model for scheduling operations in a flow network with preventive maintenance tasks, in *Proceedings of The World Congress on Engineering and Computer Science 2013, WCECS 2013*. Lecture Notes in Engineering and Computer Science, vol. 2 (San Francisco, USA, 2013), pp. 1036–1041, 23–25 Oct 2013
2. K. Quintero, E. Niel, J. Aguilar et al., Scheduling operations in a flow network with flexible preventive maintenance: a (max, +) approach. Eng. Lett. **22**, 24–33 (2014)
3. M. Alsaba, J.-L. Boimond, S. Lahaye, On the control of flexible manufacturing production systems by dioid algebra (originally in french: Sur la commande des systèmes flexibles de production manufacturière par l'algèbre des dioïdes). Revue e-STA, Sciences et Technologies de l'Automatique **4**, 3247–3259 (2007)
4. A. Nait-Sidi-Moh, M.-A. Manier, A. El Moudni et al. Petri net with conflicts and (max, +) algebra for transportation systems, in *11th IFAC Symposium on control in transportation systems*, 11 pp. 548–553
5. W. Ait-Cheik-Bihi, A. Nait-Sidi-Moh, M. Wack, Conflict management and resolution using (max, +) algebra: application to services interaction, in *Evaluation and Optimization of Innovative Production Systems of Goods and Services: 8th International Conference of Modeling and Simulation—MOSIM'10* (2010)
6. S.G. Ponnambalam, N. Jawahar, B.S. Girish, An ant colony optimization algorithm for flexible job shop scheduling problem, in *New Advanced Technologies* (2010), http://www.intechopen.com/books/new-advanced-technologies/An Ant Colony Optimization Algorithm for Flexible Job Shop Scheduling Problem. Accessed 11 Feb 2014
7. W. Moore, A. Starr, An intelligent maintenance system for continuous cost-based prioritization of maintenance activities. Computers in industry (2006), doi:10.1016/j.compind.2006.02.008
8. K. Yang, X. Liu, A bi-criteria optimization model and algorithm for scheduling in a real-world flow shop with setup times, in *Proceedings of the International Conference on Intelligent Computation Technology and Automation (ICICTA)*, vol.1, (2008), pp. 535–539
9. Z. Zhao, G. Zhang, Z. Bing, Job-shop scheduling optimization design based on an improved GA, in *Proceedings of the 10th World Congress on Intelligent Control and Automation* (*WCICA*), 2012, pp. 654–659
10. C. Zeng, J. Tang, H. Zhu, Two heuristic algorithms of job scheduling problem with inter-cell production mode in hybrid operations of machining, in *25th Chinese Control and Decision Conference (CCDC),* 2013, pp. 1281–1285
11. J. Rojas-D'Onofrio, J. González, E. Boutleux et al. Path search algorithm minimizing interferences with envisaged operations in a pipe network, in *Proceedings of the European Control Conference, ECC'09* (2009)
12. K. Quintero, E. Niel, J. Rojas-D'Onofrio, Optimizing process supervision in a flow network in terms of operative capacity and failure risk, in *15th International congress on automation, systems and instrumentation*, 2011
13. F. Baccelli, G. Cohen, G. Jan-Olsder, J.-P. Quadrat, *Synchronization and Linearity an Algebra for Discrete Event Systems*. (Wiley, New York, 2001)

14. G. Cohen, S. Gaubert, J.-P. Quadrat, Sandwich algebra (originally in french: lalg‘ebre des sandwichs). Pour la science **328**, 56–63 (2005)
15. L. Houssin, S. Lahaye, J.-L. Boimond, Just-in-time control under constraints of (max, +)-linear systems (originally in french: Commande en juste-à-temps sous contraintes de systèmes (max, +)-linéaires). Journal Européen des systèmes automatisés—JESA **39**, 335–350 (2005)
16. G. Nasri I, Habchi, R. Boukezzoula, An algebraic max-plus model for HVLV systems scheduling and optimization with repetitive and flexible periodic preventive maintenance: just-in-time production, in *9th International Conference of Modeling, Optimization and Simulation—MOSIM'12*, 2012
17. Z. Königsberg, Modeling, analysis and timetable design of a helicopter maintenance process based on timed event petri nets and max-plus algebra. Neural Parallel Sci Comput. **18**, 1–12 (2010)
18. S. Gaubert, Performance evaluation of (max, +) automata. IEEE Trans. Autom. Control **40**, 2014–2025 (1995)
19. S. Lahaye, J. Komenda, J.-L. Boimond, Modular modeling with (max, +) automata (originally in french: Modélisation modulaire à l'aide d'automates (max, +)), in *Conférence internationale francophone d'automatique*—CIFA. (Grenoble, France 2012)
20. J. Komenda, S. Lahaye, J.-L. Boimond, The synchronous product of (max, +) automata (originally in french: Le produit synchrone des automates (max, +)). Special issue of Journal Européen des Systèmes Automatisés—JESA **43**, 1033–1047 (2009)

Chapter 48
Detailed Analysis of Deformation Behavior of Plexiform Bone Using Small Specimen Testing and Finite Element Simulation

Nitin Kumar Sharma, Swati Sharma, Daya K. Sehgal and Rama Kant Pandey

Abstract Bone is a complex anisotropic and heterogeneous material. The structure of bone material is considered to be hierarchical in nature that changes from nano to macro level. Small punch testing, used in the present study, is a very useful technique to analyze the deformational behavior of materials that are difficult to obtain in a sufficient size for conventional mechanical testing. The finite element modeling (FEM) of small punch testing was carried out using ABAQUS code. The material properties of cortical bone for FE simulation were considered to be both isotropic and transversely isotropic in nature. This study shows that isotropic tensile properties of cortical bone are sufficient to predict the deformational behavior of cortical under small punch testing.

Keywords Anisotropy · Cortical bone · Finite element simulation · Heterogeneity · Small punch testing · Transversely isotropic material

N. K. Sharma (✉)
School of Technology, The Glocal University, Mirzapur Pole, Saharanpur 247001, India
e-mail: nitin@theglocaluniversity.in; enksharma@yahoo.com

S. Sharma · D. K. Sehgal · R. K. Pandey
Department of Applied Mechanics, Indian Institute of Technology Delhi,
Hauz Khas, New Delhi 110016, India
e-mail: swati.2517@gmail.com

D. K. Sehgal
e-mail: profsehgal@yahoo.com

R. K. Pandey
e-mail: rkpiitd@yahoo.com

H. K. Kim et al. (eds.), *Transactions on Engineering Technologies*,
DOI: 10.1007/978-94-017-9115-1_48,

1 Introduction

Bone is a well known biological composite material having hierarchical structure down to nano level. From material science point of view it is considered as a functional material with heterogeneous composition and anisotropic properties. The anisotropy in mechanical properties of bone has been regarded as being caused by a characteristic composite structure of constituents in bone. The main constituents of bone are rigid hydroxyapatite like minerals and pliant collagen fibril [1–5]. The mechanical properties of cortical bone are difficult to obtain due to various constrains such as complexity in bone material, shape and size of bone, availability of the specimen from bone diaphysis, testing procedures etc. [6–10]. These constraints may be overcome by employing small size specimens for mechanical testing. The advantage of small size specimens includes the possibility of sampling very small volume of material within a heterogeneous structure such as cortical bone. This may also be used for studying biological materials that are not available in volumes sufficient enough for conventional mechanical testing. The small specimen test techniques include a wide array of types and techniques as described by Lucas [11] and Cheon and Kim [12]: tensile test, micro hardness, small punch (ball, shear, and hemispherical head), bend, fracture, impact, and fatigue.

In the present investigation the mechanical behavior of cortical bone has been studies for two different orientations of loading with the help of small specimen testing and finite element (FE) simulation. The load-displacement behavior of cortical bone specimens was analyzed with the help of small punch testing technique. The small punch test is a version of the bulge test, the name given by Baik et al. [13]. Manahan et al. [14] and Huang et al. [15] were among the earliest to study the small punch test technique for the determination of mechanical properties of irradiated materials from small circular disks. These techniques largely involve loading a supported disc or coupon with an indenter or punch, deforming it to failure, and analyzing the resulting load-displacement data. Recently, the load-displacement behavior of longitudinal small cortical bone specimens was analyzed by Sharma et al. [16] with the help of small punch testing technique and FEM.

The finite element method (FEM) is a very versatile numerical technique for solving engineering problems of plasticity, impact mechanics, structural analysis and computational mechanics. The FE simulation of small punch test is a large deformation problem and was carried out using the commercially available ABAQUS code. The main advantage of applying FEM to large deformation problems is its versatility, i.e. its ability to simulate the progress of the deformation process and to obtain a complete solution to the problem including the deformation shape, nodal displacement, nodal velocity and distribution of stresses and strains. In the present study six different FE models of small punch test were developed for two orthogonal directions (longitudinal and transverse) of loading and for two different thicknesses (1 and 1.5 mm) of the bone specimens. The results of these models were compared with the corresponding experimental one to analyze the effect of heterogeneity and material properties on load-displacement behavior of cortical bone. For different FE models bone material was considered to be isotropic as well as transversely isotropic in nature.

2 Materials and Method

The study has been conducted on bovine femoral cortical bones obtained from farm raised just after animal's natural death. After removal of the soft tissue the bones were soaked in normal saline and wrapped in normal saline soaked cloth. These bones were further kept in plastic bag and stored at −20 °C before processing. The tensile properties of cortical bone in longitudinal direction (load being applied along the long axis of bone) and transverse directions (load being applied perpendicular to the long axis of bone) were evaluated using strip type dumbbell shape specimens. The longitudinal tensile specimens were prepared with gauge length 25 mm, gauge width 4 mm and total length 80 mm, whereas, the transverse tensile specimens were prepared with gauge length 8 mm, gauge width 4 mm and total length 22 mm. Poisson's ratio in each direction was tested with the help of biaxial extensometer of gauge length 25 mm. The longitudinal and transverse compressive properties of cortical bone were determined using cubic specimens of each edge equal to 5 mm. For small punch testing on cortical bone the small specimens were prepared with rectangular cross-section having 2 mm width and two different thicknesses of 1 and 1.5 mm. The length of these specimens was maintained equal to 10 mm. The small specimens were prepared from the longitudinal (with length parallel to the long axis of femur) as well as transverse (with length perpendicular to the long axis of femur) directions of the cortical bone to analyze the effect of directionality on the deformational behavior of bone. The longitudinal and transverse small specimens of cortical bone prepared for small specimen testing are shown respectively in Fig. 1a, b.

All different specimens for experimental testing were obtained from the middle diaphysis of the femoral bone and were stored at room temperature in a solution of 50 % saline and 50 % ethanol at all time until testing. Different experimental tests were conducted on MTS 858 Table Top Machine with very slow strain rates (i.e. 1.8 mm/min for tensile and compressive testing and 0.5 mm/min for small specimen testing). The fixture for small punch testing and its installation on MTS 858 table top machine is shown in Fig. 2.

The rectangular shape small specimen geometry is clearly not axisymmetric, so a three dimensional finite element model is required. In this study, a three dimensional finite element model has been proposed for simulating small punch test with rectangular shape small (fixed beam) specimen under quasi-static loading by special cylindrical tip headed rigid punch. ABAQUS standard package is used for computation and detailed computational results are obtained. For simulation the dimensions of the rectangular specimen are decided according to the inner square (4 mm × 4 mm) cross-section of the borehole of the specimen holder. The

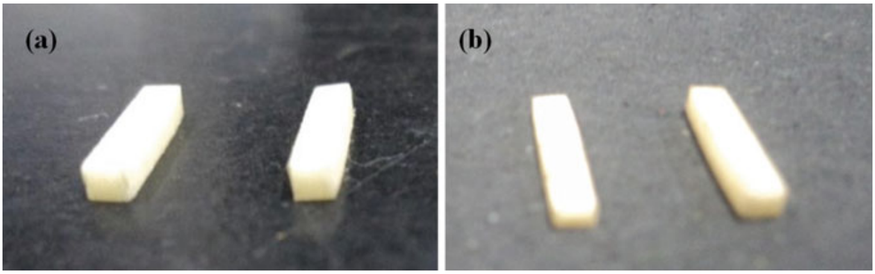


Fig. 1 Small **a** longitudinal and **b** transverse specimens of cortical bone of different thicknesses (i.e. 1 and 1.5 mm)

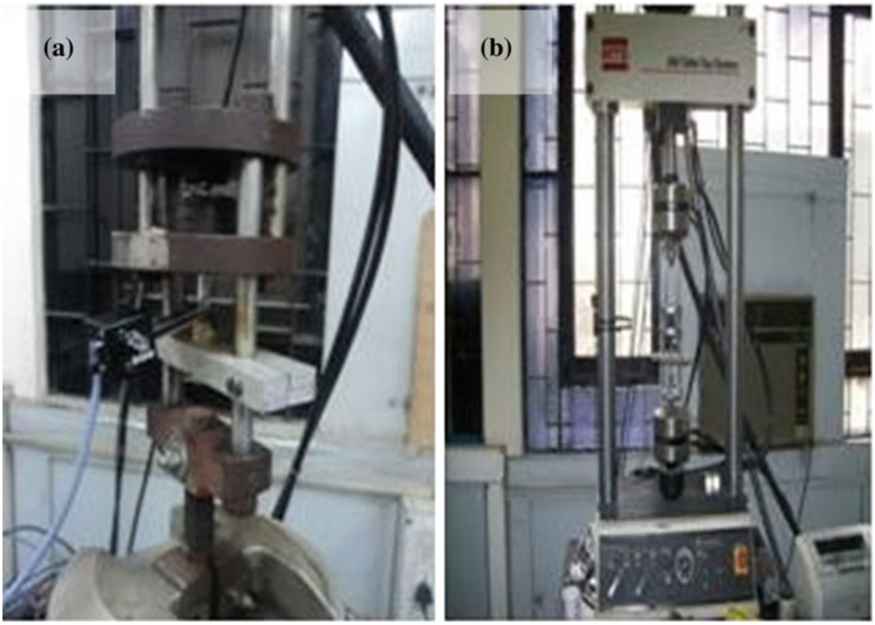


Fig. 2 Setup for small punch testing, **a** fixture for small punch testing and **b** installation of fixture on MTS 858 table top machine

three dimensional FE models of small specimen test in case of 1 and 1.5 mm thicknesses of the small cortical bone specimens are shown in Fig. 3. These models were discretized with second order 20-noded brick (hexahedra) elements.

The size of the cylindrical headed punch is considered as 1.82 mm diameter, 2.0 mm width and 4 mm height. The cylindrical headed tip rigid punch is modeled with rigid body option using rigid element from ABAQUS library. Two opposite side edges of the small specimen are considered as fixed by boundary condition option ENCASTER (motion constrained in all direction). Top surface of

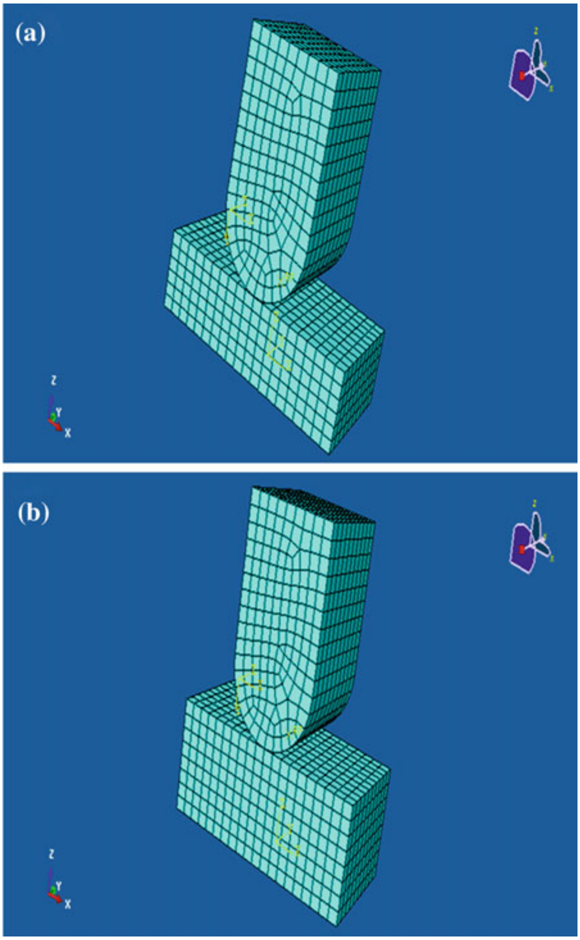


Fig. 3 Three dimensional FE models of small punch test in case of **a** 1 mm and **b** 1.5 mm thicknesses of the small specimens

rectangular specimen is defined as SLAVE surface for interaction. The cylindrical tip surface of rigid punch is defined as MASTER surface for interaction. A close surface interaction of SLAVE and MASTER surface is defined by contact algorithm. The quasi-static loading simulation is simulated with amplitude option and small incremental steps are used at reference node of cylindrical tip punch.

For FE simulation, cortical bone material was considered to be both isotropic and transversely isotropic in nature. The small specimen in shear punch testing is under pure bending situation, therefore, the upper surface of the specimen is under compression whereas the lower surface is under tension. This necessitates the partitioning of the small specimen as bone material may behave in a different manner under tensile and compressive loading. Therefore the behaviour of cortical bone specimen under small punch testing was analyzed using six independent FE models. These models were categories according to the specified material properties as; FTTI (having fully tensile mechanical properties considering bone material to be transversely isotropic), TCTI (having different behaviour under tension and compression with transversely isotropic material properties), FCTI (having fully compressive mechanical properties considering bone material to be transversely isotropic), FTI (having fully tensile mechanical properties considering bone material to be isotropic in nature), TCI (having different behaviour under tension and compression with isotropic material properties) and FCI (having fully compressive mechanical properties considering bone material to be isotropic in nature). The anisotropic yielding criterion in case of transversely isotropic material properties was specified using Hill's potential functions [17].

3 Results and Discussion

The tensile and compressive properties of cortical bone obtained from different experimental procedures are listed in Table 1. The notations used for different mechanical properties for tensile and compressive loadings and in longitudinal and transverse directions are as given below;

$$\begin{aligned} E_i^S &= \text{Elastic modulus of cortical bone in GPa} \\ \sigma_{ysi}^S &= \text{Yield stress of cortical bone in MPa} \\ \sigma_{usi}^S &= \text{Ultimate strength of cortical bone in MPa} \\ \varepsilon_{usi}^S &= \text{Fracture strain of cortical bone} \end{aligned}$$

where $i = 1, 2$ for respectively longitudinal and transverse directions of loading and $s = \text{C, T}$ for respectively the compressive and tensile loading

The experimental load-displacement diagrams of longitudinal and transverse small specimens obtained from small punch test have been compared with the corresponding diagrams obtained from the FE models in case of both 1 and 1.5 mm thicknesses of the small specimen. The comparisons of experimental load-displacement curve in case of longitudinal small specimens with those obtained from FE transversely isotropic models are shown in Fig. 4a, b respectively for 1 and 1.5 mm thicknesses of the specimens, whereas, for this case the corresponding comparisons with those obtained from FE isotropic models are shown in Fig. 5a, b respectively for 1 and 1.5 mm thicknesses.

For the case of transverse small specimens the comparisons of experimental and FE transversely isotropic curves in case of 1 and 1.5 mm thicknesses of the specimens are shown respectively in Fig. 6a, b, whereas for these specimens the comparisons of experimental curves with those obtained from FE isotropic models are shown in Fig. 7a, b respectively for 1.0 and 1.5 mm thicknesses.

The experimental and FE results of small punch test are reported in Tables 2 and 3 for respectively the longitudinal and transverse small specimens.

Table 1 Properties of cortical bone along different material orientation and loading conditions

Properties	Values n = 4
E_1^C	2.22 ± 0.8
E_2^C	5.1 ± 0.6
E_1^T	24.3 ± 1.9
E_2^T	15.3 ± 1.7
σ_{ys1}^C	89.2 ± 3.4
σ_{ys2}^C	84.5 ± 3.7
σ_{ys1}^T	117.7 ± 5.4
σ_{ys2}^T	78.6 ± 4.4
σ_{us1}^C	168.0 ± 3.4
σ_{us1}^C	204.1 ± 2.6
σ_{us1}^T	139.0 ± 4.2
σ_{us2}^T	84.2 ± 3.9
ε_{us1}^C	0.1328 ± 0.0023
ε_{us2}^C	0.1227 ± 0.0026
ε_{us1}^T	0.0228 ± 0.0043
ε_{us2}^T	0.0145 ± 0.0049

n = number of specimens tested and the values reported here are the average values

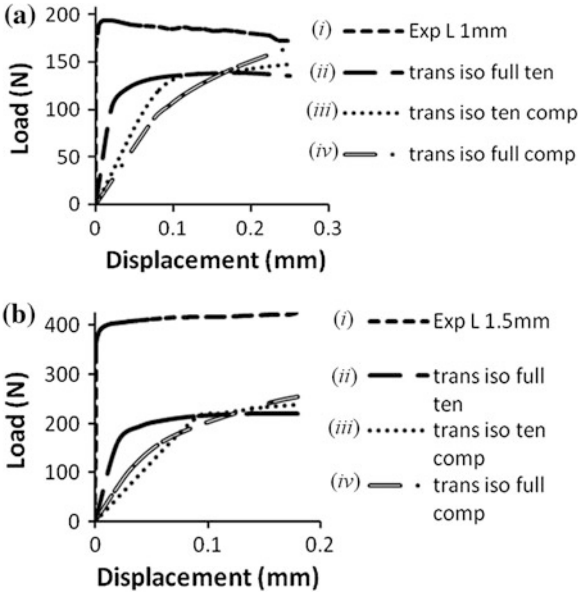


Fig. 4 Comparison between experimental and transversely isotropic finite element load-displacement diagrams for **a** 1.0 mm and **b** 1.5 mm thick longitudinal small specimen [where (*i*) = experimental curve, (*ii*) = FTTI, (*iii*) = TCTI and (*iv*) = FCTI]

As per the unpaired *t*-test analysis, for 1.0 mm thick longitudinal specimen, the failure load obtained from experimental curve was not found to be significantly different from finite element results of TCTI, FCTI, FTI, TCI, and FCI models,

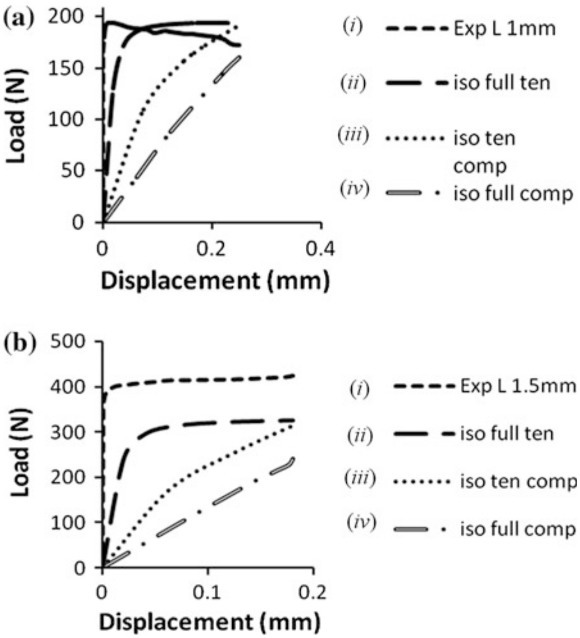


Fig. 5 Comparison between experimental and isotropic finite element load-displacement diagrams for **a** 1.0 mm and **b** 1.5 mm thick longitudinal small specimen [where (*i*) = experimental curve, (*ii*) = FTTI, (*iii*) = TCTI and (*iv*) = FCTI]

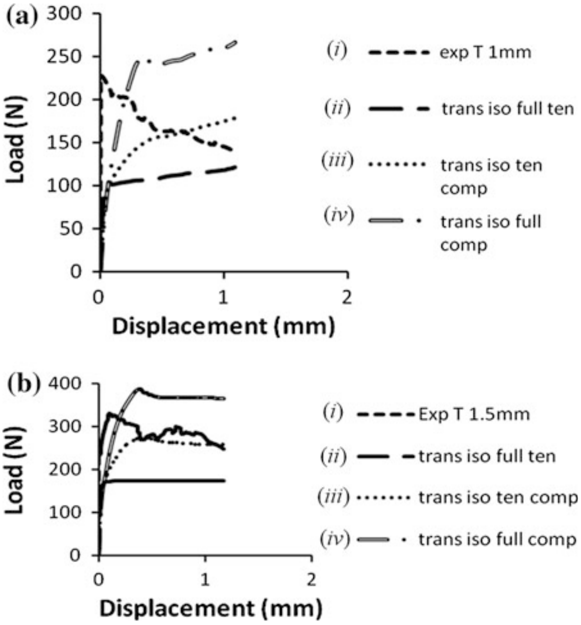


Fig. 6 Comparison between experimental and transversely isotropic finite element load-displacement diagrams for **a** 1.0 mm and **b** 1.5 mm thick small transverse specimen [where (*i*) = experimental curve, (*ii*) = FTTI, (*iii*) = TCTI and (*iv*) = FCTI]

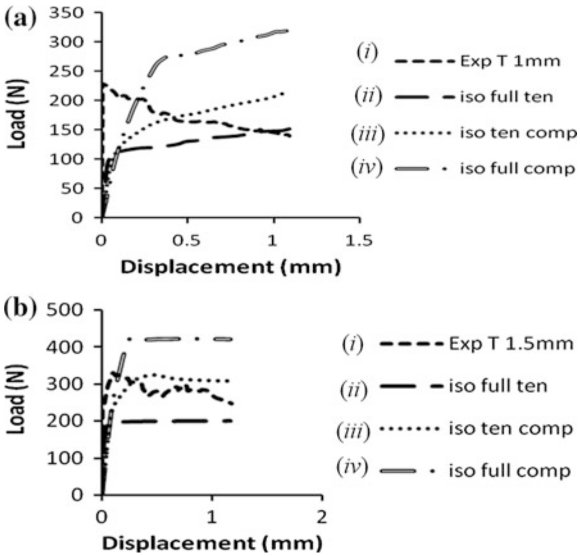


Fig. 7 Comparison between experimental and isotropic finite element load-displacement diagrams for **a** 1.0 mm and **b** 1.5 mm thick small transverse specimen [where (*i*) = experimental curve, (*ii*) = FTTI, (*iii*) = TCTI and (*iv*) = FCTI]

Table 2 Experimental and finite element results of small punch test for longitudinal small specimen

Method	Specimen thickness (mm)	Failure load (N)	Stiffness (kN/mm)	Load at break away (N)
Exp results	1.0	172.26	398.18	151.71
	1.5	425.62	525.09	166.67
FTTI	1.0	136.04	5.04	105.71
	1.5	220.49	7.99	166.67
TCTI	1.0	148.12	1.48	123.57
	1.5	237.64	2.38	211.39
FCTI	1.0	166.29	1.21	94.89
	1.5	253.65	3.45	119.24
FTI	1.0	194.19	6.92	133.69
	1.5	327.41	11.99	249.79
TCI	1.0	192.03	1.46	113.07
	1.5	314.83	2.74	177.02
FCI	1.0	160.05	0.68	146.88
	1.5	241.06	1.32	232.67

however, it was significantly different ($p < 0.05$) form FTTI model result. On the other hand, the stiffness values of experimental curves were found to be significantly different ($p < 0.00000001$) form all finite element results. The experimental load at breakaway point (load point corresponding to 0.2 % offset point) was not observed to be significantly different form finite element results of TCTI, FTI, and FCI while it was observed to be significantly different ($p < 0.01$ FCTI and $p < 0.05$ for FTTI and TCI) for FTTI, FCTI and TCI as compared to experimental result.

Table 3 Experimental and finite element results of small punch test for transverse small specimen

Method	Specimen thickness (mm)	Failure load (N)	Stiffness (kN/mm)	Load at break away (N)
Exp results	1.0	139.06	54.26	227.30
	1.5	248.28	51.40	231.80
FTTI	1.0	121.15	3.03	66.50
	1.5	174.36	7.03	155.20
TCTI	1.0	178.81	2.07	74.19
	1.5	259.34	7.34	79.83
FCTI	1.0	267.25	1.61	61.36
	1.5	366.1	5.02	114.92
FTI	1.0	151.18	2.93	86.72
	1.5	201.30	7.65	169.20
TCI	1.0	216.77	1.81	80.70
	1.5	308.08	2.56	233.89
FCI	1.0	318.98	1.24	101.05
	1.5	122.20	2.65	232.69

For 1.5 mm thick longitudinal specimen, the experimental failure load was found to be significantly different ($p < 0.0001$ for FTTI, TCTI and FCI, $p < 0.001$ for FCTI, $p < 0.005$ for FTI and $p < 0.01$ for TCI) form all finite element results. The stiffness of experimental curve was found to be significantly different ($p < 0.00000001$) form all finite element curves. Similarly, the load at breakaway point of experimental curve was also observed to be significantly different ($p < 0.0001$ for FTTI, FCTI, TCI and $p < 0.001$ for TCTI and FCI, $p < 0.01$ for FTI) form all finite element curves.

In case of transverse 1.0 mm thick specimen, the experimentally obtained failure load was not found to be significantly different from FE results of FTTI and FTI models, however, the experimental failure load was found to be significantly different from FE results of other models ($p < 0.05$ for TCTI, $p < 0.0001$ for FCTI, $p < 0.001$ for TCI and $p < 0.00001$ for FCI). The stiffness and load at breakaway point values obtained from experimental setup were found to be significantly ($p < 0.0001$ in case of stiffness and $p < 0.001$ in case of load at breakaway point) different from FE results of different models.

For transverse 1.5 mm thick specimens, experimentally obtained failure load was found to be significantly different ($p < 0.01$ for FTTI, $p < 0.001$ for FCTI and FCI, $p < 0.05$ for FTI and TCI) from Finite element results obtained from FTTI, FCTI, FTI, TCI, and FCI, however, this load was not observed to be significantly different from FE results of TCTI models. The stiffness value of the experimental load-displacement curve was noticed to be significantly different ($p < 0.000001$) from the stiffness values obtained from all different FE models. For this case the load at breakaway point value obtained from the experimental curve was found to be significantly different ($p < 0.01$ for FTTI and FTI, $p < 0.0001$ for TCTI,

$p < 0.001$ for FCTI) from FE results of FTTI, TCTI, FCTI and FTI models, whereas, this was not observed to be significantly different from FE results of TCI and FCI models.

As per the above given results, the experimental and finite element results were found to be significantly different in case of 1.5 mm thick longitudinal specimen, whereas, for 1.0 mm thick longitudinal specimen the corresponding experimental and FE models results were not significantly different. This shows that thickness of the cortical bone specimen plays an important role during deformation of bone. The hierarchical composite like structure of cortical bone may be the main reason of this deformational behaviour. Further, bone is considered to be highly heterogeneous material due to variation in its compositional parameters such as density, porosity, mineralization etc., however, for FE simulation the material was considered to be homogenous. The mismatch between experimental and FE model results of 1.5 mm thick longitudinal specimen shows that as specimen thickness increases the material becomes more heterogeneous in nature. Therefore, for small specimen testing of longitudinal specimens the FE models of bone specimens having less than 1.5 mm thickness can predict closer results to the experimental one. It is also evident from this study that in case of longitudinal small specimens the load-displacement curve obtained from isotropic fully tensile FE mode (FTI) is closer to the experimental curve. This shows that tensile properties are dominating in case of small punch testing of longitudinal specimens and bone material behaves almost isotropic in nature for this direction of loading.

For transverse specimen testing failure occurs in the form of splitting of lamellar bone. In case of lamellar bone, lamellae are glued together with an extra fibrillar matrix which is composed of non-collagenous proteins [18]. At a small thickness i.e. 1.0 mm of transverse specimens, it may possible that amount of glue is not sufficient to resist the deformation at some higher load and this way the stiffness degradation occurs at a very fast rate for this case. This may be the reason of mismatch of the experimental and simulated curves for transverse specimens of 1.0 mm thickness.

For transverse specimens of 1.5 mm thickness the nature of load-displacement curve obtained from tensile compressive isotropic (TCI) FE model was found to be very similar to that of the experimental load-displacement curve as compared to the corresponding curves obtained from other FE models. This shows that in case of transverse specimen of 1.5 mm thickness the tensile and compressive both properties are dominating and material behaves almost isotropic in nature.

The above discussion shows that deformational behavior of small specimens for cortical bone depends on the thickness of the specimen and direction of loading. For longitudinal specimen, 1.0 mm thickness is good enough to maintain the continuum of the material and this way fully tensile isotropic (FTI) FE model gives good resemblance with the experimental results. However, for transverse specimens the results are totally different. For latter case the minimum thickness required to maintain the continuum is 1.5 mm and tensile-compressive isotropic (TCI) FE model shows good agreement with the experimental results. The important outcome of this study was that the isotropic properties were found to be sufficient in order to simulate the deformational behavior of cortical bone specimens of small size.

4 Conclusion

The deformational behavior of cortical bone was analyzed for two different orientations of bone specimens using small punch testing. This way small punch testing was conducted on longitudinal and transverse small cortical bone specimens of two different thicknesses (1.0 and 1.5 mm). The finite element simulation of small punch test was carried out using ABAQUS software to compare the load-displacement behavior of cortical bone. The deformational behavior of cortical bone under small punch testing was modeled using transversely isotropic as well as isotropic material properties of bone. Various cases have been incorporated in FE simulation of cortical bone after partitioned the small rectangular specimen from the middle. Tensile and compressive properties obtained from full size tests have been fed to the software to perform the FE simulation. For longitudinal specimens, the FE results of FTI model of 1.0 mm thick specimen were found to be in better agreement with the corresponding experimental results. The deviation in FE and experimental load-displacement curve for 1.5 mm thick longitudinal specimen was considered to be due to hierarchical and heterogeneous nature of cortical bone. For transverse specimens, the FE results of TCI model of 1.5 mm thick specimen were observed to be in better agreement with the corresponding experimental results. Therefor for transverse specimens 1.0 mm thickness was not found to be sufficient to maintain the continuum. From this investigation it was observed that isotropic tensile properties are sufficient to better simulate the small punch testing of cortical bone.

References

1. N. Sasaki, N. Matsushima, T. Ikawa, H. Yamamura, A. Fukuda, Orientation of bone mineral and its role in the anisotropic mechanical properties of bone-transversely anisotropy. J. Biomech. **22**, 157–164 (1989)
2. S. Weiner, H.D. Wagner, The material bone: structure-mechanical function relations. Ann. Rev. Mater. Sci. **28**, 271–298 (1998)
3. R.A. Robinson, S.R. Elliot, The water content of bone: I. The mass of water inorganic crystals organic matrix and ‘CO2 space’ components in a unit volume of dog bone. J. Bone Joint Surg. **39A**, 167–188 (1957)
4. R.B. Martin, Porosity and specific surface of bone. CRC Crit. Rev. (1984)
5. E. Lucchinetti, *Composite Models of Bone Properties, Bone Mechanics Handbook*, 2nd edn, chap. 3 (CRC Press, Boca Raton, 2001), pp. 12.1–12.19
6. J. McElhaney, J. Fogle, E. Byars, G. Weaver, Effect of embalming on the mechanical properties of beef bone. J. Appl. Physiol. **19**, 1234–1236 (1964)

7. R.W. McCalden, J.A. McGeough, M.B. Barker, C.M. Courtbrown, Age related-changes in the tensile properties of cortical bone: the relative importance of changes in porosity, mineralization and microstructure. J. Bone Joint Surg. Am. **75A**, 1193–1205 (1993)
8. F.G. Evans, M. Lebow, Regional differences in some of the physical properties of the human femur. J. Appl. Physiol. **3**, 563–572 (1951)
9. D.T. Reilly, A.H. Burstein, V.H. Frankel, The elastic modulus for bone. J. Biomech. **7**, 271–275 (1975)
10. N.K. Sharma, D.K. Sehgal, R.K. Pandey, Studies on locational variation of shear properties in cortical bone with Iosipescu shear test. Appl. Mech. Mater. **148–149**, 276–281 (2012)
11. G.E. Lucas, Review of small specimen test technique for irradiation testing. Metall. Trans. A **21A**, 1105–1119 (1990)
12. J.S. Cheon, I.S. Kim, Initial deformation during small punch testing. J. Test. Eval. **24**, 255–262 (1996)
13. J.M. Baik, J. Kameda, O. Buck, Small punch test evaluation of inter-granular embrittlement of an alloy steel. Scr. Metall. **17**, 1443–1447 (1983)
14. M.P. Manahan, A.S. Argon, O.K. Harling, The development of miniaturized disk bend test for the determination of post irradiation mechanical properties. J. Nucl. Mater. **103–104**, 1545–1550 (1981)
15. F.M. Huang, M.L. Hanilton, G.L. Wire, Bend testing for miniature disk. Nucl. Technol. **57**, 234 (1982)
16. N.K. Sharma, S. Sharma, D.K. Sehgal, R.K. Pandey, Studies on deformational behavior of cortical bone using small punch testing and finite element simulation, in *Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering and Computer Science,* WCECS 2013, pp. 920–924, San Francisco, USA, 23–25 Oct 2013
17. N.K. Sharma, D.K. Sehgal, R.K. Pandey, R. Pal, Finite element simulation of cortical bone under different loading and anisotropic yielding situations, in *Lecture Notes in Engineering and Computer Science: Proceedings of the World Congress on Engineering and Computer Science*, WCECS 2012, vol. 2, pp. 1265–1270, San Francisco, USA, 24–26 Oct 2012
18. H.S. Gupta, W. Wagermaier, G.A. Zickler, D.R.B. Aroush, S.S. Funari, Nanoscale deformation mechanisms in bone. Nano Lett. **5**, 2108–2111 (2005)

Chapter 49
Factors Affecting the Intention to Leave Among White-Collar Employees in Turkey

Ecem Basak, Esin Ekmekci, Yagmur Bayram and Yasemin Bas

Abstract Intention to leave refers to conscious and deliberate willfulness to leave the organization. Job satisfaction and organizational commitment are the two most important factors which play an important role in determining employees' intention to leave their job. The aim of this study is to analyze the effects of job satisfaction, organizational commitment's components (affirmative, continuance, normative), perceived organizational support, and job stress on white-collar employees' intention to leave in Turkey. A structural equation-modelling approach was applied to identify the variables that significantly affect the intention to leave. Using LISREL, data collected from 225 employees were used to test the proposed model. Results indicated that seventy two percentage of white-collar employees' intention to leave is explained by job satisfaction, affective commitment, and normative commitment. Among them, job satisfaction has the strongest effect.

Keywords Affective commitment · Continuance commitment · Intention to leave · Job satisfaction · Job stress · Normative commitment · Perceived organizational support

E. Basak (✉) · E. Ekmekci · Y. Bayram · Y. Bas
Management Faculty, Industrial Engineering Department, Istanbul Technical University,
Macka, 34367 Istanbul, Turkey
e-mail: basake@itu.edu.tr

E. Ekmekci
e-mail: eekmekci@itu.edu.tr

Y. Bayram
e-mail: bayramy@itu.edu.tr

Y. Bas
e-mail: basy@itu.edu.tr

H. K. Kim et al. (eds.), *Transactions on Engineering Technologies*,
DOI: 10.1007/978-94-017-9115-1_49,

1 Introduction

Employee turnover has been an important issue for several different domains because high turnover ratio has significant negative effects on an organization such as increase in the cost of recruiting and training new employees, and decrease in organizational performance, organizational employee continuity, organizational stability, and the sales performance due to the lack of experience of the new employee [1, 2]. Therefore, turnover is an undesirable event in the organizations, because "long-term productivity is affected not only by hiring the best qualified personnel, but keeping them in the organization for long periods of time" [3].

The reasons behind the turnover decision have been investigated for years. The literature review shows that the main factor that affects employees to quit their current jobs is the intention itself, and for this reason many studies focus on the impacts of certain factors on turnover intention [1, 2, 4–6]. The research fields vary; from studies focusing on particular industries such as call centers [1], fast food industry [2] and IT industry [5] to studies focusing on certain positions [6]. There are also some researches which acknowledge the turnover intention to be the main driver for the actual turnover and have findings that sometimes the turnover behavior is driven by desirability and ease of movement more than the intention itself [7]. According to the Theory of Reasoned Action, an individual's behavior is determined by his or her behavioral intention [8]. The more an individual shows intention to perform a particular behavior, the more he or she is expected to act it [8]. For this reason, we should emphasize on the employees' intention to leave.

Intention to leave (ITL) refers to "conscious and deliberate willfulness to leave the organization" [9]. Job satisfaction (JSat) and organizational commitment are the two most important factors which play an important role in determining employees' ITL their job [9]. High levels of JSat and organizational commitment provide unwillingness of quitting work. There are several studies that examine the effect of both factors on ITL. Reference [5] modeled JSat and organizational commitment as the antecedents of ITL, and also examined the effect of JSat on organizational commitment. Information technology professionals' intention to quit their jobs in Turkey was questioned in order to investigate the reasons behind turnover. In the study of Ref. [10], turnover intention of white-collar and blue-collar employees working in a manufacturing firm was examined exploring the relationships among JSat, organizational commitment's components (affective, continuance, normative) and ITL. JSat and three components of organizational commitment were found to be important in determining an employee's ITL. Reference [11] also structured the relationships among JSat, three components of organizational commitment, and ITL as the core of the research model, which was conducted with the hospital employees in Iran.

Apart from the core of the model, there are also some other factors which contribute to the determination of turnover intention. Perceived organizational support (POS) and job stress (JS) have been investigated in many studies.

References [11, 12] studied the POS, whereas JS was examined in the studies of Refs. [4, 6, 13] in order to explore its effect on ITL and its related factors.

This study which is an extended and revised version of Ref. [14] aims to analyze the effects of organizational commitment's components (affective, continuance, normative), JSat, POS, and JS on the ITL of white-collar employees in Turkey.

2 Research Model And Hypotheses

2.1 Organizational Commitment

Organizational Commitment is defined as "the relative strength of an individual's identification with and involvement in a particular organization" [15]. Employees who feel strong commitment to the organizations are less likely to quit their jobs [16]. On the other hand, there are also less-committed employees in the organization. They perceive their current jobs as a temporary employment, and when they get a better opportunity outside the organization, they may have a favorable intention to quit [12]. However, one dimensional commitment is found to be insufficient so that three-component model of organizational commitment is proposed to better identify the psychological state which leads employees to feel committed to the organization [17]. Three components may be defined as follows: affective commitment (AC), which refers to the emotional attachment of an employee to the organization, continuance commitment (CC), which is related to the perceived costs that will be the consequence of leaving the organization, and normative commitment (NC), which refers to the obligation felt by an individual to the organization [17]. Furthermore, the effects of organizational commitment's components on ITL are also examined in other studies [10, 11, 18].

Therefore, we hypothesize as follows:

H1: AC influences ITL negatively.
H2: CC influences ITL negatively.
H3: NC influences ITL negatively.

2.2 Job Satisfaction

Job satisfaction (JSat) is defined as "a pleasurable or positive emotional state resulting from the appraisal of one's job or job experiences" [19]. Employees show positive attitude toward their jobs as they experience that their jobs fulfill values which are important to them [19]. Satisfied employees are more likely to perform their jobs better. The more they are satisfied, the more they are committed and unintended to quit [4]. The low level of JSat causes employees to feel a poor sense of belonging to the organization and search for alternative jobs [20]. Therefore, JSat has an important role in maintaining commitment and influencing ITL [5]. Furthermore, the effect of JSat on organizational commitment's components and ITL is also examined in other studies [11, 18, 21].

Therefore, we hypothesize as follows:

H4: JSat influences AC positively.
H5: JSat influences CC positively.
H6: JSat influences NC positively.
H7: JSat influences ITL negatively.

2.3 Perceived Organizational Support

Perceived Organizational Support (POS) is defined as employees' "global beliefs about the extent to which the organization cares about their well-being and values their contributions" [22]. The more employees perceives the organizational support, the more they feel that they are respected and esteemed in the organization, and expect that their superior performance will be rewarded [23]. This attitude breeds a strong sense of belonging to the organization and puts aside the feelings of entrapment in the organization [24]. Hence, employees with high POS are more likely to devote themselves to their organizations. They perceive firm's successes and failures as their own [25]. Moreover, their overall satisfaction related with their job increases if the organization meets their socio-emotional needs, answers their call for help in an emergency, and rewards their increased performance [24]. Furthermore, the effect of POS on organizational commitment's components and JSat is also examined in other studies [21, 26, 27].

Therefore, we hypothesize as follows:

H8: POS influences AC positively.
H9: POS influences CC positively.
H10: POS influences NC positively.
H11: POS influences JSat positively.

2.4 Job Stress

Job stress (JS) refers to "the harmful physical and emotional responses that occur when the requirements of the job do not match the capabilities, resources, or needs of the worker" [28]. JS cannot be matched with general stress because it occurs as a result of job settings. In job settings, the work task, the work place, the job characteristics, role conflict, or worker capabilities are the possible reasons that play a role in forming stress [29]. Employees usually encounter stress in the workplace because of the excessive demands of organizations for better job outcomes. Some negative effects such as a lack of interest to the job, a lack of concern for the organization, or a loss of responsibility can occur [30]. The experience of job related stress causes a decrease in the JSat of employees. Therefore, there is a negative relationship between JS and JSat [4]. Stressful individuals feel dissatisfied with their jobs, and end up quitting from the organization [31]. Furthermore, the effect of JS on JSat is also examined in other studies [5, 6, 13].

Therefore, we hypothesize as follows:

H12: JS influences JSat negatively.

3 Methodology

A survey methodology was used in this study to gather data. Target population is the white-collar employees who are working in different companies. The questionnaire was formed by two main parts. The first part consisted of demographic questions designed to solicit information about gender, age, working industry, working position, full-time professional experience, full-time working experience in the current firm. The 225 questionnaires were collected from different companies. The 53.33 percentage of respondents were male, and the average age of respondents was 30.22 years. The summary of demographic profiles of respondents is given in Table 1.

The second part consisted of the items measuring behavioral ITL [32], organizational commitment [33], JSat [34], POS [22], and JS [4]. The items for the constructs and their corresponding sources can be seen in the Appendix A. A five-point Likert-scale type was used to measure all these items. In a five-point Likert-scale type, one represents "*strongly disagree*" and five represents "*strongly agree*".

4 Results

The model was tested using LISREL 8.80 [35] with LISREL project.

4.1 Measurement Model

The measurement model included 44 items, describing seven constructs: ITL, JSat, AC, CC, NC, POS, JS. According to the results of reliability analysis, the items whose factor loadings are lower than 0.5 and the items with excessive standard errors were dropped from the model one by one. A total of 7 items were dropped from the model.

After re-analyzing the measurement model, reliability and validity of the model is examined by factor loadings, t-values, composite reliability (CR), and average variance extracted (AVE). The results can be seen in Tables 2, 3, and 4. The factor loadings which are higher than 0.50 are considered practically significant [36].

Table 1 Demographic profiles of the respondents

Gender (%)		
Female: 46.67	Male: 53.33	
Age (years)		
Max: 55	Min: 22	
Educational status (%)		
Graduate: 67.11	Post-graduate: 32.09	
Full-time professional experience (%)		
<6 months: 2.22	6 months–1 year: 2.66	1–3 years: 24.44
3–5 years: 22.67	5–10 years: 23.56	>10 years: 24.44
Work experience in the current company (%)		
<6 months: 2.22	6 months–1 year: 11.56	1–3 years: 36.00
3–5 years: 18.67	5–10 years: 16.44	>10 years: 10.22
Department (%)		
IT: 18.67	Marketing/Sales: 9.78	Financial services: 8.00
Production: 5.78	R&D: 5.33	Project management: 4.89
Accounting: 3.11	Engineering: 3.11	Consultancy: 2.67
ERP: 2.67	Quality/Control: 2.67	Others: 24.01
Position (%)		
Specialist: 27.55	Department manager: 14.22	Engineer: 13.77
Supervisor: 7.55	Assistant specialist: 4.88	Director: 4.88
Job analyst: 4.88	Project manager: 4.44	Consultant: 3.11
Others: 14.22		

Retained items in the measurement model are reasonably explained by related factors. In order to evaluate the reliabilities of the constructs, CR and AVE are examined. According to the results, each construct has a greater CR than the recommended value of 0.70 [37]. The constructs except NC and ITL have also greater AVE values than the recommended value of 0.50 [37]. However, AVE values of NC (0.48) and ITL (0.49) are close enough to 0.50. Therefore, each construct indicates an acceptable level of reliability.

4.2 Structural Model

The relationships between constructs are indicated in the structural model [33]. As seen in Table 5, Chi-square to degrees of freedom ratio at 2.32, Root Mean Squared Error of Approximation (RMSEA) at 0.077, Comparative Fit Index (CFI) at 0.96, Normed Fit Index (NFI) at 0.93 are all within the recommended values [36, 38, 39]. This implies that the model provides a reasonably good fit to the data.

The model explains substantial variance in ITL ($R^2 = 0.67$), AC ($R^2 = 0.53$), CC ($R^2 = 0.05$), NC ($R^2 = 0.50$), and JSat (R2 = 0.26). Additionally, all the hypotheses are supported except H2 (CC-ITL), H5 (JSat-CC), H6 (JSat-NC), and

Table 2 Factor loadings of the items

Code	λ	Code	λ	Code	λ	Code	λ	Code	λ
ITL01	0.58	CC02	0.79	JSat01	0.78	POS01	0.83	JS01	0.74
ITL02	0.61	CC03	0.79	JSat02	0.80	POS02	0.84	JS02	0.81
AC01	0.60	CC04	0.62	JSat03	0.85	POS03	0.81	JS03	0.82
AC02	0.55	NC02	0.60	JSat04	0.77	POS04	0.88	JS04	0.83
AC03	0.89	NC03	0.78	JSat05	0.80	POS05	0.69	JS05	0.74
AC04	0.88	NC04	0.71			POS06	0.73	JS06	0.72
AC05	0.80	NC05	0.75			POS07	0.67	JS07	0.76
AC06	0.63	NC06	0.59			POS08	0.73	JS08	0.65

Table 3 T-values of the items

Code	t	Code	t	Code	t	Code	t	Code	t
ITL01		CC02		JSat01		POS01	15.07	JS01	12.72
ITL02	8.57	CC03	8.77	JSat02	18.97	POS02	15.24	JS02	14.32
AC01		CC04	8.05	JSat03	20.45	POS03	14.51	JS03	14.77
AC02	7.63	NC02		JSat04	14.80	POS04	16.58	JS04	14.90
AC03	11.63	NC03	8.77	JSat05	17.22	POS05	11.52	JS05	12.59
AC04	11.46	NC04	8.18			POS06	12.43	JS06	12.09
AC05	10.64	NC05	8.48			POS07	11.11	JS07	13.05
AC06	8.61	NC06	7.18			POS08	12.55	JS08	10.55

Table 4 Reliability analysis

Latent variable	CR	AVE
Intention to leave	0.71	0.48
Affective commitment	0.88	0.53
Continuance commitment	0.78	0.55
Normative commitment	0.82	0.49
Job satisfaction	0.94	0.76
Perceived organizational support	0.92	0.60
Job Stress	0.92	0.58

H9 (POS-CC). The results show that ITL is explained by JSat, AC, and NC. However, their standardized estimates are different from each other so that their explanatory power is not the same. Appendix B shows the standardized path coefficients and the explanatory power of each construct.

Among constructs, JSat is found to be the main predictor of ITL. Subsequently, JSat is explained by both POS and JS. POS is a relatively stronger determinant of JSat. Similarly, POS also has a high effect on AC, whereas JSat has a low effect. Finally, according to the results, NC is only explained by POS, and the effects of POS and JSat on continuance commitment are found to be insignificant to explain it. Table 6 shows the direct, indirect, and total effects of each construct on the ITL. As shown in Table 6, JSat has the highest direct and total effect on ITL.

Table 5 Fit indices

Fit index	Recommended value	Observed value
$\chi 2$/df ($\chi 2$; df)	≤ 3	2.32 (1433; 616)
RMSEA	≤ 0.08	0.077
CFI	≥ 0.90	0.96
NFI	≥ 0.90	0.93

Table 6 Direct, indirect and total effects on intention to leave

Dependent variable	Independent variables	Direct effects	Indirect effects	Total effects
Intention to leave	Job satisfaction	−0.67	−0.154	−0.827
	Affective commitment	−0.27	–	−0.27
	Normative commitment	−0.37	–	−0.37
	Continuance commitment	–	–	0
	Perceived organizational support	–	−0.69	−0.69
	Job stress	–	0.87	0.87

5 Conclusion and Future Work

This study analyzes the effects of JSat, AC, CC, NC, POS, and JS in determining the white-collar employees' ITL their job. For this, 225 questionnaires have been collected from white-collar employees working in the different firms.

The explanation rate of ITL which is 0.67, is relatively high in this study compared to the other studies [4, 5, 11]. The results also show that ITL is explained by JSat, AC and NC. It is found that JSat is the most important antecedent of ITL. Consistent with the findings of Ref. [10], JSat has stronger effect than other factors on white-collar employees' ITL their jobs. The results imply that the satisfied employees will be less likely to quit their jobs. However, inconsistent with the findings of Ref. [10, 11], the effect of CC is found to be insignificant whereas AC and NC have significant effects on ITL.

Another result of this study is that AC is explained by POS and JSat, whereas NC is only explained by POS. However, both of them is found to be insignificant to explain CC. According to the results of the analysis, POS plays a more effective role than JSat in explaining the AC. These results are partially supported by the findings of Ref. [21, 26]. In these studies, all three components of organizational commitment are explained by POS and JSat. However, similar to our findings, perceived organizational has the highest influence on AC compared to the other factors.

The other result indicates that JSat is explained by POS and JS. Of the two, POS has a higher impact on JSat, but its effect is found to be relatively low compared to Ref. [21]. On the other hand, in the study of Ref. [13] JS' effect on JSat is also found to be low.

Although the findings of this study provide a better understanding of the factors affecting ITL among white-collar employees in Turkey, the effects of demographic attributes were not analyzed in this study. A similar study including demographic characteristics, such as full-time experience, age, gender of the respondents, department, and position may be a subject for future research. Moreover, by increasing the size of the data collected, group differences among engineer, manager, analyst, and consultant, specialist, and director, also among different departments such as IT, production, marketing, and accounting analyzed as a further study.

Finally, the results of the current study may be shared with the managers to have a further understanding about the findings. Firms in Turkey should consider the findings of the study for the continuity of their employees in their organizations.

Appendix A

Latent variable	Reference	items
Intention to quit	[32]	If I have a good opportunity, I would like to find another job
		I do not enjoy this job and have been searching for other positions
		*I hope that I can find another job in the same industry
		*Layoffs are a typical occurrence around here
		*People often get fired from this organization without good reason
Organizational commitment	[33]	I would be very happy to spend the rest of my career with this organization
		I really feel as if this organization's problems are my own
		I do not feel a strong sense of "belonging" to my organization (R)
		I do not feel "emotionally attached" to this organization (R)
		I do not feel like "part of the family" at my organization (R)
		This organization has a great deal of personnel meaning for me
		*Right now, staying with my organization is a matter of necessity as much as desire
		It would be very hard for me to leave my organization right now, even I wanted to
		Too much of my life would be disrupted if I decided I wanted to leave my organization right now
		*I feel that I have too few options to consider leaving this organization
		*If I had not already put so much of myself into this organization, I might consider working elsewhere
		*One of the few negative consequences of leaving this organization would be the scarcity of available alternatives
		*I do not feel any obligation to remain with my current employer (R)
		Even if it were to my advantage, I do not feel it would be right to leave my organization now
		I would feel guilty if I left my organization
		This organization deserves my loyalty
		I would not leave my organization right now because I have a sense of obligation to the people in it
		I owe a great deal to my organization

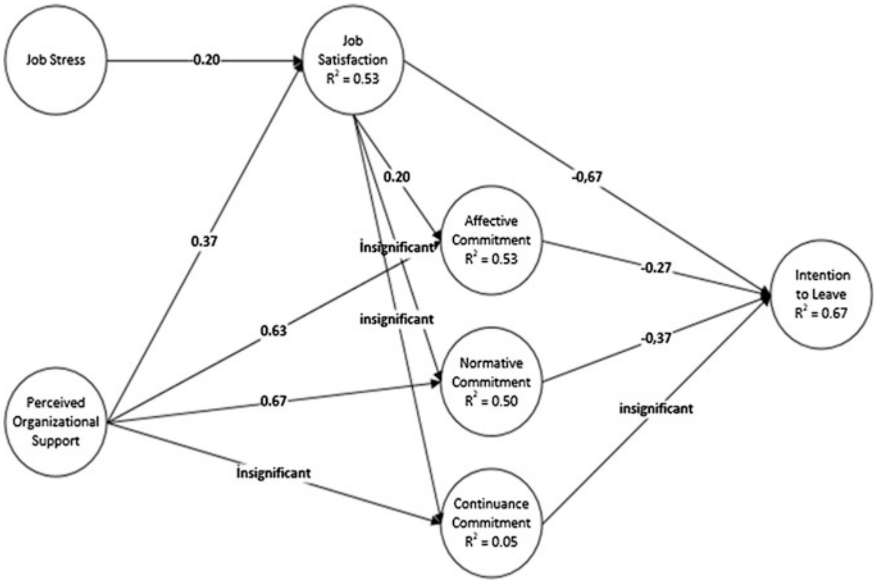
(continued)

(continued)

Latent variable	Reference	items
Job satisfaction	[34]	I consider my job pleasant I feel fairly-well satisfied with my present job I definitely like my work My job is pretty interesting I find real enjoyment in my work
Job stress	[4]	I feel emotionally drained by my job I feel burned-out by my job I feel frustrated at my job I feel tense at my job I lose my appetite because of my jobrelated problems Job-related problems keep me awake at night Job-related problems make my stomach upset
Perceived organizational support	[22]	The organization values my contribution to its well-being The organization fails to appreciate any extra effort from me (R) The organization disregards my best interests when it makes decisions that affect me (R) The organization really cares about my well-being Even if I did the best job possible, the organization would fail to notice (R) The organization cares about my general satisfaction at work The organization shows very little concern for me (R) The organization takes pride in my accomplishments at work

(*) items extracted for structural model analysis
(R) items reverse-coded

Appendix B


Job Stress
0.20
Job Satisfaction $R^2 = 0.53$
-0,67
0.20
Affective Commitment $R^2 = 0.53$
insignificant
0.37
-0.27
Intention to Leave $R^2 = 0.67$
0.63
insignificant
Normative Commitment $R^2 = 0.50$
-0,37
0.67
Perceived Organizational Support
insignificant
Insignificant
Continuance Commitment $R^2 = 0.05$

References

1. Z.M.B. Siong, D. Mellor, K.A. Moore, L. Firth, Predicting intention to quit in the call centre industry: does the retail model fit? J. Manag. Psychol. **21**(3), 231–243 (2006)
2. R. Kumar, C. Ramendran, P. Yacob, A study on turnover intention in fast food industry: Employees fit to the organizational culture and the important of their commitment. Int. J. **2**(5), 9–42 (2012)
3. R. H. Rasch, An investigation of factors that impact behavioral outcomes of software engineers, in *Proceedings of the SIGCPR, SIGCPR* (New York, USA, 1991), pp. 38–53
4. L. Firth, D.J. Mellor, K.A. Moore, C. Loquet, How can managers reduce employee intention to quit? J. Manag. Psychol. **19**(2), 170–187 (2004)
5. F. Calisir, C.A. Gumussoy, I. Iskin, Factors affecting intention to quit among it professionals in turkey. Pers. Rev. **40**(4), 514–533 (2011)
6. C.A. Veloutsou, G.G. Panigyrakis, Consumer brand managers' job stress, job satisfaction, perceived performance and intention to leave. J. Mark. Manage. **20**(1–2), 105–131 (2004)
7. L. Stanley, C. Vandenberghe, R. Vandenberg, K. Bentein, Commitment profiles and employee turnover. J. Vocat. Behav. **82**, 176–187 (2013)
8. I. Ajzen, T.J. Madden, Prediction of goal-directed behavior: Attitudes, intentions, and perceived behavioral control. J. Exp. Soc. Psychol. **22**(5), 453–474 (1986)
9. R.P. Tett, J.P. Meyer, Job satisfaction, organizational commitment, turnover intention, and turnover: path analyses based on metaanalytic findings. Pers. Psychol. **46**(2), 259–293 (1993)
10. I. Yucel, Examining the relationships among job satisfaction, organizational commitment, and turnover intention: an empirical study. Int. J. Bus. Manage. **7**(20), 44 (2012)
11. A.M. Mosadeghrad, E. Ferlie, D. Rosenberg, A study of the relationship between job satisfaction, organizational commitment and turnover intention among hospital employees. Health Serv. Manage. Res. **21**(4), 211–227 (2008)
12. A. Yahaya, N. Yahaya, K. Arshad, J. Ismail, S. Jaalam, Occupational stress and its effects towards the organization management. J. Soc. Sci. **5**(4), 390–397 (2009)
13. A. Elangovan, Causal ordering of stress, satisfaction and commitment, and intention to quit: a structural equations analysis. Leadersh. Organ. Dev. J. **22**(4), 159–165 (2001)
14. E. Basak, E. Ekmekci, Y. Bayram, Y. Bas, Analysis of factors that affect the intention to leave of white-collar employees in Turkey using structural equation modelling, in *Proceedings of the World Congress on Engineering and Computer Science 2013, WCECS*. Lecture Notes in Engineering and Computer Science (San Francisco, USA, 2013), pp. 1058–1063, 23–25 Oct 2013
15. R.T. Mowday, L.W. Porter, R.M. Steers, Employee-organization linkages: the psychology of commitment, absenteeism, and turnover. Am. J. Sociol. **88**(6), 1315–1317 (1983)
16. J.P. Meyer, D.J. Stanley, L. Herscovitch, L. Topolnytsky, Affective, continuance, and normative commitment to the organization: A meta-analysis of antecedents, correlates, and consequences. J. Vocat. Behav. **61**(1), 20–52 (2002)
17. N.J. Allen, J.P. Meyer, The measurement and antecedents of affective, continuance and normative commitment to the organization. J. Occup. Psychol. **63**(1), 1–18 (1990)
18. S. Aydogdu, B. Asikgil, An empirical study of the relationship among job satisfaction, organizational commitment and turnover intention. Int. Rev Manage. Mark. **1**(3), 43–53 (2011)
19. E.A. Locke, in *The nature and causes of job satisfaction.* ed. by M.D. Dunnette. Handbook of Industrial and Organizational Psychology, (Rand, Chicago, 1976)
20. S.A. Reed, S.H. Kratchman, R.H. Strawser, Job satisfaction, organizational commitment, and turnover intentions of United States accountants: The impact of locus of control and gender. Acc. Auditing Acc. J. **7**(1), 31–58 (1994)
21. U. Colakoglu, O. Culha, H. Atay, "The effects of perceived organizational support on employees affective outcomes: Evidence from the hotel industry. Tourism Hospitality Manag. **16**(2), 125–150 (2010)

22. R. Eisenberger, R. Huntington, Perceived organizational support. J. Appl. Psychol. **71**(3), 500–507 (1986)
23. R. Eisenberger, J. Cummings, S. Armeli, P. Lynch, Perceived organizational support, discretionary treatment, and job satisfaction. J. Appl. Psychol. **82**(5), p812 (1997)
24. L. Rhoades, R. Eisenberger, Perceived organizational support: a review of the literature. J. Appl. Psychol. **87**(4), 698–714 (2002)
25. R. Loi, N. Hang-Yue, S. Foley, Linking employees' justice perceptions to organizational commitment and intention to leave: The mediating role of perceived organizational support. J. Occup. Organ. Psychol **79**(1), 101–120 (2006)
26. D. Ucar, A.B. Ötken, Perceived organizational support and organizational commitment: The mediating role of organization based self-esteem. Dokuz Eylul Univ. Fac. Econ. Adm. Sci. J. **25**(2), 85–105 (2010)
27. V. LaMastro, Commitment and perceived organizational support. Natl. Forum Appl. Edu. Res. J. **13**(3), 1–13 (2000)
28. NIOSH, Stress at Work. NIOSH Publication, 1999
29. R.C. Jou, C.W. Kuo, M.L. Tang, A study of job stress and turnover tendency among air traffic controllers: The mediating effects of job satisfaction. Transp. Res. Part E **57**, 95–104 (2013)
30. D. Pathak, Role of perceived organizational support on stress satisfaction relationship: An empirical study. Asian J. Manage. Res. **3**(1), 153–177 (2012)
31. P. Paille, Perceived stressful work, citizenship behaviour and intention to leave the organization in a high turnover environment: Examining the mediating role of job satisfaction. J. Manage. Res. **3**(1), 2010
32. M.A. Lahey, Job *Security: Its Meaning and Measure*. (Kansas State University, 1984)
33. J.P. Meyer, N.J. Allen, C.A. Smith, Commitment to organizations and occupations: Extension and test of a three-component conceptualization. J. Appl. Psychol. **78**(4), p538 (1993)
34. D.J. Schleicher, J.D. Watt, G.J. Greguras, Reexamining the job satisfaction-performance relationship: The complexity of attitudes. J. Appl. Psychol. **89**(1), 165–177 (2004)
35. K. Joreskog, D. Sorbom, LISREL 8.80. Scientific Software International, 2006
36. J.F. Hair, W.C. Black, B.J. Babin, R.E. Anderson, R.L. Tatham, *Multivariate data analysis*, 7th edn. (Prentice Hall Upper Saddle River, NJ, 2010)
37. C. Fornell, D.F. Larcker, Evaluating structural equation models with unobservable variables and measurement error. J. Marketing Res. 39–50, 1981
38. E.K. Kelloway, *Using LISREL for Structural Equation Modeling: a Researcher's Guide* (SAGE Publications, Incorporated, 1998)
39. K. Schermelleh-Engel, H. Moosbrugger, H.M̈uller, Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. Methods Psychol. Res. Online, **8**(2), 23–74 (2003)

Chapter 50
Risk Appraisal in Engineering Infrastructure Projects: Examination of Project Risks Using Probabilistic Analysis

Jane Lai, Lihai Zhang, Colin Duffield and Lu Aye

Abstract Understanding the significant implication of uncertainty is an important step in infrastructure project appraisals. A detailed discussion and application of a risk-based, cost-benefit analytical framework with focus on the analysis of likelihood of risks is presented in this chapter. Three risk analysis tools (i.e. Monte Carlo simulation, Latin Hypercube sampling, and engineering reliability analysis) are presented and compared based on their efficiency and accuracy. Likelihood of risk was represented by a project's probability of investment loss. The framework was applied to a residential property in Melbourne, Australia, with house price as an uncertain variable. It was shown that engineering reliability analysis was the most accurate and efficient in calculating a probability of loss in a 3-year investment duration. In addition, Latin Hypercube sampling, requiring 50 to 100 iterations for convergence, was superior to Monte Carlo simulation which needed 500 to 1000 iterations. Finally, an integrated model is presented to estimate the project risk in term of expected loss.

Keywords Engineering reliability analysis · Expected loss · Latin hypercube sampling · Monte Carlo simulation · Probability of loss · Risk analysis

J. Lai (✉) · L. Zhang · C. Duffield · L. Aye
Department of Infrastructure Engineering, The University of Melbourne,
Parkville, VIC 3010, Australia
e-mail: yklai@unimelb.edu.au

L. Zhang
e-mail: lihzhang@unimelb.edu.au

C. Duffield
e-mail: colinfd@unimelb.edu.au

L. Aye
e-mail: lua@unimelb.edu.au

H. K. Kim et al. (eds.), *Transactions on Engineering Technologies*,
DOI: 10.1007/978-94-017-9115-1_50,

1 Introduction

Project appraisals assess the feasibility of specific options and forecast outcomes by considering aspects of economic, environmental and social attributes. Investment decisions may rest on a few key variables and assumptions. Allowance for uncertainty is a significant concern, particularly in long term projects. Uncertainty may be non-cognitive or cognitive [1] and impact on project elements including scope, cost and quality [2]. Despite comprehensive planning, materialization of some risks is still often unforeseeable [3], however their impact on project goals can be minimized if managed properly [4]. In large infrastructure projects, like public private partnerships, risks are examined and attempts made to identify potential cost impact based on likelihood and allowance made for event occurrence. Risk assessments are often simplistic and do not consider the full description of risk.

The study presented in this chapter discussed and compared methods to quantify the probability of failure. Many of these techniques are well understood mathematically but rarely applied in practice in the area of engineering project risk assessment. The techniques were discussed as part of a proposed framework, as illustrated in Fig. 1. They were applied and tested on a residential property in Melbourne, Australia. The three main risk analysis techniques considered are Monte Carlo simulation (MCS), Latin Hypercube sampling (LHS), and engineering reliability analysis (ERA). These are integrated with the commonly adopted cost benefit analysis (CBA).

The framework is outlined in more detail in Fig. 2. Step 1 establishes the project scope, followed by identifying and collecting variables related to project benefit and cost in Step 2. The variables are forecasted for the specified timeframe in Step 3, which are discounted back to present values. Step 4 evaluates uncertainty by employing one of the three risk analysis techniques: MCS, LHS and ERA. The project is evaluated in Step 5 by accounting for likelihood and consequence of risk. Alternative scenarios are compared where Step 4 is repeated to generate different sets of outcomes.

2 Basic Benefit Cost Model for Infrastructure Project Appraisal

2.1 Measuring Feasibility with Cost Benefit Analysis

CBA is commonly applied to appraise the feasibility of infrastructure projects by examining relevant whole of life financial impacts. Projects with forecasted life cycle variables employ net present value (NPV), which is the sum of discounted future cash flows as shown in Eq. (1).

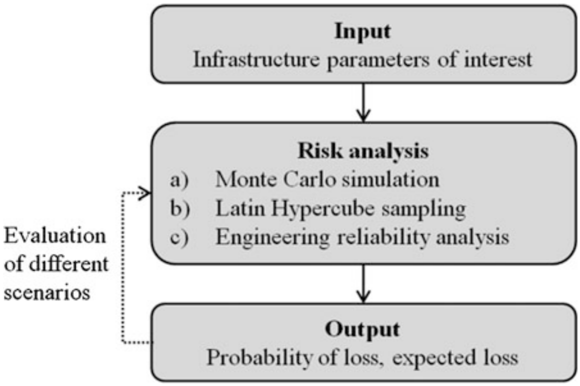

Input
Infrastructure parameters of interest
Risk analysis
a) Monte Carlo simulation
b) Latin Hypercube sampling
c) Engineering reliability analysis
Evaluation of different scenarios
Output
Probability of loss, expected loss

Fig. 1 Overall framework of proposed risk model

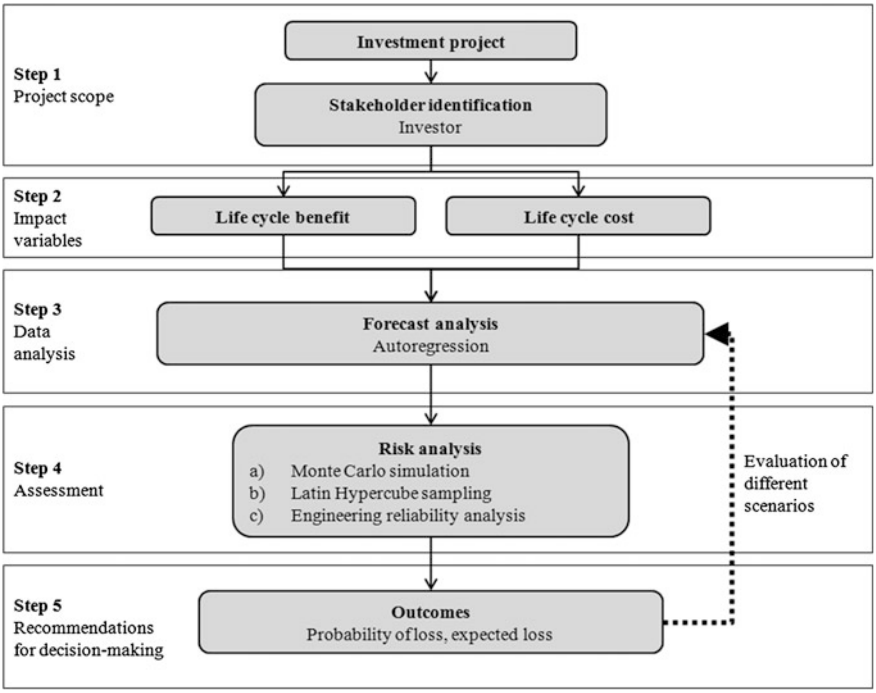

Step 1
Project scope
Investment project
Stakeholder identification
Investor
Step 2
Impact variables
Life cycle benefit
Life cycle cost
Step 3
Data analysis
Forecast analysis
Autoregression
Step 4
Assessment
Risk analysis
a) Monte Carlo simulation
b) Latin Hypercube sampling
c) Engineering reliability analysis
Evaluation of different scenarios
Step 5
Recommendations for decision-making
Outcomes
Probability of loss, expected loss

Fig. 2 Detailed framework for risk management

$$\mathrm{NPV} = \mathbf{B} - \mathbf{C} \tag{1}$$

where **B** is the present value of all benefit, $\Sigma \mathrm{B_j}(1 + \mathrm{i})^{-\mathrm{t}}$; **C** is the present value of all cost, $\Sigma \mathrm{C_j}(1 + \mathrm{k})^{-\mathrm{t}}$; t is time of the cash flow in years; i is discount rate for benefit, k is discount rate for cost; B_j are benefit variables, where j = 1, 2, 3, …; and C_j are cost variables, where j = 1, 2, 3, … $\mathbf{B} > \mathbf{C}$ or NPV > 0 indicates a project is financially feasible. Project options are often prioritized using this approach.

2.2 Uncertainty Analysis

2.2.1 Project Risk

Uncertainty affects the ability, timing and extent of a project's capacity in achieving its objectives [5]. The impact of uncertainty on objectives is risk [5]. Risk is thus present when the decision maker is unsure of the final outcome [6]. The magnitude of risk is the combination of its consequence and its associated likelihood [5, 7, 8], as shown in Eq. (2), where i is a risk event, R is risk, H is likelihood, and Q is consequence. Q may be a project's NPV or risk premium depending on the scope of risk analysis.

$$R_i = f(H_i, Q_i) \tag{2}$$

2.2.2 Assessing Risk

This study aims to model H_i from Eq. (2), represented by probability of loss (P_f). It occurs when cost exceeds benefit or when NPV is less than zero, as shown in Eq. (3)

$$P(\mathbf{B} < \mathbf{C}) = P(\mathrm{NPV} < 0) \tag{3}$$

In comparing the three methods employed to model P_f, the efficiency and accuracy are examined. It is noted that non-monetary variables may have an impact on project decision. These include social and environmental factors that may influence an investor to proceed despite NPV < 0 based on non-financial growth. However, this is outside the scope of this study.

3 Modeling Risk

3.1 Introduction

Simulation creates a range of possible values of variables based on a distribution. The solution generated tend towards the theoretical solution with convergence after large number of iterations [9]. MCS is commonly applied for complex risk management problems [10] with multiple variables of significant uncertainties [7]. In projects with a large number of uncertain variables or complex deterministic systems, it may require much more trial runs to attain an acceptable level of theoretical accuracy, yet the accuracy of the input data may not be commensurate with the analysis. Thus, it may be time-consuming and impractical in some cases [1]. LHS was first introduced by McKay, Conover and Beckman [11] and uses stratified sampling to reduce iterations. Furthermore, MCS might lead to clustering as there is no assurance of sampling the entire sample space of the variable [12]. In contrast, LHS samples are forced to be taken from specified regions [9]. ERA is a risk-based concept that allows uncertainties to be included in the design framework. It does not require simulation if the disparity of variables are known or approximated, and thus can provide a solution almost instantaneously. The following sections briefly discuss the background of MCS, LHS and ERA for completeness.

3.2 Monte Carlo Simulation

MCS involves sampling random variables with multiple iterations according to the random probabilistic characteristic of interest [13]. It utilizes input variables' probability functions to produce distribution function for its output [14]. The outcome is a probabilistic representation of expected value and range of the combined effect of various risks [7]. Projects typically have multiple key inputs with each exhibiting uncertainty simultaneously. These inputs are analyzed based on their distributions instead of point estimates. Iterations are repeated for a large number of times to produce a possible set of random values for the inputs, and thus its outputs.

The $\mathrm{P_f}$ of a project using MCS is calculated with Eq. (4), where $\mathrm{N_{loss,MCS}}$ is the total number of loss, and $\mathrm{N_{total,MCS}}$ is the number of iterations.

$$P_{f,MCS} = \frac{N_{loss,MCS}}{N_{total,MCS}} \tag{4}$$

Complex probabilistic problems commonly employ MCS as it is relatively straight forward to implement and able to generate a range of output values to draw different scenarios. However, as sampling occurs in a random manner, the number of iterations required for accurate representation is high. For complex models, computational time can be extensive.

3.3 Latin Hypercube Sampling

LHS is a simulation method in which a variable is divided into arbitrary intervals or strata of equal probability [12] to ensure all components of the variable is represented [11]. Let $z(\mathbf{X}) = z(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_i)$ where $\mathbf{X}$ is the vector of variables, and X_k is an input variable where $k = 1, 2, \ldots, K$, in $\mathbf{X}_i$. To ensure the whole sample space, S, of variable X_k is represented, each dimension of the input distribution of X_k is divided into equal probability $1/n$ of n strata on the cumulative curve. Within every dimension, only one sample is drawn randomly from each n strata forming n^K number of cells that span across S. Suppose the sample is $X_{k,p}$ where $p = 1, 2, \ldots, n$, it forms the components of X_k. The procedure is repeated for each of the X_i and the components of the X_i's are randomly matched.

The P_f of a project using LHS is calculated with Eq. (5), where $N_{loss,LHS}$ is the total number of loss, and $N_{total,LHS}$ is the number of iterations.

$$P_{f,LHS} = \frac{N_{loss,LHS}}{N_{total,LHS}} \tag{5}$$

LHS is sampling without replacement and avoids clustering of random sampled values that may lead to misrepresentation of distribution. It also allows tail-end samples to be analyzed by forcing simulation to sample from the entire distribution, where this may be overlooked in MCS. LHS is especially useful for small number of iterations or for complex models in which a large number of iterations is not feasible. Whilst LHS is an improvement compared to MCS, it is however still an iterative method and requires computational time.

3.4 Engineering Reliability Analysis

ERA summarizes probability of events created by computationally intensive models [15]. It has traditionally been used in structural analyses involving resistance and loading [1, 16]. In financial analysis structural resistance and loading are analogous to **B** and **C** respectively. The variables are assumed to be statistically independent normally distributed random variables characterized by their means μ, standard deviations σ, and probability density functions $f_{\mathbf{B}}(b)$ and $f_{\mathbf{C}}(c)$, as shown in Fig. 3. For projects to be feasible, yield conditions require $\mathbf{B}_N > \mathbf{C}_N$, where N represents the nominal values. The shaded area is P_f, which can be represented by

$$P_{f,ERA} = P(\text{loss}) = P(\mathbf{B} < \mathbf{C}) = P(NPV < 0) = \int_0^{\infty} F_{\mathbf{B}}(c) f_{\mathbf{C}}(c) dc \tag{6}$$

where $F_{\mathbf{B}}(c)$ is the cumulative distribution function of **B** at 'c'. The limit state or failure surface is where $NPV = 0$.

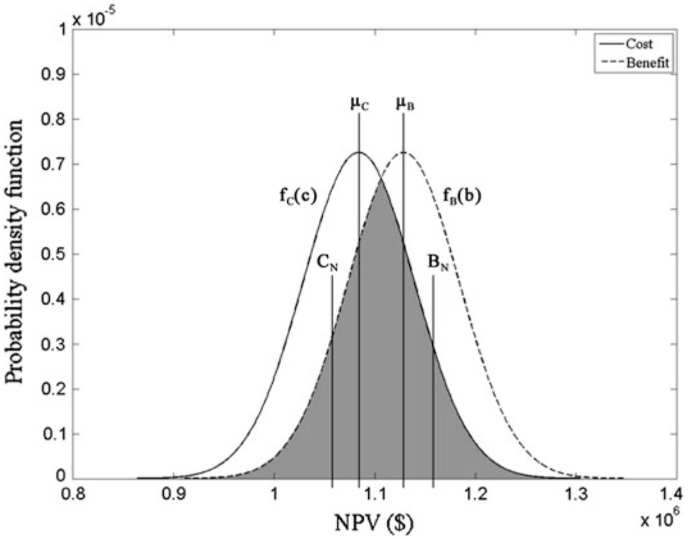


Fig. 3 Concept of ERA. The shaded region represents P_f

Using the means and standard deviations of variables, first-order reliability methods (FORM) is a method to obtain analytical approximations of the integral [1, 15] in Eq. (6). The first-order second-moment (FOSM), a type of FORM, and its derivation can be pursued in its application on a desalination plant case study [17].

3.4.1 Hasofer-Lind Method

The Hasofer-Lind method, or advanced first-order second-moment (AFOSM), is an expansion to FOSM. An advantage of using AFOSM over FOSM is that FOSM does not consider distributional information of variables when it is known and is also less accurate with non-linear performance functions [1]. Further, FOSM exhibits an invariance problem whereby generating different β when safety margins are mathematically equivalent [1, 16]. As AFOSM calculation is dimensionless, it also has the potential to evaluate non-monetary variables such as social and environmental factors. The employment of AFOSM to infrastructure financial appraisals was discussed by Lai et al. [18, 19].

Figure 4 illustrates the AFOSM system with intercepts $\left[-\left(\frac{\mu_B-\mu_C}{\sigma_B}\right), 0\right]$ and $\left[0, \frac{\mu_B-\mu_C}{\sigma_C}\right]$. Variables are standardized to $X_i' = \frac{X_i-\mu_{X_i}}{\sigma_{X_i}}$ in order to evaluate the failure surface in a reduced system, where $X_i' \sim N(0,1)$. The design point is the most probable point of loss and occurs when the planar of performance function is 0 (i.e. $\mathbf{B} = \mathbf{C}$) along the failure surface or limit state. It represents the minimum distance between the reduced failure surface and origin, and is the point of minimum reliability [20] or most probable failure [1]. As the reduced failure surface moves closer to the origin, the loss region is larger. The minimum distance is the Hasofer-Lind method reliability index, β_{HL}, as shown in Eq. (7), where * denotes the design point $x_i'^{*} = -\alpha_i \beta_{HL}$, and α is the directional cosines of the reduced coordinate X_i' in Eq. (8)

$$\beta_{HL} = \sqrt{(\boldsymbol{x}'^{*})^t(\boldsymbol{x}'^{*})} = -\frac{\sum_{i=1}^{n} x_i'^{*}\left(\frac{\partial g}{\partial X_i'}\right)^{*}}{\left[\sum_{i=1}^{n}\left(\frac{\partial g}{\partial X_i'}\right)^{2*}\right]^{\frac{1}{2}}} \tag{7}$$

$$\propto_i = \frac{\left(\frac{\partial g}{\partial X_i'}\right)^{*}}{\left[\sum_{i=1}^{n}\left(\frac{\partial g}{\partial X_i'}\right)^{2*}\right]^{\frac{1}{2}}} \tag{8}$$

$P_{f,ERA}$ can be rewritten as

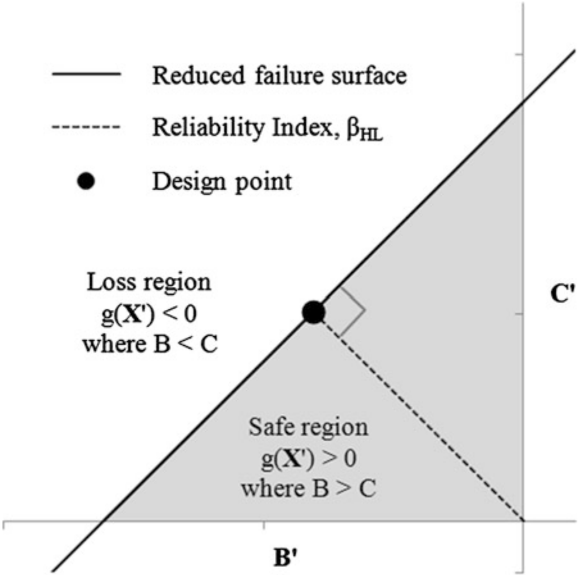
$$p_{f,ERA} = \Phi(-\beta_{HL}) = 1 - \Phi(\beta_{HL}) \tag{9}$$

where Φ is cumulative distribution function of the standard normal variate. ERA is not an iterative technique and does not require extensive computational time to obtain P_f. It is useful for projects with large reliable historical database or expert opinion based on past experience which can infer input distributional characteristics. Without accurate estimations of key parameters, P_f should only be used as an indicator.

3.5 Results and Discussion

The framework is applied to a one-storey urban residential building located in Melbourne, Australia, with a floor area of 222.4 m [2, 21]. The building has two bathrooms and a kitchen with additional data sources summarized in the Appendix. The uncertain variable is house price and has been forecasted with autoregression using historical data. The benefit and cost variables included are shown in Eqs. (10) and (11) where * and † specify the variable at year of purchase and at year of sale respectively. Selling price, SP, is determined by house price and construction cost. Repayments of loan, L, include interest and principle. Tax is regarded as a negating benefit.

Fig. 4 A system of reduced coordinate with standardized benefit, B', and cost, C'



$$B_t = f(Rental\ income,\ SP^{\dagger}, -tax)_t \tag{10}$$

$$C_t = \left\{ \begin{array}{c} h^*, construction\ cost^*, stamp\ duty^*, \\ land\ tax, L, council\ rates \end{array} \right\}_t \tag{11}$$

The variation of parameters is dependent on project duration. In this study, the property is assumed to have been purchased in 2013 and sold 3 years later. During the investment period, rental income is expressed as a function of yield. In simulating input variables **B** and **C**, their standard deviations are arbitrarily fixed at $55,000 [19] for model illustration.

3.5.1 Efficiency

MCS and LHS are used to simulate the project's benefit and cost. Figures 5 and 6 illustrates the cost distribution generated from MCS and LHS respectively through iterations at n = 10, 50, 100, 500, and 1000 compared with the ERA method. It is evident that LHS is superior over MCS in all iterations. For instance even at the lowest iteration of n = 10, the distribution of LHS follows the distribution of ERA closer than the distribution of MCS. Convergence is achieved faster with less iteration for LHS. LHS converges between n = 50 and 100 iterations, however MCS requires n = 500 to 1000 iterations. This is significant for models with complex multiple probabilistic variables as LHS provides increased accuracy, and

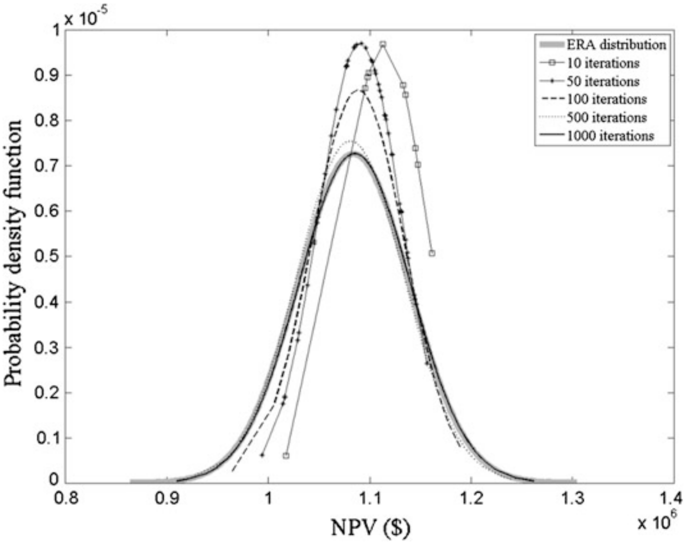


Fig. 5 Cost distribution generated by MCS (converged between 500 and 1000 iterations)

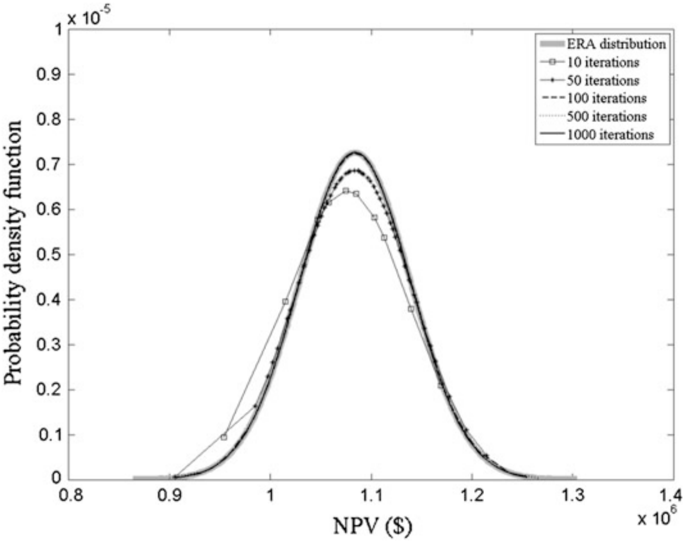


Fig. 6 Cost distribution generated by LHS (converged between 50 and 100 iterations)

is computationally less demanding and thus less time-consuming. At large iterations (n = 1000), the three distributions exhibit little difference.

It is important to note the clustering effect from MCS sampling. Sampled points focus about the mean and neglect tail-end events for iterations n = 10, 50 and 100.

Table 1 Comparison of P_f between MCS and LHS with ERA

Iterations	MCS		LHS	
	$P_{f,MCS}$	$P_{f,MCS}$ error from $P_{f,ERA}$(%)	$P_{f,LHS}$	$P_{f,LHS}$ error from $P_{f,ERA}$(%)
10	0.600	111.51	0.200	−29.50
50	0.420	48.06	0.300	5.75
100	0.340	19.86	0.290	2.23
500	0.266	−6.23	0.282	−0.59
1000	0.275	−3.06	0.284	0.11

In contrast, the distribution is well represented beyond n = 10 in LHS, enabling low probability events to be included in the analysis. This is useful for projects requiring evaluation of catastrophic events that involve high loss but low probability of occurrence such as mining disasters and global financial crisis. ERA does not require simulation if the variations of the variables are known or supplied by reliable experts. In such cases, ERA is considered most efficient as it provides output almost instantaneously.

3.5.2 Accuracy

The P_f indicates the likelihood of a project experiencing a loss. $P_{f,MCS}$ and $P_{f,LHS}$ are evaluated in instances where $\mathbf{B} < \mathbf{C}$ or NPV < 0 within the run of iteration. $P_{f,ERA}$ is 0.284 over a 3-year period. That is, there is a 28.4 % probability that the project will experience a loss in 3 years. Table 1 shows the P_f simulated by MCS and LHS with the number of iterations required for an appropriate representation of $P_{f,ERA}$. It is clear that the higher the iteration, the higher the accuracy of P_f obtained. For instance, increasing the iteration by 10 fold from 10 to 100 runs improves the percentage error from −29.50 % to 2.23 % for LHS. In order to obtain similar P_f from MCS and LHS as ERA, the iterations required are much lower for LHS, thus computationally more efficient than MCS in achieving similar accuracy. The acceptable percentage error depends on the nature of the investment and the risk preference of the investor. If the acceptable percentage error is targeted at 5 %, MCS and LHS require between 500 and 1000 iterations, and 50 to 100 iterations respectively. Lower tolerance to risk that requires increased accuracy in P_f can be achieved by an increased number of iteration runs.

As $P_{f,ERA}$ is calculated directly from the benefit and cost distribution without iterative simulation, ERA estimation is considered accurate assuming variations of variables are also accurate.

Table 2 ERA expected loss

	L_v	E(L)
P_{20}	$20,975	$5,950
P_{10}	$55,193	$15,657
P_5	$83,452	$23,673

Table 3 Comparison between MCS, LHS and ERA of expected loss

		50 iterations		1000 iterations	
		MCS	LHS	MCS	LHS
P_{20}	L_v	$40,741	$20,889	$19,480	$21,592
	E(L)	$17,111	$6,267	$5,357	$6,132
	E(L) error from ERA	187.58 %	5.32 %	–9.97 %	3.06 %
P_{10}	L_v	$57,106	$45,087	$50,271	$52,701
	E(L)	$23,985	$13,526	$13,825	$14,967
	E(L) error from ERA	53.19 %	−13.61 %	−11.70 %	–4.41 %
P_5	L_v	$91,998	$76,177	$81,799	$81,800
	E(L)	$38,639	$22,853	$22,495	$23,231
	E(L) error from ERA	63.22 %	–3.46 %	–4.98 %	−1.87 %

4 An Integrated Risk Model

4.1 Introduction

Risk, R_i, is a measure of likelihood, H_i, and consequence, Q_i, of an event occurring, as described in Eq. (1). This section highlights how P_f may be used to determine the overall financial risk of a project. The consequence associated with P_f is fundamentally taken as monetary loss (L_v) and is a value when $\mathbf{B} < \mathbf{C}$ or NPV < 0. A range of L_v is used to demonstrate project risk and is estimated from the output distribution of NPVs generated from the three methods associated with each P_f. Expected loss, E(L), as a product of P_f and L_v, is used as an estimate of R_i, as shown in Eq. (12).

$$\mathrm{E(L)} = \mathrm{P_f} \times \mathrm{L_V} \tag{12}$$

4.2 Expected Loss

The range of L_v is estimated at the 5th (P_5), 10th (P_{10}) and 20th (P_{20}) percentile of the NPV distribution as $P_{f,ERA} = 0.284$. The comparison between ERA and two sets of iterations from MCS and LHS is presented in Tables 2 and 3. Iteration 50 was chosen due to LHS convergence. Iteration 1000 showed convergence of both methods. Using ERA, at P_{20} it is estimated that L_v will be \$20,975, with an E(L) of \$5,950 when likelihood is considered. In all cases, LHS is more accurate compared to MCS especially when iteration is low. For instance at P_{20} with 50 iterations, the difference of E(L) between MCS and ERA is 187.6 %, while it is 5.3 % between LHS and ERA. At 1000 iterations, the difference is less as both simulations have reached convergence. For instance at P_{10}, the differences for MCS and ERA are −11.7 and −4.4 % respectively at 1000 iterations.

Although the focus of this study is the evaluation of E(L), it should be noted that the difference of L_v is fairly high between MCS and ERA at 50 iterations. This is as expected because the sampling method of MCS requires a high number of iterations to achieve accuracy.

5 Summary

This study presented a risk modeling framework on a residential property with focus on evaluating P_f. The three methods compared were MCS, LHS and ERA. In simple projects with modest number of variables, MCS is suitable as it is relatively straightforward to implement and able to generate theoretical solution at large number of iterations. For larger projects involving more variables of complex nature, computational intensity could be reduced with stratified sampling in LHS. Projects with reliable historical data or expertise from similar past projects could employ ERA to remove the need for sampling altogether.

For this particular residential property case study, it was found that over a three year investment period the investor would expect a P_f of 0.284 from the ERA approach. The same answer was reached using MCS with iterations between 500 and 1000 for convergence. LHS was found to show improved accuracy and efficiency, only requiring between 50 and 100 iterations. E(L) was used as a measure of R_i and was calculated from P_f and L_v. A range of L_v was applied to calculate E(L). For instance at P_{20}, the L_v and E(L) was \$20,975 and \$5,950 respectively. Further work could investigate the capability of the model in other industry sectors, and to extend the analysis to non-monetary aspects such as social and environmental factors of project appraisals.

Appendix

Historical median house price were obtained from The Real Estate Institute of Victoria [22]. Construction costs were extracted from items 13.1.2.5, 13.1.3.7, 13.1.3.8 in the Australian Construction Handbook by Rawlinsons from years 1983 to 2013. For simplicity, linear regression was used to forecast construction costs as the focus of this study is on house price uncertainty. Land tax [23] and stamp duty specifications [24] were taken from the State Revenue Office Victoria. Council rates were from the City of Melbourne, year 2013 to 2014 [25]. Rental income was the average yield from 1993 to 2009 from Westpac Property Outlook. Taxable income involved rental income and capital gain minus interest payment and stamp duty. Tax specifications were extracted from Australian Taxation Office [26]. The average cash rate between 1993 and 2013 was from the RBA interest rates and yields table series FIRMMCRT [27] and used as the discount rate of 5.2 %. All costs and benefits are in Australian dollars. The base year is 2013 and the loan for house purchase and construction was assumed to be 80 % with an interest rate of 6 %.

References

1. A. Haldar, S. Mahadevan, *Probability, Reliability and Statistical Methods in Engineering Design* (Wiley, New York, 2000)
2. S.E. Chia, Risk assessment framework for project management, in *EEE* (2006), pp. 376–379
3. N. Gil, B.S. Tether, Project risk management and design flexibility: analysing a case and conditions of complementarity. Res. Policy **40**(3), 415–428 (2011)
4. A. Nieto-Morote, F. Ruz-Vila, A fuzzy approach to construction project risk assessment. Int. J. Project Manage. **29**(2), 220–231 (2011)
5. Standards Australia, in *AS/NZS ISO31000:2009 Risk Management—Principles and Guidelines*, Sydney, Australia (2009)
6. A. Riabacke, Managerial decision making under risk and uncertainty. IAENG Int. J. Comput. Sci. **32**(4), 453–459 (2006)
7. H.M.S. Treasury, *The Green Book: Appraisal and Evaluation in Central Government* (TSO, London, 2003)
8. P. Iskanius, Risk management in ERP project in the context of SMEs. Eng. Lett. **17**(4), 266–273 (2009)
9. P. Bhattacharjee, K. Ramesh Kumar, T. Janardhan Reddy, Structural safety evaluation using modified latin hypercube sampling technique. Int. J. Perform. Eng. **9**(5), 515–522 (2013)
10. J. Imai, K.S. Tan, Dimension reduction approach to simulating exotic options in a Meixner Lévy market. IAENG Int. J. Appl. Math. **39**(4), 265–275 (2009)
11. M.D. McKay, W.J. Conover, R.J. Beckman, Comparison of three methods for selecting values of input variables in the analysis of output from a computer code. Technometrics **21**(2), 239–245 (1979)
12. S.S. Drew, T. Homem-de-Mello, Some large deviations results for Latin hypercube sampling. Methodol. Comput. Appl. Probab. **14**(2), 203–232 (2012)
13. A.Z. Grzybowski, Monte Carlo analysis of risk measures for blackjack type optimal stopping problems. Eng. Lett. **19**(3), 147–154 (2011)
14. J.P.C. Kleijnen, Sensitivity analysis and related analyses: a review of some statistical techniques. J. Stat. Comput. Simul. **57**(1–4), 111–142 (2007)
15. D. Straub, A. Der Kiureghian, Bayesian network enhanced with structural reliability methods: methodology. J. Eng. Mech. **136**(10), 1248–1258 (2010)
16. M. Tichý, *Applied Methods of Structural Reliability*, vol. 2 (Kluwer Academic Publishers, The Netherlands, 1993)
17. J. Lai, L. Zhang, C. Duffield, L. Aye, Economic risk analysis for sustainable urban development: validation of framework and decision support technique. Desalin. Water Treat, **52**(4–6), 1109–1121 (2014)
18. J. Lai, L. Zhang, C. Duffield, L. Aye, Financial risk analysis for engineering management: a framework development and testing, in Lecture Notes in Engineering and Computer Science,

Proceedings of the World Congress on Engineering and Computer Science 2013, WCECS 2013, San Francisco, USA, 23–25 Oct, (2013), pp. 1042–1046

19. J. Lai, L. Zhang, C. Duffield, L. Aye, Engineering reliability analysis in risk management framework: development and application in infrastructure project. IAENG Int. J. Appl. Math. **43**(4), 242–249 (2013)
20. S. Adarsh, M.J. Reddy, Reliability analysis of composite channels using first order approximation and Monte Carlo simulations. Stoch. Env. Res. Risk Assess. **27**(2), 477–487 (2013)
21. Australian Bureau of Statistics, in *Average Floor Area Of New Dwellings, Building Approvals, 8731.0*, Canberra, Australia. (2003), p. 38–40
22. The Real Estate Institute of Victoria, in *Market History*, Melbourne, Camberwell, Victoria, Australia, 31 July 2013
23. State Government of Victoria, in *Land Tax Act 2005, Authorised Version No. 045, No. 88 of 2005, Chief Parliamentary Counsel, Editor, State Government of Victoria*, Melbourne, Victoria, Australia (2013)
24. State Government of Victoria, in *Duties Act 2000, Authorised Version No. 090, No. 79 of 2000, Chief Parliamentary Counsel, Editor, State Government of Victoria*, Melbourne, Victoria, Australia (2012)
25. City of Melbourne, in *How Your Rates are Calculated,* Melbourne, Victoria, Australia, 20 Feb 2014 (2013)
26. CPA Australia, in *Tax and Social Security Guide 2013–2014*, CPA Australia, Melbourne, Victoria, Australia (2013)
27. Reserve Bank of Australia, in *Interest Rates and Yields*, Sydney, New South Wales, Australia, 17 Sept 2013

Chapter 51
Virtual Production Intelligence: Process Analysis in the Production Planing Phase

Daniel Schilberg, Tobias Meisen and Rudolf Reinhard

Abstract To gain a better and deeper understanding of cause and effect dependencies in complex production processes it is necessary to represent these processes for analysis as good and complete as possible. Virtual Production is a main contribution to reach this objective. To use the Virtual Production effectively in this context, a base that allows a holistic, integrated view of information that is provided by IT tools along the production process has to be created. The goal of such an analysis is the possibility to identify optimization potentials in order to increase product quality and production efficiency. The presented work will focus on a simulation based planning phase of a production process as core part of the Virtual Production. An integrative approach which represents the integration, analysis and visualization of data generated along such a simulated production process is introduced. This introduced system is called Virtual Production Intelligence and in addition to the integration possibilities it provides a context-sensitive information analysis to gain more detailed knowledge of production processes.

Keywords Analysis · Digital factory · Laser cutting · Production technology · Virtual production · Virtual production intelligence · VPI

D. Schilberg (✉) · T. Meisen · R. Reinhard
Institute of Information Management in Mechanical Engineering,
RWTH Aachen University Germany, Dennewartstraße 27, 52068 Aachen, Germany
e-mail: daniel.schilberg@ima.rwth-aachen.de

T. Meisen
e-mail: tobias.meisen@ima.rwth-aachen.de

R. Reinhard
e-mail: rudolf.reinhard@ima.rwth-aachen.de

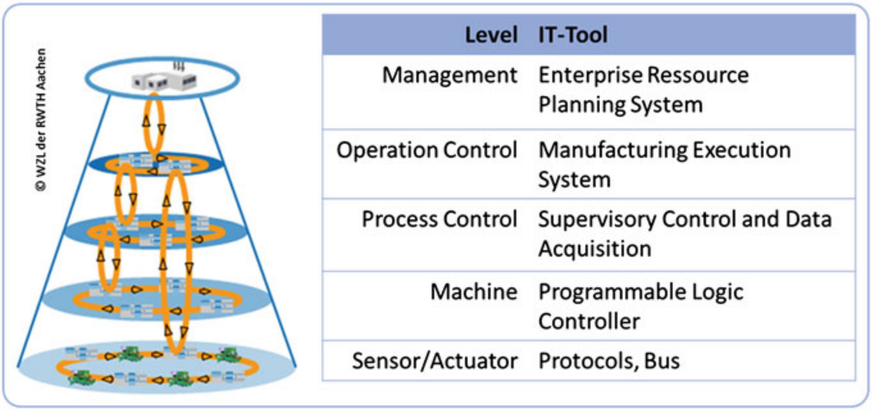
H. K. Kim et al. (eds.), *Transactions on Engineering Technologies*,
DOI: 10.1007/978-94-017-9115-1_51,

1 Introduction

Considering the individualization and increasing performance of products the complexity of products and production processes in mechanical and automatic processing is constantly growing. This, in turn, results in new challenges concerning the designing as well as the production itself. In order to face these challenges, measures are required to meet the demands which are based on higher complexity. One measure to face this challenge is a more detailed planning of the design and manufacturing of the products by the massive use of simulations and other IT tools which enable the user to fulfill the various demands on the product and its manufacturing. To a further improvement of simulations and IT tools it is important not to evaluate them separately but in their usage context: which tool is used to which planning or manufacturing process. It has to be fathomed which information on which effort between the tools are exchanged.

To formulate and execute an appropriate measure, it is necessary to create a basis which allows a holistic, integrative examination of deployed tools in the process. Aim of such an examination is an increasing product quality, efficiency and performance. Due to the rapid development of high-performance computers, the use of simulations in product design and manufacturing processes has already been well-established and enables users to map relations more and more detailed virtually. This has led to a change concerning the way to perform preparatory and manufacturing activities. Instead of an early development of physical existent prototypes, the object of observation is developed as a digital model which represents an abstraction of essential characteristics or practices. The subsequent simulation the digital model is used to derive statements concerning practices and properties of systems to be examined. The use of digital models in production processes is described by the term of Virtual Production which specifies a "mainstreaming, experimental planning, evaluation and controlling of production processes and plant by means of digital models" [1, 2].

This paper will show an integrative concept which describes an important component to achieve the objective of a Virtual Production by the integration and visualization of data, produced on simulated processes within the production technology. Taking account of the application domain of production technology and used context-sensitive information analysis, with the aim of an increasing improvement of knowledge concerning the examined processes, this concept is called Virtual Production Intelligence (VPI). The aim of this paper is to present how the VPI contributes to optimize manufacturing processes like laser cutting. The usage of the VPI in a factory planning process is shown in [3].



© WZL der RWTH Aachen

Level	IT-Tool
Management	Enterprise Ressource Planning System
Operation Control	Manufacturing Execution System
Process Control	Supervisory Control and Data Acquisition
Machine	Programmable Logic Controller
Sensor/Actuator	Protocols, Bus

Fig. 1 Automation pyramid

2 Problem

As a central issue of the Virtual Production the heterogeneous IT landscape can be identified. As indicated in the introduction, a variety of software tools to support various processes are used. Within these software tools data cannot be exchanged without effort. The automation pyramid [4] offers a good possibility to demonstrate this difficulty. The automation pyramid is depicted in Fig. 1. It shows the different levels of the automation pyramid with the corresponding IT tools and the flow of information between the levels. The level related processes are supported by the mentioned IT tools very good or at least sufficient. At the top level command and control decisions for the company management are supported by Enterprise Resource Planning (ERP) systems. Therefore, these systems allow the decision-makers in the management to monitor any enterprise-wide resources like employees, machinery or materials.

At the lower levels, the Manufacturing Execution Systems (MES), the data acquisition (Supervisory Control and Data Acquisition SCADA) and programmable logic controllers (PLC) are arranged according to the increasing complexity. The field level is the lowest level. Corresponding to protocols the data exchange is organized on this level. The used software tools are developed very well to support the corresponding processes on the appropriate level. The Association of German Engineers (VDI) addressed for the Virtual Production a unified data management as a way to use data and information across all levels, but this is not realized yet. Without a unified data management the data exchange from a PLC via the SCADA and MES up to the ERP system requires a great effort for conversions and aggregating. The goal is that the ERP can support decisions on the base down to the PLC Data and changes in the ERP system will change the input for the PLC. Currently most companies only exchange data between different levels instead of a flow of information across all levels. This is why a holistic picture of production and manufacturing process is not possible [5].

At present, a continuous flow of information is available only with the application of customized architectures and adapters to overcome the problem of heterogeneity. This involves high costs, why usually small and medium-sized enterprises (SMEs) have no integration of all available data into a system.

There are a high number of different IT tools for Virtual Production. These enable the simulation of various processes, such as in manufacturing technology, the realistic simulation of heat treatment and rolling process or the digital viewing of complex machinery such as laser cutting machines. At this juncture various independent data formats and structures have developed for a representation of the digital models. Whereas an independent simulation of certain aspects of product and manufacturing planning is possible, the integrative simulation of complex manufacturing processes involves high costs as well as an high expenditure of time because in general an interoperability between heterogeneous IT tools along the automation pyramid is not given. One approach to overcome the heterogeneity is the homogenization with the help of a definition of unified data standards. In this context a transfer of the data formats into a standardization of data by the use of specific adapters as mentioned above. However, this approach is not practical for the considered scenario for two reasons. Firstly, the diversity of possible IT tools that are used lead to a complex data standard. This is why its understanding, care and use are time and cost intensive. Secondly, the compatibility issues for individual versions of the standard are to be addressed (see STEP [6]). Therefore the standard must be compatible with older versions and enhanced constantly to reflect current developments of IT tools and to correspond to the progressive development through research [7, 8].

Another approach, which is chosen as basic in this paper, includes the use of concepts of the data and application integration, which do not require a unified standard. The interoperability of IT applications must be ensured in a different way so that no standard data format is necessary. This is done by mapping the various aspects of the data formats and structures on a so-called integrated data model or ¬ canonical data model [8, 9]. In current approaches to these concepts are extended to the use of semantic technologies. The semantic technologies enable a context-sensitive behavior of the integration system. The continuation of this approach enables the so-called adaptive application and data integration [10, 11].

The integration of all data collected in the process in a consolidated data management is only the first step to solving the problem. The major challenge that must be overcome is the further processing of the integrated data along a production process to achieve a combination of IT tools across all levels of the automation pyramid. The question of the analysis of data from heterogeneous sources is addressed in the analysis of corporate data for some time. The applications that enable integration and analysis of data are grouped under the term “Business Intelligence” (BI). BI applications have in common that they provide the identification and collection of data that arise in business processes, as well as their extraction and analysis [12, 13].

The problem in the application of BI on Virtual Production is that the implementation of the BI integration challenges of heterogeneous data and information conceptually solves in the first place which causes significant problems in the implementation of functional systems. Thus, in concept, for example, a translation of the data into a common data format and context-sensitive annotation is provided. A translation may not be achieved because it is proprietary information which meaning is not known to the annotation. This is also the reason why so many BI integrations have failed so far [14].

The following shows that the previously addressed problems should be solved by the vision of the digital factory. Because this vision is not realized yet, the section heterogeneity of simulations and solution: Virtual Production Intelligence will outline next steps towards the realization of a digital factory. The term "Virtual Production Intelligence" was selected in reference to the problem introduced in the term "business intelligence", which has become popular in early to mid-1990s. It called "business intelligence" methods and processes for a systematic analysis (collection, analysis and presentation) of a company's data in electronic form. Based on gained findings, it aims at improved operative or strategic decisions with respect to various business goals "Intelligence". In this context "Intelligence" does not refer to intelligence in terms of a cognitive size but describes the insights which are provided by collecting and preparing information.

3 Digital Factory

The digital factory is defined by the working group VDI in the VDI guideline [1] as:

> *the generic term for a comprehensive network of digital models, methods and tools—including simulation and 3D visualization—integrated by a continuous data management system. Its aim is the holistic planning, evaluation and ongoing improvement of all the main structures, processes and resources of the real factory in conjunction with the product.*

According to the VDI guideline 4499 the concept of the digital factory does not include individual aspects of the planning or production but the entire product life cycle (PLC) (Fig. 2). All processes from the onset to the point of decommissioning shall be modeled. Therefore the observation starts with the collection of market requirement, the design stages including all the required documents, project management, prototypes (digital mock-ups), the necessary internal and external logistic processes, planning the assembly and manufacturing, the planning of appropriate manufacturing facilities, installation and commissioning of production facilities, the start-up management (ramp up), series production, sales to maintenance and ends with the recycling or disposal of the product all these points should be part of the Digital Factory. Currently there is no platform which complies with this integration task. But there are already implemented some elements of the digital factory at different levels of the automation pyramid or in phases of the PLC.

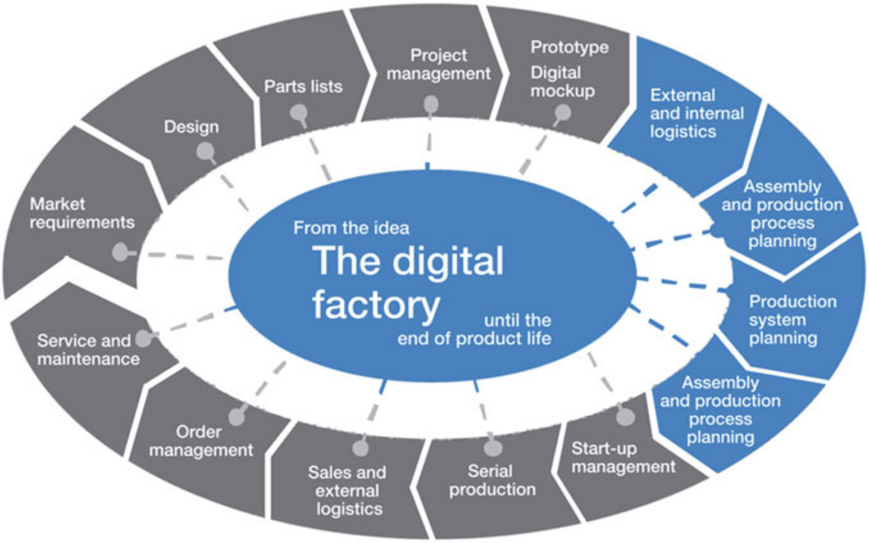

From the idea
The digital factory
until the end of product life
Market requirements
Design
Parts lists
Project management
Prototype
Digital mockup
External and internal logistics
Assembly and production process planning
Production system planning
Assembly and production process planning
Start-up management
Serial production
Sales and external logistics
Order management
Service and maintenance

Fig. 2 Product life cycle (VDI 4499) and localization of Virtual Production within the product life cycle in accordance with VDI Directive 4499

Existing PLC Software products help companies to plan, monitor and control the product life cycle in parts. However, these applications are usually only isolated solutions and enable the integration of IT tools that have the same interfaces for data exchange and are provided by the same manufacturer. The detail of the images of individual phases of the product life cycle does not reach this high spatial resolution of special applications to the description of individual phases of the product life cycle or of IT tools that focus on aspects of individual phases. Therefore the recommendation of the VDI to design data management and exchange as homogeneous as possible can only be considered for new developments. Besides there is still no approach about how to implement a standard for such a homogeneous data exchange and how to prevent or avoid the known issues of a standardization process. Therefore even a project that wants to realize the homogenization of the flow of information cannot succeed, because it is not defined what such a condition has to look like. Moreover there is no standard or efforts to standardize as for example the Standard for the Exchange of Product Model Data (STEP) compete with proprietary formats. It must be considered that the proprietary formats were also used to protect the knowledge and skills of the software provider.

With view to a visualization of the digital factory there are tools of Virtual and Augmented Reality which enable users to realize 3D models of factories with or without people as well as to interact with it and to annotate information. A real time-control of a physical existent plant via virtual representation, at which data from the operation in virtual installation are illustrated and further processed for analysis, is right now not possible. The running times of individual simulations do not meet the real-time requirement. With the present techniques, its developments and innovations the goal of digital manufacturing is to be achieved.

The Virtual Production Intelligence serves as a basic building block for the digital factory. To achieve this goal, it is not necessary to address the overall vision of the digital factory, but rather it is sufficient to focus the area of simulation-based Virtual Production (see Fig. 2). Again, the VDI guideline 4499 is cited to the definition of Virtual Production:

> is the simulated networked planning and control of production processes with the aid of digital models. It serves to optimize production systems and allows a flexible adaptation of the process design prior to prototype realization.

The production processes are here divided into individual process steps, which are described by simulations. The simulation of the individual process steps is done using modern simulation tools which can represent complex production processes accurately. Despite the high accuracy of individual simulations the central challenge in virtual manufacturing is the sum of individual process steps in a value chain.

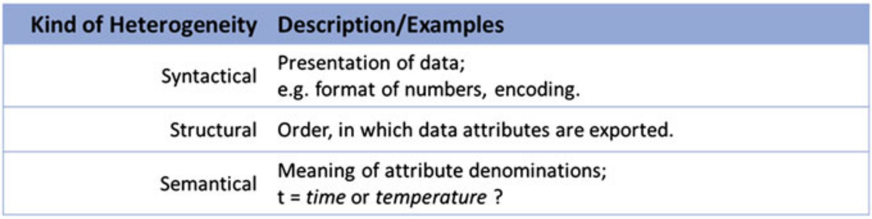
The VPI is developed to set the interoperability of IT tools in a first step with distinctly less effort than using tailored solutions mentioned above. In a second step the integrated data is consolidated, analyzed and processed. The VPI is a holistic, integrative approach to support the implementation of collaborative technology and product development. Thereby enabling optimization potentials are identified and made available for the purpose of early identification and elimination of errors in processes. To better understand the terms holistic, integrative and collaborative will be defined as follows:

- Holistic: all parts of the addressed processes will be taken into consideration.
- Integrative: use and integration of existing solutions.
- Collaborative: consideration of all processes addressed in involved roles as well as their communication

In the next section, the above-mentioned heterogeneities that should be overcome by the use of the VPI, a closer look.

4 Heterogeneity

Regarding ISO/IEC 2382-01 [15] interoperability between software applications is realized when the ability exists to communicate, to run programs, or to transfer data between functional units is possible in such a way that the user need no information about the properties of the application. Figure 3 summarizes the heterogeneities, which contribute significantly to the fact that no interoperability is achieved without using customized adapters [16–18].



Kind of Heterogeneity	Description/Examples
Syntactical	Presentation of data; e.g. format of numbers, encoding.
Structural	Order, in which data attributes are exported.
Semantical	Meaning of attribute denominations; t = *time* or *temperature* ?

Fig. 3 Types of heterogeneity of simulations

The syntactic heterogeneity describes the differences in the technical description of data, for example different coding standards such as ASCII or binary encoding, or the use of floating-point numbers as float or double. These two types of heterogeneity can be overcome relatively easy by the use of adapters. Therefore a generic approach should be applied, so that the implemented adapters are reusable. Existing libraries and solutions are available to address the problem of technical heterogeneity. Most modern programming concepts contain methods for implicit type adjustments and controlling explicit conversion of data [16–18].

Overcoming the structural and semantic heterogeneity is the much greater challenge. Structural heterogeneity differences specify the representation of information. Semantic heterogeneity describes the differences in the importance of domain specific entities and concepts used for their award. E.g. the concept of ambient temperature is used by two simulations, simulation A, uses the concept to define the room temperature of the site where the heating furnace is located. Simulation B uses the concept to define the temperature inside the heating furnace so the temperature in the immediate vicinity of the object to be heated is specified.

In the following section, the VPI is presented, which provides methods to overcome of the mentioned heterogeneity and to facilitate interoperability between applications [16–18].

5 Virtual Production Intelligence

The main objective for the use of the “Virtual Production Intelligence” is to gather results of a simulation process, to analyze and visualize them in order to generate insights that enable a holistic assessment of the individual simulation results and aggregated simulation results. The analysis is based on experts know how and physical and mathematical models. Through an immersive visualization requirements for a “Virtual Production Intelligence” are completely covered.

The integration of result data from a simulation process in a canonical data model (Fig. 4) is the first step to gain knowledge from these data and to realize the extraction of hidden, valid, useful and actionable information.

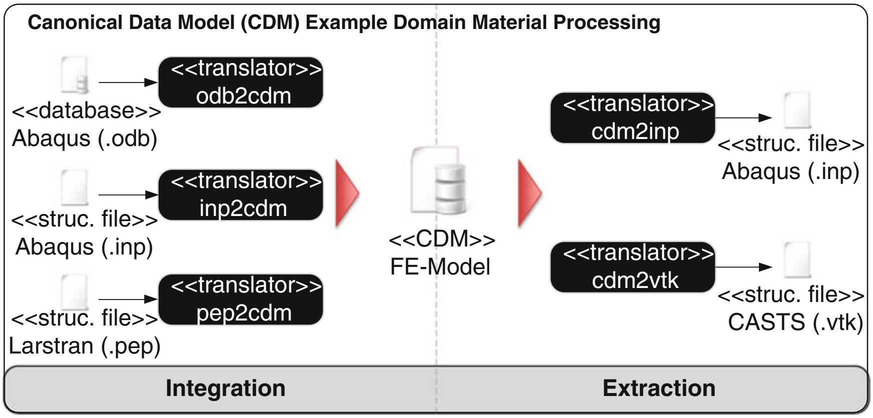


Fig. 4 Canonical data model VPI

This information includes, for example, the quality of the results of a simulation process or in concrete cases, the causes for the emergence of inconsistencies. Right now the user who has to identify such aspects has currently limited options to do so. With the realization of an integration solution a uniform view to the data gets possible. This includes on the one hand the visualization of the entire simulation process in a visualization component and on the other hand, the analysis of the data over the entire simulation process. For this purpose, different exploration methods can be used.

First, the data along the simulation process is integrated into a canonical data model. This is implemented as a relational data model, so that a consistent and consolidated view of the data is possible. Subsequently, the data is analyzed on the analysis level by the user. The user can interact in an immersive environment to explore and analyze the data. With the ability to provide feedback to the analysis component, the user can selectively influence the exploration process and make parameterizations during runtime.

In addition to a retrospective analysis by experts, it is also useful to monitor the data during the simulation process. Such a process monitoring assures compliance with parameter corridors or other boundary conditions. Therefore, if a simulation provides parameter values outside the defined parameter corridors the simulation process will be terminated. Then experts can analyze the current results in order to subsequently perform a specific adaptation of the simulation parameters. A process monitoring could also enable the extraction of point-of-interests (POI) on the basis of features that would be highlighted by the visualization (Fig. 5). The components of the "Virtual Production Intelligence" are shown in [3].

An effective optimization of different structures of production, such as determining the number of process chains and production segment is made possible only by mapping the interdependencies of different planning modules.

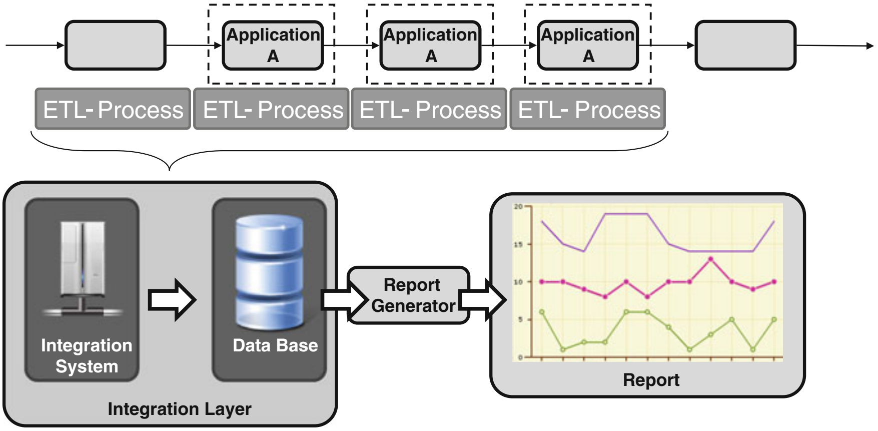


Fig. 5 Extraction of point-of-interests

6 Application Domain Laser Cutting

The VPI is used to identify relevant machine parameters to optimize laser cutting processes concerning different goals like quality or speed. At first a brief look on the process itself is given. Laser cutting is a thermal separation process widely used in shaping and contour cutting applications. Therefore the laser cutting process has some advantages over conventional cutting techniques. The cutting process is very fast and accurate.Because an optical tool is used there is no risk of additional wear.

The ablation process in fusion metal cutting is based on thermodynamics and hydrodynamics. The absorbed laser energy is converted to heat which melts the material. This melt is driven out of the cut by a gas jet that is coming out of the nozzle, coaxially aligned to the laser beam. The VPI is a simulation based tool; therefore not the real process is used for optimization but the simulated. Hence, the analysis results of the VPI strongly correlate to the quality of the simulated laser cutting process. The core of each simulation is the simulation model that is used. The modelling of a laser cutting process requires the modelling of at least three entities at the same time. The optical tool—the laser beam—must be included, the material that should be cut by the laser beam, and the gas jet separating the melt. To gain a good model it is evident that the modeling of the following quantities has to be accomplished as well as their numerical implementation:

- The cutting gas flow
- The radiation propagation
- The ablation of the material

Figure 6 shows the simulation results based on a numerical model developed by the NLD at RWTH Aachen University, Germany, for the ablation ant the beam propagation into the cut kerf.

Fig. 6 Ablation simulation for laser cutting

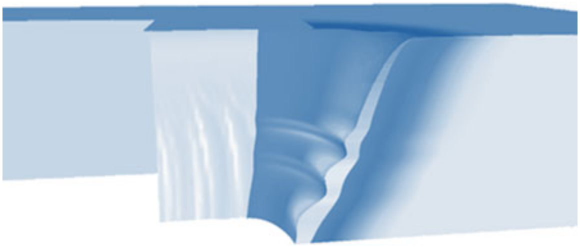
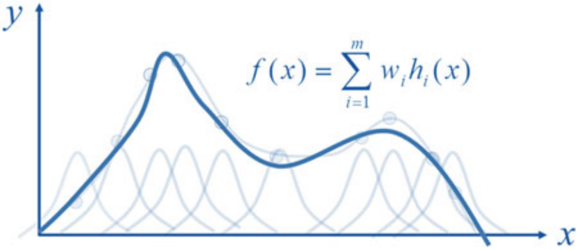


Fig. 7 Method used by the VPI for reduction of number of parameters



There are, however, gaps in understanding the dynamics of the process, especially issues related to cut quality. The user of a laser cutting machine needs to know how the surface roughness on cut edge can be better influenced. How can dross formation on the cut bottom be avoided and the inclination of cut edge be controlled? Especially it is important to understand the influence of laser parameters on those quality criteria. The most important parameters may be the wave length and the modelling of the wave length. Is a gas laser better than a solid state laser? The shape of the beam must be analysed: What is the influence of a Gaussian or a tophat shaped beam? Should the polarization circular or radial? Hence, the goal of the VPI is the reduction of numbers of parameters that are relevant for reaching a certain cutting quality. For that the correlation between chosen output or criterion and parameters or inputs must be determined. The VPI uses different methods to find the correlations (like Fig. 7). The VPI uses a sensitivity analysis:

> *The study of how uncertainty in the output of a model (numerical or otherwise) can be apportioned to different sources of uncertainty in the model input.* (Saltelli et al. 2004) [19]

The results of qualitative methods can be visualized by scatter plots and quantitative methods will be used like the computation of rank correlation coefficients between various criteria and parameters.

The VPI is used for the planning of the laser cutting process. It supports the user in three ways. At first by using the VPIs data integration possibilities the user can gather data from various sources and get a consolidated view on these data. In the second way the VPI provides an explorative visualization to present the data and to facilitate interaction. The last way is the data analysis to determine where and how to optimize process outcome.

7 Conclusion

In this chapter the Definition of the VPI was given as a holistic and integrative concept for the support of the collaborative planning, monitoring and control of core processes within production and product development in various fields of application. As a role model the idea of the Business Intelligence is used, applied to the domain of Virtual Production. The aim of the VPI is the identification and elimination of error sources in planning processes as well as detecting and taking advantage of enhancement potentials.

With the VPI an essential contribution to the realization of the vision of the digital factory can be achieved. The VPI is an integration platform that enables heterogeneous IT tools in the phase of product and production planning to interoperate with each other. Based on information processing concepts it supports the analysis and evaluation of cause-effect relationships. As product and production planning is the core area of Virtual Production, as part of the digital factory, the contribution is focused on this part. The VPI is the basis to establish interoperability. The functionality of the VPI was presented and illustrated by using the example of factory planning. The use of the VPI allows a significant reduction in engineering effort to create tailored integration and analysis tools, since the VPI is an adaptive solution. Now it is possible to start with a process-oriented and so contextual information processing. Information is now not only based on a single process step, it is related to the overall process, so that the importance and validity of information can be considered.

The future work concerning the VPI in the domain laser cutting will be the determination of machine parameters depending on desired machine states. That will be an optimization problem for a multidimensional function. The solution could be an explorative visualization based on the concept of hyperslices linked with 3D volume visualization. It is important to evaluate what cause-effect relationships can be identified through the exploration process. Furthermore, it must be examined how this information can be presented to the user in an immersive environment, and how can context information understandable and comprehensible be presented. For this purpose, there are various feedback-based techniques in which experts assess results of analysis and optimization. A bidirectional communication is needed, the user gives feedback and this feedback will be used to correct the displayed information. The system will store this feedback to avoid imprecise or erroneous statements.

References

1. VDI Guideline 4499, Sheet 1, 2008: Digital Factory
2. VDI Guideline 4499, Sheet 2, 2011: Digital Factory
3. D. Schilberg, T. Meisen, R. Reinhard, Virtual production—the connection of the modules through the virtual production intelligence, in *Lecture Notes in Engineering and Computer*

Science: Proceedings of The World Congress on Engineering and Computer Science 2013, WCECS, San Francisco, USA, 23–25 Oct, (2013), pp. 1047–1052

4. R. Lauber, P. Göhner, *Prozessautomatisierung,* 1. 3. Aufl. (Springer, Berlin, 1999)
5. H. Kagermann, W. Wahlster, J. *Helbig, Umsetzungsempfehlungen für das Zukunftsprojekt Industrie 4.0—Abschlussbericht des Arbeitskreises Industrie 4.0* (Forschungsunion im Stifterverband für die Deutsche Wissenschaft, Berlin, 2012)
6. DIN EN ISO 10303
7. M. Nagl, B. Westfechtel, *Modelle, Werkzeuge und Infrastrukturen zur Unterstützung von Entwicklungsprozessen. Symposium (Forschungsbericht (DFG)),* 1. Aufl. (Wiley-VCH, Weinheim, 2003), S. 331–332
8. C. Horstmann, *Integration und Flexibilitat der Organisation Durch Informationstechnologie*, 1. Aufl. (Gabler Verlag., Wiesbaden, 2011), S. 156–162
9. Daniel Schilberg, *Architektur eines Datenintegrators zur durchgängigen Kopplung von verteilten numerischen Simulationen* (VDI-Verlag, Düsseldorf, 2010)
10. T. Meisen, P. Meisen, D. Schilberg, S. Jeschke, Application integration of simulation tools considering domain specific knowledge, in *Proceedings of the 13th International Conference on Enterprise Information Systems* (2011)
11. R. Reinhard, T. Meisen, T. Beer, D. Schilberg, S. Jeschke, A framework enabling data integration for virtual production. in *Enabling Manufacturing Competitiveness and Economic Sustainability; Proceedings of the 4th International Conference on Changeable, Agile, Reconfigurable and Virtual production (CARV2011)*, Montreal, Canada, Hrsg. v. Hoda A. ElMaraghy, Berlin Heidelberg, 2–5 Oct (2011) S. 275–280
12. B. Byrne, J. Kling, D. McCarty, G. Sauter, P. Worcester, The Value of Applying the Canonical Modeling Pattern in SOA, IBM (The information perspective of SOA design, 4) (2008)
13. M. West, *Developing High Quality Data Models*, 1. Aufl. (Morgan Kaufmann, Burlington, 2011)
14. W. Yeoh, A. Koronios, Critical success factors for business intelligence systems. J. Comput. Inf. Syst. **50**(3), 23 (2010)
15. ISO/IEC 2382-01
16. M. Daconta, L. Obrst, K. Smith, *The Semantic Web: The Future of XML, Web Services, and Knowledge Management* (Wiley, New York, 2003)
17. D. Schilberg, A. Gramatke, K. Henning, Semantic interconnection of distributed numerical simulations via SOA, in *Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering and Computer Science 2008*, WCECS, San Francisco, USA, 22–24 Oct (2008) pp. 894–897
18. D. Schilberg, T. Meisen, R. Reinhard, S. Jeschke, Simulation and Interoperability in The Planning Phase of Production Processes, in *ASME 2011 International Mechanical Engineering Congress and Exposition*, Hrsg. v. ASME, Denver (2011)
19. A. Saltelli, S. Tarantola, F. Campolongo, M. Ratto, *Sensitivity Analysis in Practice: A Guide to Assessing Scientific Models* (Wiley, Chichester, 2004)

Chapter 52
Operating Room Scheduling Problems: A Survey and a Proposed Solution Framework

Zakaria Abdelrasol, Nermine Harraz and Amr Eltawil

Abstract Healthcare is becoming one of the fast growing industries in both, the developed and developing countries. Operating rooms provide a large portion of hospital revenue; hence, scheduling operation room surgeries is very important to maximize profits. This paper reviews the three operating room scheduling problems: the case mix problem, the master surgery scheduling problem and the surgery scheduling problem. Also, the paper introduces a research framework for an integrated planning method for the three problems.

Keywords Case-mix · Healthcare management · Literature survey · Master surgery scheduling · Operating room scheduling · Operations research · Simulation · Surgery scheduling

Z. Abdelrasol (✉) · N. Harraz · A. Eltawil
Department of Industrial Engineering and Systems Management,
Egypt-Japan University of Science and Technology (E-JUST), Alexandria, Egypt
e-mail: zakaria.abdelgawad@ejust.edu.eg

N. Harraz
e-mail: nermine.harraz@ejust.edu.eg

A. Eltawil
e-mail: eltawil@ejust.edu.eg

Z. Abdelrasol
On leave from Industrial Engineering Department, Faculty of Engineering,
Fayoum University, Fayoum, Egypt

N. Harraz · A. Eltawil
On leave from the Production Engineering Department, Faculty of Engineering,
Alexandria University, Alexandria, Egypt

H. K. Kim et al. (eds.), *Transactions on Engineering Technologies*,
DOI: 10.1007/978-94-017-9115-1_52,

1 Introduction

Healthcare is becoming one of the largest industries in the developed and developing countries. Like in many other countries all over the world, hospitals in the developing countries have faced multiple challenges in the past decade, such as the occurrence of new diseases and much more budgetary constraints. Nowadays, there are many challenges facing healthcare systems, such as the limited resources, the high cost of medical technology and medication, the high demand, the high customer expectations and the shortage in planning and management decision support tools specially with the complexity of healthcare systems. In addition to these challenges, hospitals in developing countries are facing more challenges as the limited government support. Consequently, hospitals are more and more aware of the need to use their resources as efficiently as possible, which urges healthcare organizations to increase emphasis on process optimization in order to control and minimize operating costs and improve the provided services levels.

According to the "Healthcare Financial Management Association" (HFMA), Operating rooms (ORs) present a large share of hospital care services and expenditure, at the same time, ORs result in an estimated 40 % of hospital revenue. This makes the scheduling of ORs an important problem to study in order to meet hospital goals.

In this concern, it is necessary to schedule the ORs, in such a way that the operations/surgeries are carried out with maximal efficiency. The increase in efficiency of ORs schedule has a bearing of the number of surgery cases, the cost or profit, utilization of resources and waiting time of patients, which, in turn, are widely accepted indicators of ORs efficiency.

The problem of ORs scheduling can be divided into three different and related sub-problems as shown in Fig. 1, namely (i) the Case Mix Problem (CMP), (ii) the Master Surgery Scheduling Problem (MSSP), and (iii) the Surgery Scheduling Problem (SSP). The CMP refers to the time of a resource (e.g., ORs) allocated to each surgical specialty in order to minimize the total costs or maximize the total revenues or how the available ORs time is divided over the different surgeons. This stage takes place on the strategic level of hospital management as it determines for which ailments capacity will be preserved for a long time horizon. In MSS problem, the ORs time is allocated to these surgical specialties over the scheduling window (typically, a week) in order to maximize and level resources utilization. Finally, SSP refers to assigning each surgical case a start time, a day, and an OR with the target of minimizing the waiting time and maximizing resources utilization. More details and examples about the three ORs scheduling problems and the relation between them can be found in Abdelrasol et al. [1].

There are many review papers that discussed the ORs scheduling problems [13, 14, 21]. In a recent work for the authors [1], the three operating room scheduling problems were reviewed. Furthermore, a research framework for an integrated planning method for the three problems was introduced. Moreover, the framework included a broad system dynamics model to analyze the relationship between the

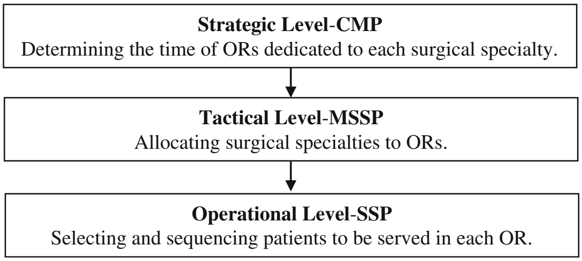

Strategic Level-CMP
Determining the time of ORs dedicated to each surgical specialty.
Tactical Level-MSSP
Allocating surgical specialties to ORs.
Operational Level-SSP
Selecting and sequencing patients to be served in each OR.

Fig. 1 The hierarchy of the three ORs scheduling problems and decision levels

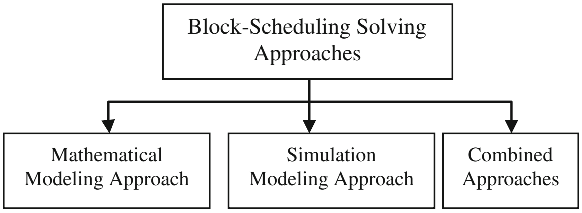

Block-Scheduling Solving Approaches
Mathematical Modeling Approach
Simulation Modeling Approach
Combined Approaches

Fig. 2 Classification of block-scheduling solution approaches

different hospital's departments and the operating rooms. In this chapter we have extended their work, by considering more references especially in the MSSP section, supporting the work with additional figures, and making some revisions (Fig. 2).

The aim of this book chapter is threefold. First, it presents a short literature review and a classification of ORs scheduling problems. The literature review and classification is structured depending on the three stages that can be distinguished in developing ORs schedules; CMP, MSSP, and SSP followed by the integration of these three problems. Figure 3 summarizes these four sections. Second, we present a gap analysis and define a recommended research direction. Third, the proposed research framework to tackle the defined problem is presented.

The remainder of this chapter will be structured as follows. Section 2 contains a focused literature on the first ORs scheduling problem, CMP. In Sect. 3 the second ORs scheduling problem will be reviewed, MSSP. The subsequent section reviewed the third ORs scheduling problem, SSP, followed by a discussion of the integrated work for the three problems. Section 6 provides a gap analysis, followed by presenting the proposed approach framework. Finally conclusion section has been provided.

2 The Case Mix Problem

This stage takes place on the strategic level of hospital management. The literature on this level is relatively sparse. The available literature can be categorized into two groups according to demand certainty. The first category contains the papers

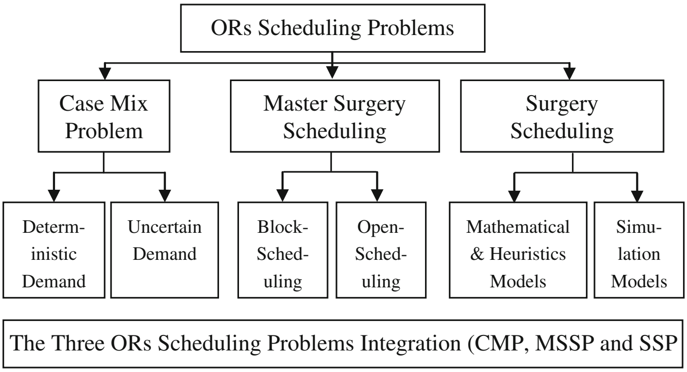


Fig. 3 ORs scheduling problems classification summary

Table 1 Case mix problem references summary

Ref.	Authors (year)	Objective function	Considered resources	Techniques
[32]	Ma and Demeulemeester (2013)	Maximize overall financial contribution	ORs' time, beds	ILP
[41]	Testi et al. (2007)	Maximize total benefit	ORs' time	BPP—binary LP
[31]	Ma et al. (2009)	Maximize profits	ORs' time, beds	ILP and B&P.
[12]	Blake and Carter (2002)	Minimize profit deviation	ORs' time, beds	LGP
[36]	Mulholland et al. (2005)	Maximize hospital total margins plus professional payments	ORs' time, ICU Beds, wards' beds	LP
[28]	Kuo et al. (2003)	Maximize professional revenues. Minimized value for hospital costs	ORs' time	LP

with deterministic demand which represent the large portion of the Case Mix Problem (CMP) literature so it will be discussed in details; the related papers are summarized in Table 1. The second category considers the variation and uncertainty on surgical demand even from a statistical viewpoint or as a newsvendor problem.

2.1 The CMP with Deterministic Demand

The literature in this category is summarized in Table 1. Further details can be found in Abdelrasol et al. [1].

2.2 The CMP with Uncertain Demand

In the CMP and due to the uncertainty in demand or SGs workloads, the optimally allocated time must balance the costs of allocating too much time, which typically translates to idle time for ORs and staff, with the costs of allocating too little time, which typically translates to overtime charges. This issue has been tackled by two approaches; from a statistical viewpoint [35] and by addressing the CMP as a newsvendor problem [45].

3 The Master Surgery Scheduling Problem

This problem is addressed at the tactical level and it is concerned with the development of the Master Surgery Schedule (MSS), which usually is cyclically constructed for a given planning period (usually 1–3 months–1 year). An MSS defines the allocated Time Blocks (TBs) of each OR to several SGs every day. Historical data and actual/forecast demand (e.g., in terms of waiting list and appointment requests for surgeries) are used as input.

The decision of dedicating ORs time to SGs can be taken by one of three strategies [38]; block-scheduling, modified block scheduling and open-scheduling. In block-scheduling, a set of TBs is assigned to specific SGs, generally for some weeks or months. Surgical cases are arranged in TBs and none of these can be released. On the contrary, open-scheduling allows assigning surgical cases to an available OR, at the convenience of surgeons. An empty schedule is filled up with surgical cases by following the order of arrival time. As a fundamental difference from block-scheduling, in open-scheduling the surgeons could choose “any workday” for a case. For this reason, open-scheduling strategy is sometimes called “any workday” strategy. These points make the open-scheduling strategy more flexible than block-scheduling. In modified block scheduling, as an integration of the block and open scheduling, the block-scheduling strategy can be modified in two ways to increase its flexibility; Some TBs are booked and others are left open, or unused TBs are released at some time before surgery [16]. However block and open scheduling strategies are mentioned clearly and applied widely in the literature, no one mentioned directly that he/she applied modified block scheduling in MSSP. Next, the most recent papers for both block-scheduling and open-scheduling are discussed.

3.1 The MSS with Block-Scheduling Strategy

Block-scheduling strategy has been applied to construct the MSS by many authors. Most recently, Ma and Demeulemeester [32] first choose the optimal patient mix then they constructed a balanced MSS in order to improve the patient service level by minimizing the total expected bed shortage. Originally Testi et al. [41] firstly solved a bin packing-like problem in order to select the optimal case-mix. Then they applied block-scheduling strategy for determining optimal time tables.

Although, many cancellations take place because the intrinsic stochastic nature of the system, a few number of authors dealt with stochastic issues. For instance; Houdenhoven et al. [26] proved that the actual OR utilization can be increased by applying the Regret-Based Random Sampling algorithm. They aimed to minimize the planned slacks while surgical case durations vary according to a normal distribution. Furthermore, the stochastic Length of Stay (LOS) for the Intensive Care Unit (ICU) and Medium Care Unit (MCU) are considered by Adan et al. [2]. They compared their approach with that of Vissers et al. [44] which considered deterministic LOSs for ICU and MCU.

As an extension to the classical MSSP, few authors integrated the problem with either operational problems like; the Surgical Case Assignment Problem (SCAP) or the nurse scheduling problem. In the SCAP surgeries are allocated and assigned to each session. Patients are selected from the waiting lists according to several parameters, including surgery duration, waiting time and priority class of the operations. Testi and Tànfani [42] proposed a model that determines, during a given planning period, the allocation of ORs' blocks to surgical sub-specialties, i.e. the so called MSSP, together with the subsets of elective patients to be operated on in each block time, i.e. the so called SCAP. Recently, Agnetis et al. [3] proposed a decomposition approach for the same integration problem. In regards to the integration with nurse scheduling problem, Beliën and Demeulemeester [6] presented a model that integrates both the nurse and the ORs scheduling process in the tactical level.

The block-scheduling literature can be further classified based on the solution approaches. There are three main approaches applied to solve this problem; Mathematical models, simulation models and combined approaches as illustrated in Fig. 2.

Many authors applied mathematical models in order to get a closed form solution. Originally, Blake et al. [10] proposed an IP model to tackle the MSSP. They described a hospital's experience using the mathematical technique of IP to solve the problem with the target of minimizing the shortfall between each group's target and actual assignment of OR time. Recently, Mannino et al. [33] tackled the MSSP and focused on balancing patient queue lengths and minimizing resort to overtime. To cope with these problems they introduced a new MILP formulation and showed its beneficial properties. Furthermore, they develop a light robustness optimization approach in order to consider surgery demand uncertainty.

Regarding the second approach, originally, Harris [24] developed a simulation model to aid decision making in the area of operating theatre time tables and the resultant hospital bed requirements. Thereafter, simulation modeling approach has not been applied as a stand a loan solving approach. Very recently Banditori et al. [5] used a simulation model in order to develop their proposed combined optimization–simulation approach.

Finally, combination approaches usually gather and utilize the power of many approaches. Blake and Donald [11] proposed an IP model and a post-solution heuristic in order to allocate operating room time. The developed heuristic has been applied to improve the schedule produced by the IP model.

In order to gain the power of both mathematical modeling approach and metaheuristics, Beliën and Demeulemeester [7] developed a number of MIP based heuristics and a Simulated Annealing (SA) algorithm to minimize the expected total bed shortage for cyclic MSSP. Posteriorly, this model is adopted and embedded in the decision support system which has been presented in Beliën et al. [8]. In the same line, the Column Generation (CG) approach has been combined with mathematical modeling in order to tackle the MSSP with uncertainty. Oostrum et al. [37] proposed a two-phase decomposition approach to deal with the uncertain duration of procedures. First, they constructed MSS plan by MILP containing probabilistic constraints. Then, they proposed a CG heuristic to maximize the operation room utilization and levels the requirements for subsequent hospital beds. Similarly, Holte and Mannino [25] proposed an implementor/adversary algorithm by combining a MILP model with a CG algorithm. The general model has been applied to compute a most robust MSS plan.

To take the advantage of simulation modeling approach especially with the existence of randomness, a combined optimization–simulation approach has been applied to fine tune the optimization model to trade-off robustness and efficiency [4]. First, they presented a MIP model considering the cases' due dates and with the objective function of maximizing the patient throughput. Secondly, they illustrated the results of a simulation study through which they tested the model solution's robustness against the randomness of surgery duration and the length of stay.

3.2 The MSS with Open-Scheduling Strategy

With regard to open-scheduling, there are no TBs that can be reserved for a particular surgeon. Many authors applied open-scheduling strategy. Recently Liu et al. [30] applied only the open-scheduling strategy and developed a heuristic algorithm. This algorithm comes from the dynamic programming (DP) idea by aggregating states to avoid the explosion of the number of states. The objective was to maximize the operating rooms' use efficiency and minimize the overtime cost. Fei et al. [19] modeled the problem of weekly ORs scheduling as ILP and solved it by a column-generation-based heuristic procedure. Then the daily scheduling problem has been treated as a two-stage hybrid flow-shop problem and solved by a hybrid genetic algorithm. The main objective functions were; minimizing the idle time and overtime and maximizing the ORs utilization.

4 The Surgery Scheduling Problem

The third and final hierarchical stage is mainly devoted to define a schedule of elective surgeries (i.e., off-line scheduling). Few papers considered on-line scheduling, aimed at modifying an existing schedule given urgent and emergency arrivals. The literature in this level is very large; it can be classified according to the solution technique to; mathematical and heuristic models and simulation models. Firstly very recent papers which considered mathematical or heuristics model are presented, followed by a detailed discussion about simulation studies.

4.1 Mathematical and Heuristics Models for the SSP

Vijayakumar et al. [43] addressed a SSP experienced at a publicly-funded hospital and conceptualize this multi-period, multi-resource, priority-based case scheduling problem as an unequal-sized, multi-bin, multi-dimensional dual bin-packing problem. A mixed integer programming (MIP) model and a heuristic based on the first fit decreasing algorithm are presented. Fei et al. [19] and Jebali et al. [27] considered surgeon agendas and recovery beds availability in open-scheduling systems.

As a trial to consider the stochastic nature of SSP, Denton et al. [15] tackled the uncertainty in surgery duration by developing a two-stage stochastic MIP model, for a daily schedule of single OR. The authors assumed that the total OR schedule duration is known in advance, whereas uncertainty in surgery duration has been represented by a set of scenarios. Hans et al. [22] presented another contribution dealing with uncertainty in surgery times by applying Simulated Annealing (SA). The emergency surgeries are not taken into account, which has been considered by Lamiri et al. [29].

4.2 Simulation Models for the SSP

Due to the stochastic nature of SSP, simulation models have been increasingly recognized as a valuable tool to handle this problem. The stochastic nature stems from many factors, for example, the dynamic arrivals especially for emergency cases, and the stochastic operating time. This tendency is confirmed by the scientific contributions reviewed in what follows.

Table 2 Simulation work with scheduling rules summary

Ref.	Authors (year)	Objective functions	Assumptions	Scheduling rules
[41]	Testi et al. (2007)	Number of operations (yearly) Overruns and shifted operations Bed utilization rates	Only elective case considered Uncertainties demand and LOS Beds and OT resources considered	LWT, LPT, SPT
[17]	Dexter et al. (1999)	ORs utilization	Elective cases	10 rules
[34]	Marcon and Dexter (2006)	Over utilized OR time Completion time Percentage days with delay	Deterministic durations	7 rules
[4]	Arnaout and Sevag (2008)	Minimizing the makespan	Stochastic operation duration. SDS	LEPST, LEPT, SEPT
[23]	Harper (2002)	ORs occupancy level Beds occupancy level	Demand Uncertainty Variability in LOS	FCFS, LTF, STF, LTSC

Simulation models are widely used in the SSP literature. The simulation studies in this problem can be classified with regards to its purpose to many categories. The main two categories are; system performance evaluation and new heuristics evaluation. There are many papers published in those two groups. Firstly, recent papers from the first category will be presented followed by the second category contributions in more details.

With regard to system performance evaluation, most recently Ma and Demeulemeester [32] have been applied a simulation model to evaluate the operational performance of the case mix and capacity decision that was obtained from CMP and MSSP under the uncertainty condition. Moreover, they studied the effect of variation on bed capacity and LOS. Another example can be shown in [39] where a discrete event simulation model is developed to evaluate the performance of ORs activity scheduling.

The use of simulation models to evaluate the performance of sequencing rules has been considered in many references. A summary of these papers can be found in Table 2. More details can be found in Abdelrasol et al. [1].

5 The Integration of the Three Problems

As mentioned previously, the ORs scheduling problem can be divided into three separated, yet related problems called; CMP, MSSP and SSP. Every decision made at a given problem influences those of the next one. This specific issue has been explicitly studied in only a few papers. In this section, only papers which integrated the three problems will be discussed. To the best of our knowledge, there are only two references that have been integrated the three problems [32] and [41]. Originally, Testi et al. [41] develop a three-phase, hierarchical approach for the three ORs scheduling problems. Most recently, Ma and Demeulemeester [32] proposed a multilevel integrative approach to the ORs planning problem. It consists of three stages, namely the CMP, the MSSP and the operational performance evaluation phase.

6 Current Gaps in the Literature

According to the literature, there are many issues that should be tackled. These issues will be classified according to the problem which they are related to.

6.1 The Case Mix Problem Related Gaps

1. Generally, the literature in the CMP is sparse and the issue of uncertain cases demand has been tackled only by statistical approaches and is still an open direction to be tackled with stochastic models.
2. Due to the nature of CMP, its solution should be integer numbers. Large number of variables results in a huge integer program even for a hospital of regular size and no commercial ILP software (e.g., ILOG CPLEX) can solve it effectively [32]. Presenting a dynamic programming and meta-heuristics models for this problem, represent an interesting field for further contributions.

6.2 The Master Surgery Scheduling Problem Related Gaps

1. Although, MSSP has attracted significant research interest; few authors introduced probabilistic constraints for tackling the uncertainty mainly in surgery demand and LOS.
2. In open scheduling, there are two decisions to be taken. Routing and scheduling rules can be imported from production management field and adapted to ORs environment, such rules can be found in Shalaby et al. [40].
3. Despite open-scheduling strategy is more flexible and tends to find a better assignment of the surgical cases than the block one [18], the open-scheduling systems are rarely adopted in healthcare [20].
4. Finally, the modified block-scheduling strategy has the advantages of both block-scheduling and open-scheduling as a combination of them; nevertheless, it has been rarely applied in literature.

6.3 The Surgery Scheduling Problem Related Gaps

1. The mathematical models presented in the SSP literature are generally difficult to be solved to optimality. Consequently, in order to address instances of large size and more realistic assumptions, future research should be devoted to developing efficient heuristic solution approaches.
2. There is a need for tackling more realistic assumptions. For example, to the best of our knowledge, no one considered breakdowns and resources uncertainty.
3. Proposing new scheduling rules considering more technical factors, for example, risk based sequencing, surgeon preferences based sequencing or a decreasing order of act severity, represent interesting fields for further contributions.
4. Finally, on-line scheduling of urgent cases and the consequent rescheduling of elective surgeries represent interesting fields for further research.

6.4 The Integration of the Three Problems Related Gaps

1. To the best of our knowledge, only two papers integrated the three problems [32] and [41]. Only one of them applied feedback loop to re-optimize the three decisions [32]. The integration problem is still an open area to be solved simultaneously.

7 A Proposed Research Framework

In this section; a proposed frame work to tackle the ORs scheduling problems will be presented. Generally, this proposal aims to consider more realistic assumptions to present efficient solutions to the considered problem. The frame work can be divided into three stages as shown in Fig. 4. In the first one, the CMP will be tackled, followed by surgeries allocation and sequencing in the second stage. Finally, investigate the results of integrating the three ORs scheduling problems by different integration approaches.

7.1 The First Stage: The Case Mix Problem

This work will be the first attempt to solve the CMP with the objective of minimizing the total cost and maximizing the total profit under demand uncertainty. A dynamic programming (DP) approach will be developed to solve this problem.

DP is a mathematical technique that is based on the idea of separating a decision making problem into smaller sub-problems [9]. Stochastic DP model will presented to consider the demand uncertainty. The optimal case mix solution will be considered as a constraint while solving the next stage problems.

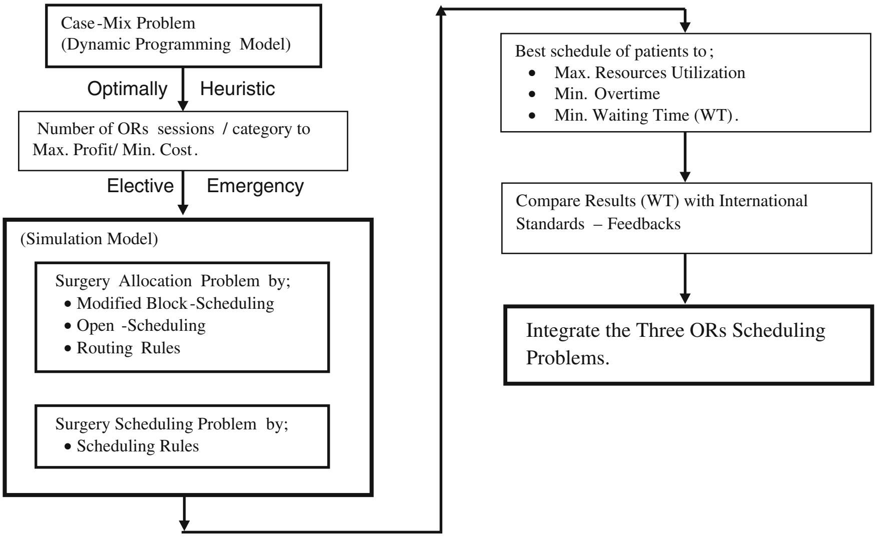

Case-Mix Problem
(Dynamic Programming Model)
Optimally
Heuristic
Number of ORs sessions / category to Max. Profit/ Min. Cost.
Elective
Emergency
(Simulation Model)
Surgery Allocation Problem by;
• Modified Block-Scheduling
• Open -Scheduling
• Routing Rules
Surgery Scheduling Problem by;
• Scheduling Rules
Best schedule of patients to;
• Max. Resources Utilization
• Min. Overtime
• Min. Waiting Time (WT).
Compare Results (WT) with International Standards – Feedbacks
Integrate the Three ORs Scheduling Problems.

Fig. 4 Basic flow chart for the proposed research frame work

7.2 The Second Stage: Surgeries Allocation and Sequencing

In this stage, there are to sub-problems; allocation problem (Case Assignment Problem) and sequencing problem. In the first one, cases are allocated to an OR among the available ORs. Modified-block scheduling and open scheduling strategies will be used to allocate elective cases to ORs, while emergency cases will be allocated by different routing rules. In the second sub-problem, cases which allocated to ORs will be sequenced according to several sequencing rules. Taking into account many realistic assumptions like, dynamic arrivals, stochastic times and equipment breakdowns, a simulation model will be built to capture the system and to test the proposed scheduling rules.

7.3 The Third Stage: The Three ORs Scheduling Problems Integration

In this stage, integration models will be built to solve the three ORs scheduling problems simultaneously. The purpose of this integration models is to link the three problems in downstream by the hierarchy relationship and also in upstream by feed backing. The objective is not only to optimize the performance of single problem but is to optimize the three ones.

8 Conclusion

This paper illustrated a comprehensive review for the different classes of operating rooms scheduling problems. Also it introduced a general research framework to solve the operating rooms scheduling problems. The proposed framework is based mainly on operations research approaches. The expected model is a generic solution that can be adapted to different cases with objective to maximize short term as well as long term profits.

Acknowledgments This research project is sponsored by the Egyptian ministry of higher education grant and the Japanese International Cooperation Agency (JICA) in the scope of the Egypt Japan University of Science and Technology (E-JUST).

References

1. Z. Abdelrasol, N. Harraz, A. Eltawil, A proposed solution framework for the operatingroom scheduling problems, in *Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering and Computer Science 2013 (WCECS 2013)*, San Francisco, USA, pp. 1149–1157 (2013)
2. I. Adan, J. Bekkers, N. Dellaert, J. Vissers, X. Yu, Patient mix optimisation and stochasticresource requirements: a case study in cardiothoracic surgery planning. Health Care Manag. Sci. **12**, 129–141 (2009)
3. A. Agnetis, A. Coppi, M. Corsini, G. Dellino, C. Meloni, M. Pranzo, A decomposition approach for the combined master surgical schedule and surgical case assignment problems. Health Care Manag. Sci. (2013). doi:10.1007/s10729-013-9244-0
4. J.P. Arnaout, K. Sevag, Maximizing the utilization of operating rooms with stochastictimes using simulation, in *Proceedings of the 40th conference on winter simulation*, pp. 1617–1623 (2008)
5. C. Banditori, P. Cappanera, F. Visintin, A combined optimization–simulation approach to the master surgical scheduling problem. IMA J. Manag. Math. **24**, 155–187 (2013)
6. J. Beliën, E. Demeulemeester, A branch-and-price approach for integrating nurse and surgery scheduling. Eur. J. Oper. Res. **189**, 652–668 (2008)
7. J. Beliën, E. Demeulemeester, Building cyclic master surgery schedules with levelled resulting bed occupancy. Eur. J. Oper. Res. **176**, 1185–1204 (2007)
8. J. Beliën, E. Demeulemeester, B. Cardoen, A decision support system for cyclic master surgery scheduling with multiple objectives. J. Sched. **12**, 147–161 (2009)
9. R. Bellman, S. Dreyfus, *Applied Dynamic Programming* (Princeton University Press, Princeton, NJ, 1962)
10. J. Blake, F. Dexter, J. Donald, Operating room managers' use of integer programming for assigning block time to surgical groups: a case study. Anesth. Analg. **94**, 143–148 (2002)
11. J. Blake, J. Donald, Mount Sinai hospital uses integer programming to allocate operating room time. Interfaces **32**, 63–73 (2002)
12. J.T. Blake, M.W. Carter, A goal programming approach to strategic resource allocation in acute care hospitals. Eur. J. Oper. Res. **140**, 541–561 (2002)
13. S. Brailsford, J. Vissers, OR in healthcare: a European perspective. Eur. J. Oper. Res. **212**, 223–234 (2011)
14. B. Cardoen, E. Demeulemeester, J. Beliën, Operating room planning and scheduling: a literature review. Eur. J. Oper. Res. **201**, 921–932 (2010)

15. B. Denton, J. Viapiano, A. Vogl, Optimization of surgery sequencing and scheduling decisions under uncertainty. Health Care Manag. Sci. **10**, 13–24 (2007)
16. F. Dexter, A strategy to decide whether to move the last case of the day in an operating room to other empty operating room to decrease overtime labor costs. Anesth. Analg. **91**, 925–928 (2000)
17. F. Dexter, A. Macario, R. Traub, Which algorithm for scheduling add-on elective cases maximizes operating room utilization?: use of bin packing algorithms and fuzzy constraints in operating room management. Anesthesiology **91**, 1491–1500 (1999)
18. H. Fei, C. Chu, N. Meskens, Solving a tactical operating room planning problem by a column-generation based heuristic procedure with four criteria. Ann. Oper. Res. **166**, 91–108 (2009)
19. H. Fei, N. Meskens, C. Chu, A planning and scheduling problem for an operating theatre using an open scheduling strategy. Comput. Ind. Eng. **58**, 221–230 (2010)
20. R. Gabel, J. Kulli, B.S. Lee, D. Spratt, D. Ward, *Operating Room Management* (Butterworth-Heinemann, London, 1999)
21. F. Guerriero, R. Guido, Operational research in the management of the operating theatre: a survey. Health Care Manag. Sci. **14**, 89–114 (2011)
22. E. Hans, G. Wullink, M.V. Houdenhoven, G. Kazemier, Robust surgery loading. Eur. J. Oper. Res. **185**, 1038–1050 (2008)
23. P. Harper, A framework for operational modelling of hospital resources. Health Care Manag. Sci. **5**, 165–173 (2002)
24. R. Harris, Hospital bed requirements planning. Eur. J. Oper. Res. **25**, 121–126 (1985)
25. M. Holte, C. Mannino, The implementor/adversary algorithm for the cyclic and robust scheduling problem in health-care. Eur. J. Oper. Res. **226**, 551–559 (2013)
26. M.V. Houdenhoven, J.V. Oostrum, E. Hans, G. Wullink, G. Kazemier, Improving operating room efficiency by applying bin-packing and portfolio techniques to surgical case scheduling. Anesth. Analg. **105**, 707–714 (2007)
27. A. Jebali, A. Hadjalouane, P. Ladet, Operating rooms scheduling. Int. J. Prod. Econ. **99**, 52–62 (2006)
28. P.C. Kuo, R.A. Schroeder, S. Mahaffey, R. Bollinger, Optimization of Operating Room Allocation Using Linear Programming Techniques. The Am. Coll. Surg. **197**, 889–895 (2003)
29. M. Lamiri, X. Xie, A. Dolgui, F. Grimaud, A stochastic model for operating room planning with elective and emergency demand for surgery. Eur. J. Oper. Res. **185**, 1026–1037 (2008)
30. Y. Liu, C. Chu, K. Wang, A new heuristic algorithm for the operating room scheduling problem. Comput. Ind. Eng. **61**, 865–871 (2011)
31. G. Ma, J. Beliën, E. Demeulemeester, L. Wang, Solving the strategic case mix problem optimally by using branch-and-price algorithms, in *International Conference on Operational Research Applied to Health Services (ORAHS)*, vol. 35. Belgium (2009)
32. G. Ma, E. Demeulemeester, A multilevel integrative approach to hospital case mix and capacity planning. Comput. Oper. Res. **40**, 2198–2207 (2013)
33. C. Mannino, E. Nilssen, T. Nordlander, A pattern based, robust approach to cyclic master surgery scheduling. J. Sched. **15**, 553–563 (2012)
34. E. Marcon, F. Dexter, Impact of surgical sequencing on post anesthesia care unit staffing. Health Care Manag. Sci. **9**, 87–98 (2006)
35. I. Moore, D. Strum, L. Vargas, D. Thomson, Observations on surgical demand time series: detection and resolution of holiday variance. Anesthesiology **109**, 408–416 (2008)
36. M.W. Mulholland, P. Abrahamse, V. Bahl, Linear programming to optimize performance in a department of surgery. The Am. Coll. Surg **200**, 861–868 (2005)
37. J.V. Oostrum, M.V. Houdenhoven, J. Hurink, E. Hans, G. Wullink, G. Kazemier, A master surgical scheduling approach for cyclic scheduling in operating room departments. OR Spectrum. **30**, 355–374 (2008)

38. P. Patterson, What makes a well-oiled scheduling system. OR Manag. **12**, 19–23 (1996)
39. A. Sciomachen, E. Tànfani, A. Testi, Simulation Models for optimal schedules of operating theatres. Int. J. Simul. Modell. **6**, 26–34 (2005)
40. M.A. Shalaby, T.F. Abdelmaguid, Z.Y. Abdelrasol, New routing rules for dynamic flexible job shop scheduling with sequence-dependent setup times, *Proceedings of the 2012 International Conference on Industrial Engineering and Operations Management* (Istanbul, Turkey, 2012), pp. 747–756
41. A. Testi, E. Tànfani, G. Torre, A three-phase approach for operating theatre schedules. Health Care Manag. Sci. **10**, 163–172 (2007)
42. A. Testi, E. Tànfani, Tactical and operational decisions for operating room planning: efficiency and welfare implications. Health Care Manag. Sci. **12**, 363–373 (2009)
43. B. Vijayakumar, P.J. Parikh, R. Scott, A. Barnes, J. Gallimore, A dual bin-packing approach to scheduling surgical cases at a publicly-funded hospital. Eur. J. Oper. Res. **224**, 583–591 (2013)
44. J. Vissers, I. Adan, J. Bekkers, Patient mix optimization in cardiothoracic surgery planning: a case study. IMA J. Manag. Math. **16**, 281–304 (2005)
45. R. Wachtel, F. Dexter, Review of behavioral operations experimental studies of newsvendor problems for operating room management. Anesth. Analg. **110**, 1698–1710 (2010)

Chapter 53
Improving the Logistics Operations of the Hospital Pharmacy Using a Barcode-Rfid Identification System

Alejandro Romero

Abstract Medicine represents a critical component of healthcare but impacts negatively the quality of healthcare systems because it faces serious issues such as medical errors, adverse impacts on the environment and the growing presence of counterfeit products. Despite medicines specifications and inefficiencies could contribute to their increasing cost in the healthcare sector, very little research has been published on the main issues affecting the management of medicines and how technology systems could collaborate to improve their management. Within the scope of this paper, we will concentrate on analyzing the main inefficiencies of logistics processes executed by the hospital pharmacy and identifying how an RFID-barcode identification system could improve pharmacy operations. Based on process mapping and qualitative data obtained from semi-structured interviews, this paper describes six main inefficiencies namely incorrect inventory management, medicine shrinkage, intensive manual labor, long procurement cycles, time-consuming product recalls and improper use of technology. By identifying cases and bundles with RFID technologies and primary and secondary packages with barcode labels, hospital pharmacy could ensure that medicines don't affect severely the sustainability of healthcare system in general and the patient security in particular.

Keywords Barcode · Logistics process · Pharmacy hospital · Process mapping · RFID technology · Supply chain management

A. Romero (✉)
Université du Québec à Trois-Rivières, 3351 boul. des Forges, C.P. 500,
Trois-Rivières, QC G9A 5H7, Canada
e-mail: romeroto@uqtr.ca

H. K. Kim et al. (eds.), *Transactions on Engineering Technologies*,
DOI: 10.1007/978-94-017-9115-1_53,

1 Introduction

Healthcare systems around world are facing a period of deep transformations. Governments have to adapt their national health systems to their economic, demographic, technological and organizational contexts. In economic terms, spending on health has known a strong growth. In 2010, United States spent more than 17.4 % of their GDP in medical services while Canada spent 11.7 % [1]. Demographically, the cost of this industry increases with the rise of the elderly population. On the technological background, equipment innovations and more sophisticated medicines permit to improve the quality of healthcare services; however it has imposed important economic costs. On the organizational level, healthcare system of OECD countries are considered as dysfunctional. For example, staff shortages limit the patient's accessibility to health services and actual processes don't allow the optimal use of medical resources. Thus, the quality of health care is seen severely limited [2].

Medicine represents a critical component of healthcare but impacts negatively the quality of healthcare systems because medicines face serious issues such as medical errors, adverse impacts on the environment and the growing presence of counterfeit products. One of the most important issues affecting the sustainability of healthcare service is medicine cost. They represent the third largest budget item for the Canadian health care system, 32 billion in 2011 [3], just after the hospitals and medical staff. An aging population, an over-reliance on prescription medicines as well as an over-utilization of these pharmaceuticals may explain such substantial costs. Even if medicines benefit to health population, their management is not necessarily optimal. Throughout their supply chain, medicines must be manufactured, packaged, distributed, stored, prepared, administered and disposed following rigorous procedures [4, 5]. This leads to major expenses for such management because it requires specialized personnel, sophisticated technologies and control procedures [5]. Despite medicines specifications and inefficiencies could contribute to their increasing cost in the healthcare sector, very little research has been published on the main issues affecting the management of medicines and how technology systems could collaborate to improve their management. This paper attempts to gain a better understanding of this under investigated issue.

Within the scope of this paper, we will concentrate on analyzing the hospital pharmacies processes related to the medicine management. More specifically, it is proposed to analyse the main inefficiencies of logistics processes executed by the hospital pharmacy and identify how an RFID-barcode solution could improve pharmacy operations. The paper is structured as follows. The next section offers a review of medicines processes throughout their supply chain. The research design is exposed in Sect. 3. The main results obtained from process mapping and interviews are presented and discussed in Sect. 4, while the last section offers some concluding remarks.

2 Medicines Management

From the point of view of healthcare sector, medicine management is the ability of healthcare and pharmaceutical organizations to optimize the pharmaceuticals use [6]. Some authors state that an appropriate management of medicines must ensure patient safety in accordance with the five “R”: the right product or the right service to the right patient at the right time using the right way and in the right quantity [6, 7]. Shaeffer [7] and Dumitru [8] add three new “R” to this principle: right communication, a right reason and a right documentation.

In order to better understand the different management processes for medicine use, we have opted to analyze the medicine supply chain into a “product centric” approach. Figure 1 shows a simplified and generic product value chain for medicines that includes the full range of activities from their production to their administration to the patients in the hospital. This paper focuses on logistics processes of the hospital pharmacy (central part of the Fig. 1).

The following sections present the main issues related to the management of medicine at the pharmaceutical supply chain level and at the hospital pharmacy level.

2.1 Medicines Supply Chain

From the healthcare perspective, the supply chain management is characterized “by the information, goods and money necessary to purchase and transfer the goods and services from the supplier to the final user in order to control costs” [9, p. 10]. It is estimated that the healthcare supply chain management spends hundreds of millions of dollars per year [10], which suggests that hospital financial priorities must be re-defined.

The healthcare supply chains are more complex and more immature compared to other industries [9, 11, 12]. This can be explained by different reasons. First, supply chain management has an impact on human health requiring adequate and accurate medical supply conforming to the patients’ needs [13]. If medical supplies are out-of stock, distributed to the wrong patient or are prepared inadequately, patients may experience adverse events, and in some cases death [6, 7]. In fact, it is estimated that one million medication errors occur each year in US [14] accounting for 7,000 deaths and entailing a cost of about $2 billion a year [15]. Second, medical products, medicines and equipment are not totally standardized. Medical professionals are responsible for their selection, but their choice depends on the physical characteristics and health status of each patient [9]. Indeed, they can request different kinds of products for patients undergoing the same treatment. Consequently, several products, medicines and equipment are required, resulting in differentiated and complex health services and generating negative impacts on the hospital finances [9]. Third, hospital operations must deal with a complex distribution network composed of several storerooms and warehouses where different medical supplies are stored following a variety of regulations [16]. Fourth,

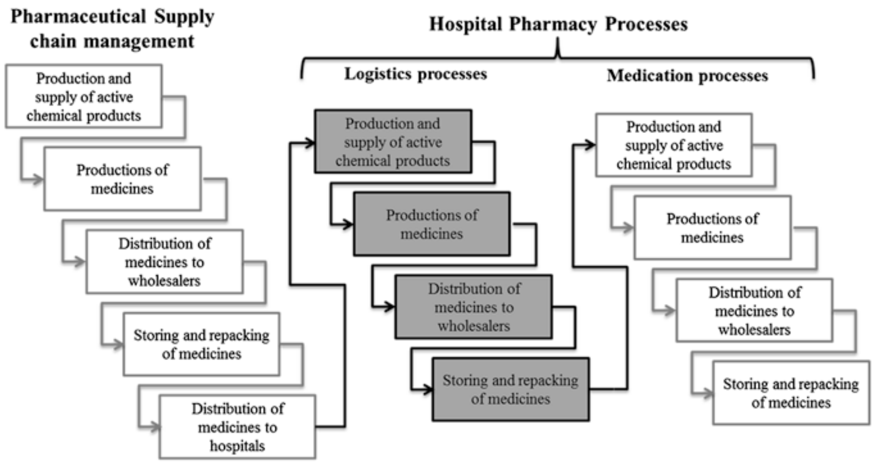

Pharmaceutical Supply chain management
Production and supply of active chemical products
Productions of medicines
Distribution of medicines to wholesalers
Storing and repacking of medicines
Distribution of medicines to hospitals
Hospital Pharmacy Processes
Logistics processes
Medication processes
Production and supply of active chemical products
Productions of medicines
Distribution of medicines to wholesalers
Storing and repacking of medicines
Production and supply of active chemical products
Productions of medicines
Distribution of medicines to wholesalers
Storing and repacking of medicines

Fig. 1 Medicines supply chain

caregivers conduct a staggering number of logistics activities that do not fall under their formal responsibilities. For instance, Landry and Philippe [17, p. 3] estimated that "nursing staff will spend on average 10 % of their time performing logistics tasks instead of taking care of patients, which can not only have cost and care implications, but in countries where there is a shortage of healthcare professionals, social implications as well, such as stress-related diseases." Fifth, healthcare supply chains are characterized by multiple stakeholders that work together in order to ensure the flow of products and services. Inside and outside hospital, medicine management requires a wide variety of human intensive processes which are poorly supported by technology [9, 12]. This results in an increased of workload and a higher possibility of errors. Sixth, healthcare supply chains are high regulated and must respect a number of standards and procedures [5]. In fact, national and international healthcare organizations and government have defined several standards for the distribution, storage, preparation and administration of medical products and materials [4]. Finally, healthcare supply chains are vulnerable to terrorism and criminal facts. According to many observers, this industry experiences a strong possibility of being affected by the presence of counterfeited products [18–20]. From the above-mentioned reasons, one can conclude that healthcare supply chains are indeed inherently complex.

2.2 Hospital Pharmacy Processes

The hospital pharmacy plays a vital role in patient care. It focuses on ensuring that the prescribed medication is precisely and timely dispensed to the intended patient [21]. The hospital pharmacy must purchase, store and distribute medicines. These activities are known as pharmacy logistics processes (central part of the Fig. 1 and focus of this paper), which are under the responsibility of specialized staff because medicines must be managed under specific conditions and standards. Logistics activities include (i) planning of medicine supply, (ii) request of purchase order, (iii) reception of medicines, (iv) validation of package delivery, (v) fitting and sorting of medicine packages, (vi) storage, (vii) preparation for distribution, (viii) distribution of medicines to the primary and secondary pharmacies and to automated equipment, and (ix) reverse logistics.

Hospitals in general and hospital pharmacies in specific look forward to reducing operation costs while ensuring the patient security [22]. However, pharmacy logistics processes are related to several issues that impact negatively the cost and quality of the medication services. Several studies show different inefficiencies, namely out-of-stock [23, 24, 25, 26], high costs [27], excessive manual labour [28, 29], shrinkage [23, 28, 30], high frequency of reorders [23, 28], counterfeit products [31–33] and product recalls [31].

Improving the efficiency of this logistics function is an indispensable option for ensuring the profitability of the healthcare organizations. Past research work shows that hospital pharmacy can adopt several managerial approaches such as Just-In-Time [23, 25], Virtual Inventory [23], Stockless Materials Management Programs or Vendor Managed Replenishment (VMR) [34, 35], Collaborative Planning, Forecasting and Replenishment (CPFR) [25, 29, 35], simulation and outsourcing [28]. These managerial transformations must be supported by information technologies, namely Exchange Data Interchange (EDI) [23, 25, 29], e-commerce [25, 29, 35], barcode and RFID [36, 37].

2.3 Barcode and RFID For Medicine Identification

To improve efficiency, healthcare organisations have focused on developing traceability systems to identify and trace main assets such as medicines. This innovation, also known as track and trace system, could bring important benefits to the hospital pharmacy by enabling real-time and accurate identification of medicines and doses [38]. The implementation of this system requires the mass serialization of medicines, in other words, to assign a unique identification to each medicine and dose [33]. The healthcare entities use two different data carriers to ensure the mass serialisation namely barcode and RFID technology. While barcodes are a mature and well establish technology, RFID technology is described as "one of the ten greatest contributory technologies of the 21st century" [39].

Barcodes have been used for many years in the healthcare sector. Most solutions are focused on the identification of objects (assets and medical supplies) and people (staff and patients). Barcode enables the manual identification of medicines by the lecture of one-dimension or two-dimension barcodes such as Data Matrix [33]. Nevertheless, barcodes present several disadvantage such as low storage capacity, lecture problems and long manual processes. RFID technology has great potential to improve supply chain efficiency [37] because it allows real-time and automated multi-lectures. Despite its technological superiority, the adoption of RFID technology is not as fast as many researchers had expected [37]. The technical limits of RFID technologies, such as cost, lecture interferences and lack of standards, do not allow mitigating the strength of barcodes.

The healthcare sector not only focuses on either barcode or RFID. Indeed, the industry players are assessing the combination of both technology to enable tracking and tracing of medicines. Also known as hybrid solutions, this configuration could overcome the limits of both technologies by completing their advantages [40]. Hybrid solutions fit with the medicine logistic context because hospital pharmacy must manage different level of medicine packages, namely box or case, and primary or secondary package. As shown in the Table 1, cases and bundles of medicines could be labelled with RFID tags while primary and secondary packages could hold a barcode label [41].

Since hybrid solutions have been poorly studied in the literature, this paper aims to assess its capacity to improve logistics inefficiencies of the hospital pharmacy.

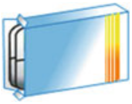
3 Methodology Strategy

3.1 Research Site

Hospital A represents the primary research site but other healthcare entities, government institutions, associations and technology organizations also gave valuable inputs and insights (see Table 1). In total, eighth organizations and 38 healthcare professionals and key managers participated to the field research study.

The hospital A has undertaken a review process program in order to improve their medical services. Hospital pharmacy has been chosen because of its importance and the quantity of different pharmaceutical products, which are received, stored and distributed. Hospital B, C and D are also reviewing their pharmacy processes in order to standardize their activities and ensure a future consolidation. All hospitals involved in this research offer their services in North America. The government institution is involved in different programs for decreasing medication errors by automating the medication process and is also involved in the technological and managerial projects undertaken in hospital A. The pharmacist association represents the perspective of pharmacists and pharmaceutical scientists. The technology provider works with different healthcare organizations in order to develop new equipment and review medical processes. This provider is involved in the projects undertaken in hospital A.

Table 1 Profile of participants

Package level	Description	Illustration	Data carrier
Primary and secondary package	Medicine doses are packed on primary and, in some cases, secondary packaged for its storage in the pharmacy shelves. For administration purposes, this package level is broken down at the medicine dose		Barcode labels
Bundles or cases	For distribution purposes, primary packages are packed on cases or bundles. Before sending medicines to the warehouse, this package level is broken down at the primary package level		RFID tags

3.2 Participants

The vast majority or 76 % of individuals who were interviewed are well aware of the characteristics of medication processes and are involved in the improvement of hospital pharmacy processes. 58 of participants are involved into the logistics function of the hospital pharmacy. In contrast, fewer participants (21 %) are knowledgeable about pharmaceutical supply chain processes. Table 2 shows the profile of participants.

3.3 Data Collection Strategies

We rely on multiple sources of empirical evidence in order to allow triangulation and strengthen the validity of results [42]. Data collection was based on:

(i) Mapping process permits us to analyze the logistics function of the hospital pharmacy and identify the main sources of inefficiencies.
(ii) Multiple on-site observations allowed us to carry out the logistics process mapping.
(iii) Semi-structured interviews were conducted for the validation of the medication process mapping and for the analysis of the different inefficiencies.

The process mapping has been built on the mapping procedures undertaken by the National Health Service Modernization Agency and was based on process flow observations from the field research. The process mapping at the macro level and more detailed level was then validated with key participants in two points in time

Table 2 Profile of participants

	Number of beds	Participants	Number	Related to		
				External logistics	Logistics operation	Medication process
Organization						
Hospital A	630	Chief pharmacist	5		√	√
		Pharmacist	6			√
		IT project manager	2			√
		Material manager	3	√	√	
		Physicians and nurses	6			√
		Pharmacy clerk	4	√	√	
Hospital B	490	Chief pharmacist	2		√	√
		Project manager	1		√	√
Hospital C	320	Chief pharmacist	2		√	√
		Project manager	1			√
Hospital D	230	Chief pharmacist	1		√	√
		Project manager	1		√	
Total			**34**	**7**	**19**	**26**
Other organization						
Govern-ment entity	N.A.	Medical technology director	1		√	√
		Medical project manager	1		√	√
Pharmacy association	N.A.	President	1	√		
Technology provider	N.A.	Project manager	1		√	√
Total			**4**	**1**	**3**	**3**

and served as an anchor for assessing the relative inefficiencies of the hospital pharmacy. In particular, AS-IS process maps were developed for each of the sub-process. The AS-IS process map reflects the actual situation of the pharmacy logistics function. By analyzing the AS-IS process map, we were able to assess the relative issues related to the use of medicine in hospitals. These inefficiencies were then discussed, validated, and compared through three rounds of the semi-structured interviews in order to capture systematically their relative importance across interviewees.

To assess the main advantage of adopting an hybrid barcode-RFID solution for medicine management, we developed TO-BE process maps which represent the process changes generated by the adoption of the hybrid solution. By analyzing and comparing the AS-IS process map and the TO-BE one, we were able to assess the relative benefits derived from this technological innovation. These benefits were also discussed, validated, and compared through three rounds of the semi-structured interviews.

4 Results

4.1 Logistics Inefficiencies

The mapping of the hospital pharmacy logistics processes and the subsequent validations with the key participants have allowed uncovering several inefficiencies of the pharmacy logistics function. Figure 2 shows an example of map process for the delivery of urgent medication to the hospital pharmacy.

This map process gives us important insights of logistics inefficiencies of the pharmacy. There are a total of 10 manual technical tasks or activities which are very time-consuming, inaccurate, inefficient and error-prone processes. Semi-automated and automated activities are less time-consuming, more accurate, more efficient and less error-prone than manual processes. Nevertheless, actual situation for delivery urgent medicine just includes three semi-automated and one automated task. Transportation activities are also time-consuming activities and could introduce several logistics inefficiencies. Technical assistant could lose or incorrectly delivery medicines when they are executing the transportation. In the case of the process map showed in the Fig. 2, medicines must be transported from unloading dock to the pharmacy warehouse before reaching the hospital pharmacy.

Through a thorough content analysis of five different map processes and the comments of the 38 key participants, the following six main issues were identified:

1. *Incorrect inventory management:* It is essential to maintain adequate inventory levels that ensure zero stock outs. One chief pharmacist stated “in most hospitals with clinical operations, a pharmacy inventory of the order of $100 to $200 per bed is considered as reasonable.” However, hospital pharmacies must hold enough medicines to guard against fluctuations in demand, to take advantage of bulk discounts and to withstand fluctuations in supply and, as a result, the pharmacy stocks higher levels of pharmaceuticals than necessary, even if medicines can become obsolete. The pharmacy warehouse clerks manage expiration dates and storage conditions of medicines at the lot level because actual processes cannot support the management of medicine at the primary package level, even less at the unit level. This results in a poor inventory management. A pharmacist pointed out that “even if we are trying to improve the management of our inventories, we may find a few medicines that have expired or will expire soon.” Another pharmacist mentioned that “clerks can update the inventory system with the wrong information resulting in an inexact control of medicines and in an incorrect planning for supplies.” An inappropriate inventory management can produce over- and under-procurement,

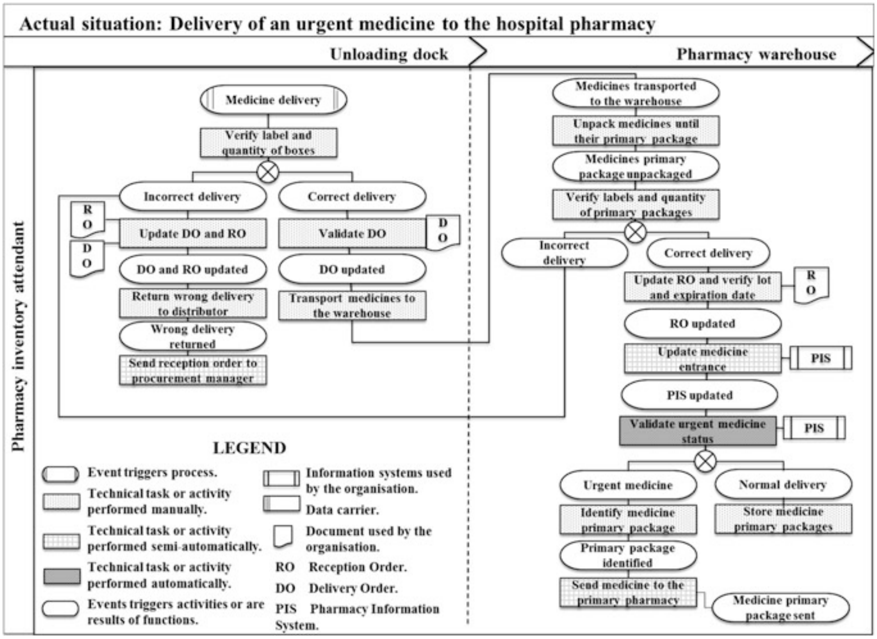

Actual situation: Delivery of an urgent medicine to the hospital pharmacy
Unloading dock
Pharmacy warehouse
Pharmacy inventory attendant
Medicine delivery
Verify label and quantity of boxes
Incorrect delivery
Correct delivery
R O
D O
Update DO and RO
Validate DO
D O
DO and RO updated
DO updated
Return wrong delivery to distributor
Transport medicines to the warehouse
Wrong delivery returned
Send reception order to procurement manager
Medicines transported to the warehouse
Unpack medicines until their primary package
Medicines primary package unpackaged
Verify labels and quantity of primary packages
Incorrect delivery
Correct delivery
Update RO and verify lot and expiration date
R O
RO updated
Update medicine entrance
PIS
PIS updated
Validate urgent medicine status
PIS
Urgent medicine
Normal delivery
Identify medicine primary package
Store medicine primary packages
Primary package identified
Send medicine to the primary pharmacy
Medicine primary package sent
LEGEND
Event triggers process.
Technical task or activity performed manually.
Technical task or activity performed semi-automatically.
Technical task or activity performed automatically.
Events triggers activities or are results of functions.
Information systems used by the organisation.
Data carrier.
Document used by the organisation.
RO Reception Order.
DO Delivery Order.
PIS Pharmacy Information System.

Fig. 2 Map process for delivery of an urgent medicine to the hospital pharmacy

out-of-stock, medicine shortage and multiple and unnecessary storage locations such as storage in care units or in the physicians' and nurses' offices.

2. *Excessive losses*: Poor inventory control may lead to misplaced medicines and to theft. There is however a consensus from the key participants that losses occur mainly from the fact that "medicines could become obsolete before its utilization because they were expired or they were not stored under the proper conditions, such as, for example, in amber, air-tight and moisture-resistant containers."
3. *Intensive manual labour*: Several processes must be executed manually. A technical assistant stated that "approximately 60 % of medicines require repackaging and most of the inventory controls are conducted manually." The same situation is observed for the inventory control. As pointed by one pharmacist "even if pharmacy relies on visual, periodic/cycle counting or perpetual inventory systems, it requires the pharmacist to look manually at the number of units in inventory and compare them with a listing."
4. *Lengthy procurement cycles*: Patients could be affected if their medication doses cannot be delivered in time to the care unit. Several pharmacists mentioned that the procurement cycle might be too lengthy for two main reasons. First, at the reception point, hospital could receive medicines that do not correspond to the purchase order or, in rare but documented cases, could receive altered or counterfeited products. Second, it could take a rather long

time to distribute a medicine from the hospital dock to the care unit if the pharmacy staff cannot properly and immediately identify the medicines.

5. *Time-consuming product recalls*: Due to quality problems or safety issues, pharmaceutical laboratories and governmental agencies could request hospital pharmacies to return some medicines. Known as medicine recalls, this procedure represents a critical issue for the hospital pharmacy due to the complexity of tracing medicines and its associated cost. For recovering medicines, several actions can be undertaken. If the required batch of medicines is still in the pharmacy warehouse (level 3 and 2), the pharmacy clerks place them in a special container. However, if the required batch of medicines has been distributed to the primary and secondary hospital pharmacies or went through automated equipment, the pharmacy clerks and technical pharmacists must retrace the flow of medicines. According to several pharmacists and chief pharmacists, retracing entails high costs since the pharmacy staff must verify manually the inventory of medicines in primary and secondary pharmacies.
6. *Improper technology use*: Even if medicines are delivered to the hospital in cases or in bundles with a linear barcode for its identification, most of the hospitals do not use the same barcode for supporting their internal medicines logistics processes. A pharmacy clerk stated "if he had a barcode reader during the reception and management of boxes, he would use the barcode label for their automatic identification and verification." However, current practices are as follows: at the reception point, the pharmacy clerks identify and manage medicines by reading the label in characters placed on the cases or bundles. Pharmacy clerks validate the received packages by verifying manually their correspondence with the purchase order. A pharmacist indicated: "Several errors may occur with the manual verification because clerks can confuse or misinterpret the medicine information. Unfortunately, these errors are identified once medicines are distributed to the primary or secondary pharmacies."

4.2 Combining RFID and Barcode into a Hybrid Solution

Potential advantages of adopting RFID and barcode for medicine identification were analyzed by developing project map process (TO-BE). Figure 3 shows the TO-BE process map of implementing RFID identification for boxes and cases of medicines while barcode labels for primary and secondary packages.

By comparing AS-IS and TO-BE map process for the delivery of a urgent medicine to the hospital pharmacy, we assessed that RFID tags on cases or bundles allow accurate, real-time and automatic identification because manual technical task are reduced from 10 to 4 while automated activities increased from 1 to 7. Therefore, this hybrid solution brings opportunity to reduce the intensive manual labor and errors during reception and inventory management.

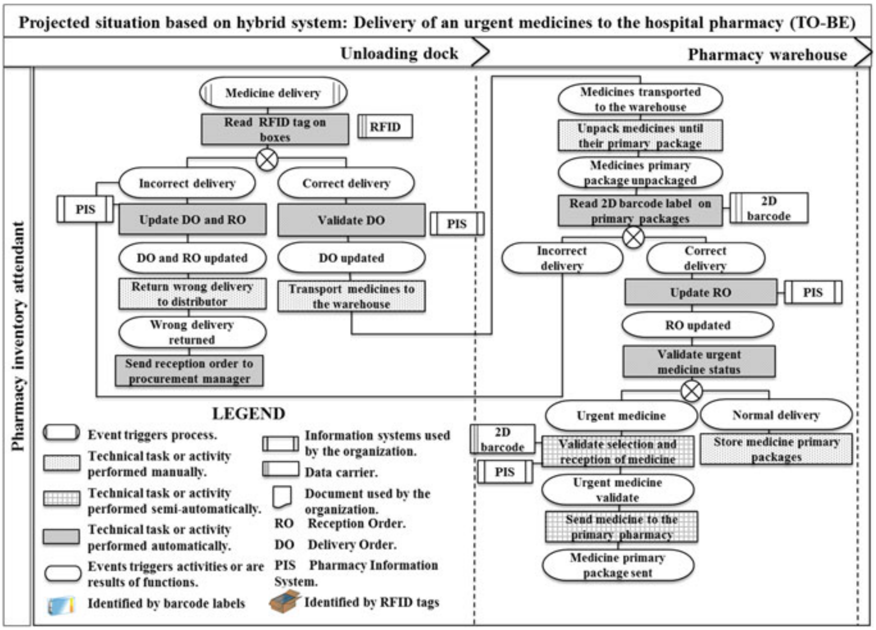

Projected situation based on hybrid system: Delivery of an urgent medicines to the hospital pharmacy (TO-BE)
Unloading dock
Pharmacy warehouse
Pharmacy inventory attendant
Medicine delivery
Read RFID tag on boxes
RFID
Incorrect delivery
Correct delivery
PIS
Update DO and RO
Validate DO
PIS
DO and RO updated
DO updated
Return wrong delivery to distributor
Transport medicines to the warehouse
Wrong delivery returned
Send reception order to procurement manager
Medicines transported to the warehouse
Unpack medicines until their primary package
Medicines primary package unpackaged
Read 2D barcode label on primary packages
2D barcode
Incorrect delivery
Correct delivery
Update RO
PIS
RO updated
Validate urgent medicine status
Urgent medicine
Normal delivery
2D barcode
Validate selection and reception of medicine
Store medicine primary packages
PIS
Urgent medicine validate
Send medicine to the primary pharmacy
Medicine primary package sent
LEGEND
Event triggers process.
Technical task or activity performed manually.
Technical task or activity performed semi-automatically.
Technical task or activity performed automatically.
Events triggers activities or are results of functions.
Identified by barcode labels
Information systems used by the organization.
Data carrier.
Document used by the organization.
RO Reception Order.
DO Delivery Order.
PIS Pharmacy Information System.
Identified by RFID tags

Fig. 3 Map process of the project hybrid solution for the delivery of an urgent medicine to the hospital pharmacy

If the technical assistants control the information placed on primary packages that are then packed into the cases or bundles (see Table 1) and send the information to the hospital, the hospital pharmacy would gain a more granular and a quick update of medicines at the reception and for their inventories. This feature could decrease time consuming verification during product recalls and increase the control of medicines for avoiding losses. A pharmacist stated that more granular and accurate records could be obtained if pharmacy staff controls, by scanning barcode labels. This way, the pharmacy could accelerate cycle times when searching an urgent medicine or when conducting product recalls.

5 Conclusions

Based on qualitative and quantitative methods, this study identifies from the reception of medicines at hospital docks to their distribution to the hospital pharmacy an important number of inefficiencies, namely incorrect inventory management, medicine shrinkage, intensive manual labor, long procurement cycles, time-consuming product recalls and improper use of technology. These inefficiencies affect severely the sustainability of healthcare system in general and the patient security in particular. Track and trace systems could decrease the impact of the above described inefficiencies. Combining barcode levels to identify primary and secondary packages of medicines with RFID tags for bundles and cases identifications permits to accurately manage and control medicines resulting in a better inventory management, less medicines losses and decrease manual labor, long procurement cycles and recall activities.

Several lines of actions can be envisioned. First, healthcare organisations are faced with huge losses estimated to be roughly at least one-tenth of their revenues, with increased medical errors and with tighter regulations. Efforts have been made to strengthen the efficiency of the pharmacy logistics function but the knowledge, analysis and process review have to be improved. Logistics activities could also be streamlined. In particular, existing relabeling and repackaging practices need to be closely examined in order to avoid medical errors. Second, technology can help by ensuring the efficiency of logistics processes. Track and trace systems for medicines could bring interesting benefits in order to improve inventory management, decrease procurement cycles and automate time-consumers processes. Nevertheless, the observed inefficiencies cannot be totally resolved by a track and trace system. Hospital pharmacy logistics function must be supported with other information technologies, the redesign of business processes, and the commitment of hospital staff.

References

1. OECD, Manuel d'Oslo: principes directeurs pour le recueil et l'interprétation des données sur l'innovation. Les éditions de l'OCDE, 3e édition (2005)
2. OECD, OECD Health Data 2011, Health Expenditure and Financing, 2011, http://stats.oecd.org/index.aspx?DataSetCode=HEALTH_STAT
3. ICIS, *Tendances de dépenses nationales en santé,* 1985–2011, Institut Canadien d'information sur la santé, Ottawa (Ont.) (2012)
4. M. Potdar, E. Chang,V. Potdar, Applications of RFID in pharmaceutical industry, Presented at *IEEE International Conference on Industrial Technology,* pp. 2860–2865, (2009)
5. Y. Meiller, S. Bureau, Logistics Projects: How to assess the right system? The case of RFID solution in healthcare, in *Proceedings Americas Conference on Information Systems (AMCIS).* California, p. 14 (2009)
6. F. Boulet, Les erreurs médicamenteuses ou l'épée de Damoclés. Pharmactuel **34**(6), 161–165 (2001)
7. R. Shaeffer, Closing the medication safety loop. Comput. Healthc. **30**(3), 30–32 (2009)
8. D. Dumitru, *The Pharmacy Informatics Primer*, Bethesda, Maryland: American Society of Health-System Pharmacists, p. 251, (2009)
9. E. Schneller, L. Smeltzer, L. Burns, *Strategic Management of the Health Care Supply Chain* (Jossey-Bass, San Francisco, Calif, 2006)
10. Ontario Buys & Healthcare Supply Network, Supply Chain Modernization in Ontario Health Care, Improving Patient Care, Enhancing Service Levels and Reducing Costs: A Report on the E-Supply Chain Project. *Ontario Ministry of Finance*, Toronto, report, 2007
11. J. Langabeer, *Health Care Operations Management: A Quantitative Approach to Business and Logistics* (Jones and Bartlett Publishers, Sudbury, 2007)
12. N.H. Mustaffa, A. Potter, Healthcare supply chain management in Malaysia: a case study. Supply Chain Manag. Int. J. **14**(3), 234–243 (2009)

13. F.J. Beier, The management of the supply chain for the hospital pharmacies: a focus on inventory management practices. J. Bus. Logistics **16**(2), 153–173 (1995)
14. G.J. Kuperman, A. Bobb, T.H. Payne, Medication- related clinical decision support in computerized provider order entry systems: a review. J. Am. Med. Inf. Assoc. **14**(1), 29–40 (2007)
15. L. Kohn, J. Corrigan, M. Donaldson, *To err is human: building a safer health system* (National Academy Pr., 2000)
16. H. Rivard-Royer, S. Landry, M. Beaulieu, Hybrid Stockless: a case study, lessons for Healthcare supply chain integration. Int. J. Oper. Prod. Manag. **22**(4), 412–424 (2002)
17. P.S. Landry, R. Philippe, *4U2C or How Logistics can Service Healthcare* (Ecole des Hautes Commerciales, Montreal, Quebec, 2002)
18. M. Brooks, K. Button, Market Structures and Shipping Security. Marit. Econ. # 38; Logistics, **8**, 00–120 (2006)
19. S. Palaniswami, L. Jenicke, P. Okonkwo, H. Kang, Risk assessment and security measures in supply chain networks. Int. J. Procurement Manag. **3**(1), 1–11 (2010)
20. R. Srinivasan, Supply chain immunity: a methodology for risk management. Int. J. Serv. Sci. **3**(1), 1–20 (2010)
21. C.A. Pedersen, P.J. Schneider, D.J. Scheckelhoff, ASHP national survey of pharmacy practice in hospital settings: dispensing and administration—2008. Am. J. Health-Syst. Pharm. **66**(10), 926–946 (2009)
22. J. Scott-Cawiezell, R.W. Madsen, G.A. Pepper, A. Vogelsmeier, G. Petroski, D. Zellmer, Medication safety teams guided implementation of electronic medication administration records in five nursing home. Jt. Comm. J. Qual. Patient Saf. **35**(1), 29–35 (2009)
23. K. Danas, P. Ketikidis, A. Roudsari, A virtual hospital pharmacy inventory: an approach to support unexpected demand. J. Med. Mark. Device Diagn. Pharmacetical Mark. **2**(2), 125–128 (2002)
24. D.S. West, Purchasing and inventory control, ed. by R. Jackson. *Effective Pharmacy Managem*ent, 9th edn. (National Community of Pharmacists Association, Sec. 17, Alexandria, VA, 2003)
25. L. Breen, H. Crawford, Improving the pharmaceutical supply chain: Assessing the reality of e-quality through e-commerce application in hospital pharmacy. Int. J. Qual. Reliab. Manag. **22**(6), 572–590 (2004)
26. K. Dongsoo, An integrated supply chain management system: a case study in healthcare sector, in *proceedings of the E-commerce and Web Technologies: Sixth International Conference*, (Copenhagen, Denmark, 2005)
27. D.S. West, *NCPA=Pharmacia Digest* (National Community of Pharmacists Association, Alexandria, VA, 2002)
28. M.D. Rosseti, D. Marek, S. Prabhu, A. Bhonsle, S. Sharp, Y. Liu, *Inventory management issues in health care supply chains*, Center of innovation in healthcare logistics, (2008). http://cihl.uark.edu/Inventory_Management_Issues_in_Health_Care_Final.pdf
29. H. Dreyer, J. Strandhagen, A. Romsdal, A. Hoff, principles for real-time, integrated supply chain control: an example from distribution of pharmaceuticals. Adv. Prod. Manag. Syst. Challenges, Approaches, **1**, 187–194 (2010)
30. A.R. Vila-Parrish, J.S. Ivy, R.E. King, A simulation-based approach for inventory modeling of perishable pharmaceuticals. *Simulation Conference, 2008. WSC 2008*. Winter, pp. 1532—1538 (2008)
31. E. Schuster, S. Allen, D. Brock, *Global RFID: The Value of the EPCglobal Network For Supply Chain Management.* (Springer, Berlin, 2007)
32. N. Basta, *Product Security Perspective: Protecting the Brand, Pharmaceutical Commerce*. Pharmaceutical Commerce website, (2008). http://www.pharmaceuticalcommerce.com/frontEnd/991-serialization_anticounterfeitin_pedigree_RFID_taggants_barcode.html
33. E. Lefebvre, A. Romero, L.-A. Lefebvre, C. Krissi, Technological strategies to deal with counterfeit medicines: the European and North-American perspectives, Int. J. Edu. Inf. Technol. **5**(3), 275–284 (2011)

34. S. Landry, R. Philippe, How logistics can service healthcare. Supply Chain Forum **5**(2), 24 (2004)
35. C. Chandra, The case for healthcare supply chain management: insights from problem-solving approaches. Int. J. Procurement Manag. **1**(3), 261–279 (2008)
36. J.F. Bussières, D. Lebel, Utilisation des code-barres dans le cadre du circuit du médicament en établissement de santé. Pharmactuel **42**(2), 131–137 (2004)
37. E. Jones, M. Henry, D. Cochran, T. Frailey, RFID pharmaceutical tracking: from manufacturer through in vivo drug delivery, J. Med. Devices **4,** 015001-1-015001-7 (2010)
38. A. Romero, Managing medicines in the hospital pharmacy: logistics inefficiencies,in *Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering and Computer Science 2013, WCECS 2013,* San Francisco, USA, pp. 1120–1125 23–25 Oct 2013
39. C.-C. Chao, Y. Jiann-Min, J. Wen-Yuan, Determining technology trends and forecasts of RFID by a historical review and bibliometric analysis from 1991 to 2005. Technovation **27**(5), 268–279 (2007)
40. A. Romero, E. Lefebvre, L.-A.Lefebvre, Breaking the Barcode and RFID myth: Adoption paths for improving the medication process. WSEAS J. Comput. Commun. **4**(5), 223–235 (2011)
41. A. Romero, E. Lefebvre, Combining barcode and RFID into a hybrid solution for improving the hospital pharmacy logistics processes, Int. J. Inf. Technol. Manag. Forthcoming (2014)
42. R.K. Yin, *Case study Research: Design and Methods, Newbury Park* (Sage Publications, CA, 2003)

Chapter 54
Algorithms for the Circle-Packing Problem via Extended Sequence-Pair

Shuhei Morinaga, Hidenori Ohta and Mario Nakamori

Abstract The circle-packing problem is a problem of packing circles into a two dimensional area such that none of them overlap with each other. Each of the former methods has its own difficulty; some of them are only applicable to the case that the area the circles are to be packed into has a special shape; some of them require different search technique according as the shape of the area. Also, most of the former methods search in a restricted neighbor. In addition, there exist unsearchable location of circles. These facts mean former methods cannot assure global optimization. Hence, in the present paper, we propose sequence-pair for circle packing (SPC), a method of representing relative location of circle pairs, which is an extended version of sequence-pair for rectangles. We propose also a method of obtaining an approximate solution of the circle-packing problem, where all constraints are replaced by approximate linear inequalities.

Keywords Circle-packing · Linear approximation · Linear programming · Non-linear programming · Packing density · Reducing redundancy · Sequence-pair

1 Introduction

The circle-packing problem is a problem of packing circles into a two dimensional area such that none of them overlap with each other. This problem is NP-hard and has a wide variety of application, e.g., fiber packing in a tube or transportation of

S. Morinaga (✉) · H. Ohta · M. Nakamori
Tokyo University of Agriculture and Technology, 2-24-16 Naka-cho, Koganei-shi
Tokyo 184-8588, Japan
e-mail: 50012646136@st.tuat.ac.jp

H. Ohta
e-mail: ota@fjlab.ei.tuat.ac.jp

M. Nakamori
e-mail: nakamori@cc.tuat.ac.jp

H. K. Kim et al. (eds.), *Transactions on Engineering Technologies*,
DOI: 10.1007/978-94-017-9115-1_54,

pipes by a ship, since they are equivalent to the problem of packing rigid cylinders.

One idea of solving the circle-packing problem is to formulate the problem as a nonlinear programming problem and solve it by some nonlinear optimization solver. This is an ideal method, because it assures the exact optimal solution especially if the given circles are of the same size [1]. If the sizes of the circles are not the same, however, the constraints are often very complex, and obtaining the optimal solution in practical computational time is almost impossible [5].

Thus, it is widely considered to be practical to obtain a quasi-optimal solution rather than the exact optimal solution for the case that the sizes of the circles are not the same. Most of existing methods are based on heuristic search that locates circles sequentially; some of them are followed by relocation via beam search or simulated annealing [2–5].

Each of the above methods, however, has its own difficulty; some of them are only applicable to the case that the area the circles are to be packed into has a special shape; some of them require different search technique according as the shape of the area. Also, most of the above methods search in a restricted neighbor. In addition, there exist unsearchable location of circles. These facts mean the above methods cannot assure global optimization.

Apart from the circle-packing problem, many promising algorithms have been proposed for the rectangle packing problem. There are two main streams in the existing rectangle packing algorithms; locating sequentially rectangles and locating via relative position. The boundary method [6] belongs to the former, whereas the sequence-pair method [7] belongs to the latter.

It is a natural extension to apply these methods to the circle-packing problem. As previously mentioned, most of the existing methods of the circle packing are based on sequentially locating. That is, these methods are classified into the extension of the boundary method, which cause the above difficulties.

Hence, in the present paper, we propose sequence-pair for circle packing (SPC), a method of representing relative location of circle pairs, which is an extended version of sequence-pair for rectangles. We propose also a method of obtaining an approximate solution of the circle-packing problem, where all constraints are replaced by approximate linear inequalities. Further, we propose searching various methods using simulated annealing. The preliminary version was presented at [8, 9].

This paper constitute as follows: in Sect. 2 we introduce SPC for circle representation; in Sect. 3 we propose algorithms to obtain dense packing solution from SPC code; in Sect. 4 we report the computational result; in Sect. 5 we conclude and discuss the related topics.

2 Circle Arrangement Representation Method: SPC

Our problem is to locate given circles in a convex region such that all circles do not overlap with each other. We denote the center of circle and radius of circle i by (x_i, y_i) and r_i. Since any circles a and b do not overlap, we have

$$\sqrt{(x_a - x_b)^2 + (y_a - y_b)^2} \geq (r_a + r_b). \tag{1}$$

In order to represent relative location of circle pairs by two sequences of circle names, we adopt sequence-pair for circle packing (SPC) [8]. We denote an SPC as (Γ_+, Γ_-), where Γ_+ and Γ_- are sequences of circles. Given an SPC, the relative location of circles are determined as follows:

- If two circles appear in the same order in Γ_+ and Γ_-, i.e., both Γ_+ and Γ_- are of the form $(\ldots, a, \ldots, b, \ldots; \ldots, a, \ldots, b, \ldots)$, then centers of circles a and b satisfy

$$x_a \leq x_b, \tag{2}$$

$$y_a \leq y_b. \tag{3}$$

- If two circles appear in the opposite order in Γ_+ and Γ_-, i.e., both Γ_+ and Γ_- are of the form $(\ldots, a, \ldots, b, \ldots; \ldots, b, \ldots, a, \ldots)$ then centers of circles a and b satisfy

$$x_a \leq x_b, \tag{4}$$

$$y_a \geq y_b. \tag{5}$$

It is possible to transform the relative location among circles by the above Γ_+ and Γ_- to a grid representation in the same way to synthesize the relative location among rectangles in sequence-pair representation through oblique grid [7]. A grid representation of SPC is a rotated form by 45° clockwise of the oblique grid for a sequence-pair. An example of the grid representation for an SPC (132456; 214635) is shown in Fig. 1a.

Suppose a vertex a in the grid representation is the origin of x and y axes. Then centers of circles corresponding to the vertices in the first (second, third, or fourth) quadrant are in up-right (up-left, down-left, or down-right, respectively) of the center of the circle a. Thus, from Fig. 1a, we can see that the centers of circles 5 and 6 are up-right, the center of circle 3 is up-left, and the centers of circles 1 and 2 are down-left of circle 4.

SPC can represent any location of circles. In a similar way as sequence-pair we can obtain from any SPC the corresponding consistent constraints of relative location among circles. Unlike sequence-pair, however, we have to make elaborate computation of mathematical optimization in order to obtain the most dense location of circles, since x and y direction of the constraints of relative location corresponding to an SPC. Especially the constraint (1) that prevents overlapping of circles is nonlinear, our problem of obtaining the most dense location of circles is that of nonlinear optimization. Figure 1b shows the most dense location of circles corresponding SPC (132456; 214635).

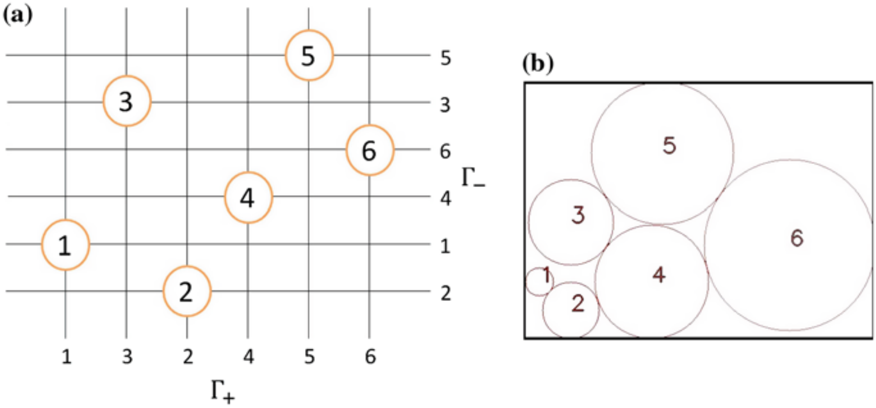


Fig. 1 Grid representation and location of circles corresponding SPC (132456; 214635)

3 Search Method of the Circle Arrangement Using SPC

Since we are going to search the optimal location of circles by simulated annealing as is often done for sequence-pair, we have to decode SPC as quick as possible. Therefore, we propose methods of performing a search in practical time by devising a search method to reduce the constraints in this section.

3.1 Search Methods Using SPC

We propose algorithms to be able to solve within practical time as follows.

3.1.1 Linear Approximation of Constraints

We obtain an approximate optimal location corresponding the given SPC by linear programing solver by expressing constraints with linear inequality.

Let us denote by $\theta_{i,j}$ the angle of the line passing through the centers of circles i ad j and the x axis, as shown in Fig. 2. Using this notation, we can translate constraints (2), (3), (4), and (5) with (1) as

$$(r_a + r_b)\cos(\theta_{a,b}) \le x_b - x_a, \tag{6}$$

$$(r_a + r_b)\sin(\theta_{a,b}) \le y_b - y_a. \tag{7}$$

Note that if circles a and b appear in the same order in Γ_+ and Γ_-, i.e., $(\ldots, a, \ldots, b, \ldots; \ldots, a, \ldots, b, \ldots)$, then $0 \le \theta_{a,b} \le \frac{\pi}{2}$; if circles a and b appear in

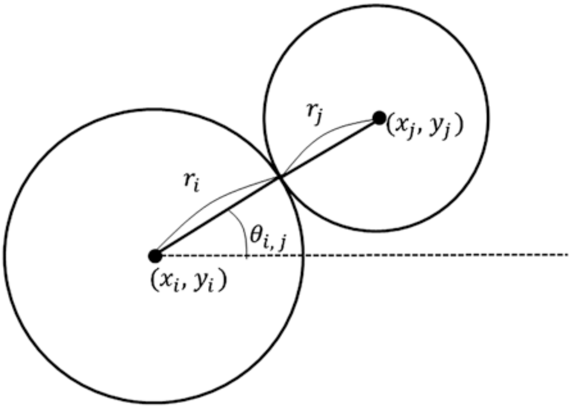


Fig. 2 Angle $\theta_{i,j}$: the angle of the line passing through the centers of the circles *i*,*j*

the reverse order in Γ_+ and Γ_-, i.e., $(\ldots, a, \ldots, b, \ldots; \ldots, b, \ldots, a, \ldots)$, then $-\frac{\pi}{2} \leq \theta_{a,b} \leq 0$.

Now we approximate trigonometric function by piecewise linear function in order to apply linear programming. The idea is as follows. Let ϑ_1 satisfy $0 \leq \theta \leq \frac{\pi}{2}$.

1. When $0 \leq \theta \leq \theta_1$, we replace $\sin\theta$ by following Eq. (8).

$$f_1(\theta) = \alpha_1 \cdot \theta + \beta_, \tag{8}$$

2. When $\theta_1 < \theta \leq \frac{\pi}{2}$, we replace $\sin\theta$ by following Eq. (9).

$$f_2(\theta) = \alpha_2 \cdot \theta + \beta_2. \tag{9}$$

Figure 3 shows an instance of approximation of $\sin\theta$ with $\theta_1 = \frac{\pi}{4}$.

In order to select appropriate approximation we introduce a 0–1 variable. For example, the above (8) and (9) are expressed as

$$\begin{aligned} M \cdot P + \theta - \theta_1 \geq 0, \\ M \cdot (1 - P) + f_1(\theta) - \alpha_1 \cdot \theta - \beta_1 \geq 0, \\ M \cdot (1 - P) - f_1(\theta) + \alpha_1 \cdot \theta + \beta_1 \geq 0, \\ M \cdot (P - 1) + \theta_1 - \theta \geq 0, \\ M \cdot P + f_2(\theta) - \alpha_2 \cdot \theta - \beta_2 \geq 0, \\ M \cdot P - f_2(\theta) + \alpha_2 \cdot \theta + \beta_2 \geq 0, \end{aligned}$$

where $P = 1$ implies $0 \leq \theta \leq \theta_1$, and $P = 0$ implies $\theta_1 < \theta \leq \frac{\pi}{2}$. Also, M is a sufficiently large number. In the same way we can approximate $\cos\theta$ as linear constraints. If we use more 0–1 variables, we can approximate nonlinear function by

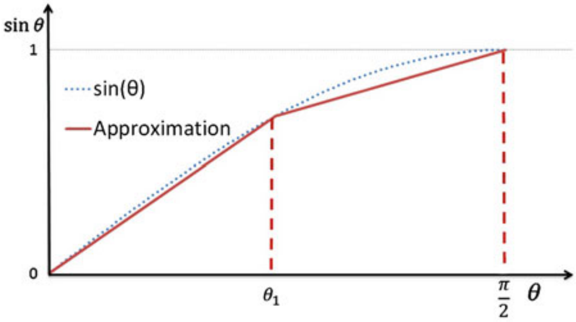


Fig. 3 Linear approximation of sin θ

linear constraints with the range of variable into more than two subsets, to obtain more accurate approximation.

In the present paper we call the way of dividing uniformly the range of variable as m division. When the number m of division grows, the approximation will be more accurate, but at the same time the number of constraints becomes large and computational time required grows larger.

3.1.2 Circle Packing in SPC by Nonlinear Optimization

We proposed in Sect. 3.1.1 a quick method a quick method of obtaining an approximate optimal solution of the circle-packing problem, where all constraints are replaced by approximate linear inequalities. This method, however, does not always give a feasible solution. In this section we propose an algorithm using nonlinear optimization to remove the defects of the above approximation.

Since existing nonlinear optimization algorithms require considerable amount of computational time as compared as linear optimization, we first search by simulated annealing on linear approximation. This does not always give a feasible solution, so next we execute nonlinear optimization, which removes infeasibility and will give a nearly optimal solution in practical computational time.

It is known that appropriate initial solution will accelerate nonlinear optimization algorithm. Thus, after linear approximation search we do not make use of only the SPC code but also the coordinate data of location of circles.

As mentioned before, different circles do not overlap in an SPC representation yields constraints (2), (3), (4), (5), and (1).

3.2 Reducing Constraints

In order to reduce the computation time, we propose methods to reduce the number of constraints (Fig. 4, Table 1).

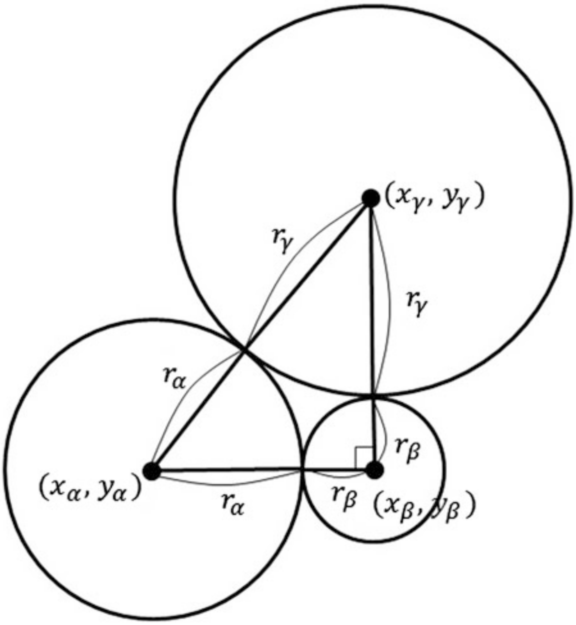


Fig. 4 The case that constraints for the pair of circles α, γ are not necessary

Table 1 Difference of calculation time for linear optimization

Number of circles	10 (s)	15 (s)	20 (s)
$m = 3$ (Redundant constraints eliminated)	0.44	0.75	1.05
$m = 3$	0.60	1.43	2.73
$m = 4$ (Redundant constraints eliminated)	0.90	3.50	9.46
$m = 4$	1.35	6.03	26.30

3.2.1 Elimination of Redundant Constraints

When for three circles α, β, and γ such that SPC $(\ldots, \alpha, \ldots, \beta, \ldots, \gamma; \ldots, \alpha, \ldots, \beta, \ldots, \gamma, \ldots)$ or SPC $(\ldots, \alpha, \ldots, \beta, \ldots, \gamma; \ldots, \gamma, \ldots, \beta, \ldots, \alpha, \ldots)$ satisfy

$$r_\alpha + r_\gamma \leq \sqrt{(r_\alpha + r_\beta)^2 + (r_\beta + r_\gamma)^2},$$

then constraints for the pair of circles α, γ are not necessary, because these constraints are transitive result of those for the pair of circles α, β and those for the pair of circles β, γ Fig. 4. Therefore, we can eliminate these redundant constraints to make the computational time short Table 1.

3.2.2 Method of Deleting Constraints for Distant Circle Pairs

Constraints of distant circle pairs have no influence on the solution. It is expected, therefore, that better solution will be obtained under limited computational time by removing redundant constraints.

In order to find redundant constraints we introduce distance coefficient $D_{distance}$. If inequality 10 is satisfied for circle a and b, we neglect constraints for circles *a* and *b*. We execute this operations for all pair of circles and search by simulated annealing on linear approximation and next make nonlinear optimization. We repeat this process by gradually increasing $D_{distance}$ until feasible solution is obtained.

$$\sqrt{(x_a - x_b)^2 + (y_a - y_b)^2} \geq D_{distance} \times (r_a + r_b). \tag{10}$$

Unnecessary constraint is removed by this method. Then, it is possible to obtain a higher density packing.

4 Computational Experiment

As for numerical experiment we carried out our circle packing algorithm. As for linear optimization solver we used CPLEX 12.4 and as for nonlinear optimization we used quasi Newton method. Our computational environment includes CPU (Intel Core i7-2600 @ 3.4 GHz) and memory of 4 GB.

The neighborhood of search by simulated annealing in SPC was obtained by random selection from the following three operations:

1. choose two circles at random and exchange the position in Γ_-;
2. choose two circles at random and exchange the position in Γ_+;
3. choose two circles at random and exchange the position both in Γ_+ and in Γ_-.

As for the linear approximation of trigonometric function mentioned in 1, we tried two types of approximation as we have done in our former research [8]. Figure 5a, b show these two types of approximation. Note that there results small overlap among circles by approximation 2, because sine curve is approximated with a little underestimation.

4.1 Circle Packing by Linear Approximation Method

In order to evaluate the performance of our algorithm we tried to pack circles in a rectangle and an equilateral triangle.

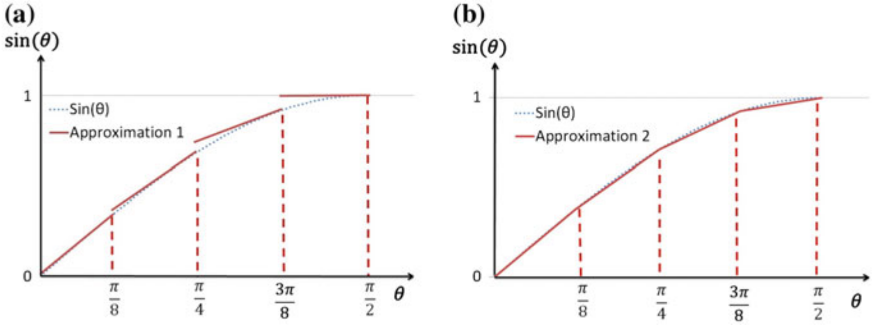


Fig. 5 Approximation 1 and approximation 2

4.1.1 Circle Packing in a Rectangle with Fixed Width *W*

Let us consider circle packing in a rectangle with fixed width *W*. This problem is known as strip packing problem. Constraints are

$$r_i \leq x_i \leq W - r_i,$$
$$r_i \leq y_i \leq H - r_i,$$

for every circle *i* and the objective function is *H* to be minimized. Problem instances are from benchmark CODP: SY2 [10].

Figure 6 shows the correspondence between calculation time and packing density for circle packing in a rectangle. Both approximation give good solutions after long calculation time.

Figure 7a shows a location by approximation 1 with packing density 77.23 %, and Fig. 4b shows a location by approximation 2 with packing density 84.06 %.

Approximation 1, however, causes redundant space between circles, whereas approximation 2 causes overlap of circles.

4.1.2 Circle Packing in an Equilateral Triangle

Let us consider circle packing in an equilateral triangle. Constraints are

$$\begin{aligned} r_i \leq y_i &\leq \sqrt{3}x_i - 2r_i, \\ r_i \leq y_i &\leq -\sqrt{3}x_i - 2r_i + Y, \end{aligned} \tag{11}$$

for each circle *i*, and the objective function is *Y* to be minimized. Problem instances are from benchmark CODP: SY2.

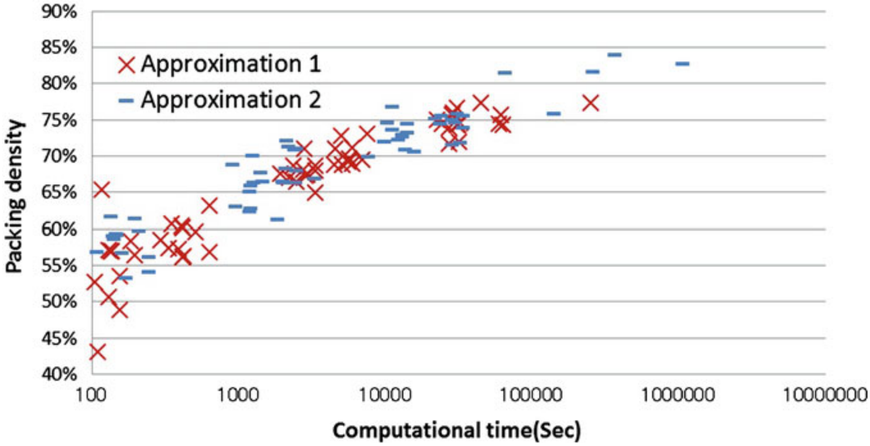


Fig. 6 Comparison of calculation time and packing density for circle packing in a rectangle

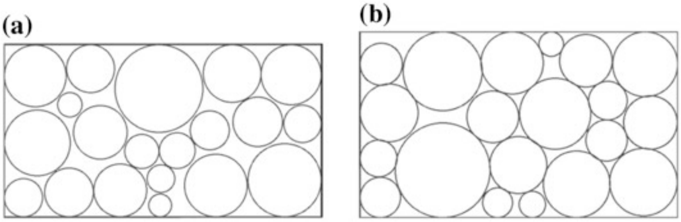


Fig. 7 Circle packing in a rectangle by **a** approximation 1 (*calculation time* 253655 s, *packing density* 77.23 %), **b** approximation 2 (*calculation time* 197865 s, *packing density* 84.06 %)

Figure 8 shows correspondence between calculation time and packing density. Figure 9a shows a location by approximation 1 with packing density 75.93 %, and Fig. 9b shows a location by approximation 2 with packing density 81.83 %.

Similar result of the difference of approximation is observed as in the case of packing in a rectangle.

4.2 Circle Packing by Nonlinear Optimization

Linear approximation method is a quick method of obtaining an approximate optimal solution of the circle-packing problem, where all constraints are replaced by approximate linear inequalities. This method, however, does not always give a feasible solution.

Next experiment improve solution using nonlinear optimization to remove the defects of the above approximation.

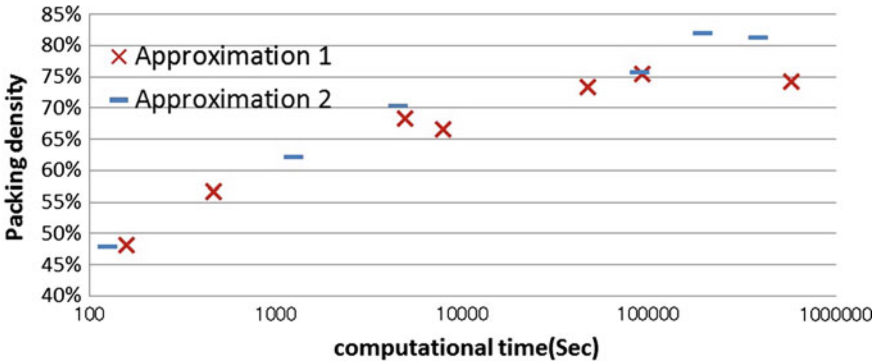


Fig. 8 Comparison of calculation time and packing density for circle packing in an equilateral triangle

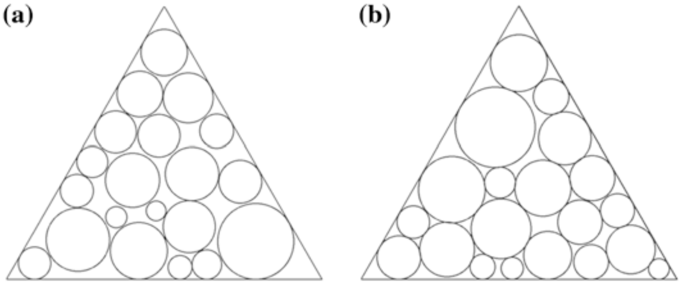


Fig. 9 Circle packing in an equilateral triangle by **a** approximation 1 (*calculation time* 71581 s, *packing density* 75.93 %), **b** approximation 2 (*calculation time* 18501 s, *packing density* 81.83 %)

4.2.1 Circle Packing in a Rectangle with Fixed Width *W*

Figure 10a shows a location by approximation 1 with packing density 77.91 %, and Fig. 10b a location obtained by nonlinear optimization with packing density 79.81 % where Fig. 10a is used as an initial solution. Figure 11a shows a location by approximation 2 with packing density 84.06 %, and Fig. 11b a location obtained by nonlinear optimization with packing density 79.96 % where again Fig. 8a is used as an initial solution. We can observe that in Fig. 10b redundant space resulted by approximation 1 is removed and in Fig. 11b overlap of circles resulted by approximation 2 is removed.

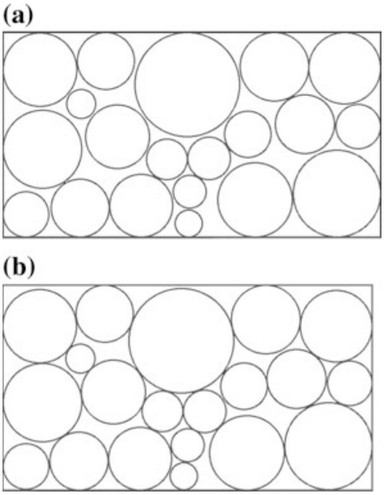


Fig. 10 Circle packing in a rectangle by approximation 1 and nonlinear optimization **a** *packing density* 77.91 %, **b** *packing density* 79.81 %

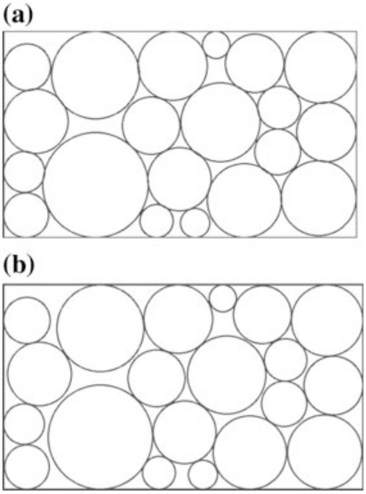


Fig. 11 Circle packing in a rectangle by approximation 2 and nonlinear optimization **a** *packing density* 84.06 %, **b** *packing density* 79.96 %

4.2.2 Circle Packing in an Equilateral Triangle

Figure 12a shows a location by approximation 1 with packing density 77.96 %, and Fig. 12b a location obtained by nonlinear optimization with packing density 79.82 % where Fig. 12a is used as an initial solution. Figure 13a shows a location by approximation 2 with packing density 81.78 %, and Fig. 13b a location obtained by nonlinear optimization with packing density 80.51 % where again Fig. 13a is used as an initial solution. Similar result of the difference of approximation is observed as in the case of packing in a rectangle.

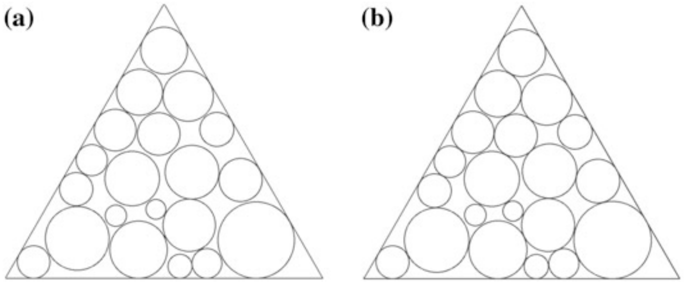


Fig. 12 Circle packing in an equilateral triangle by approximation 1 and nonlinear optimization **a** *packing density* 77.915 %, **b** *packing density* 79.819 %

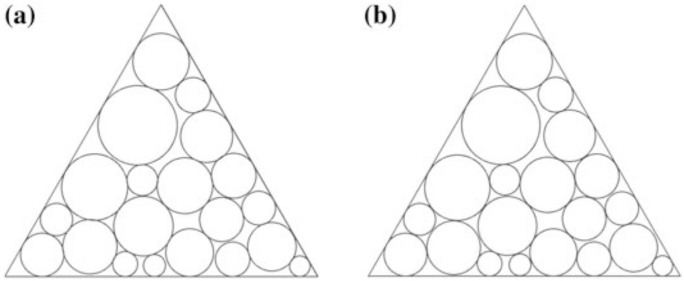


Fig. 13 Circle packing in an equilateral triangle by approximation 2 and nonlinear optimization **a** *packing density* 81.78 %, **b** *packing density* 80.51 %

4.3 A method Deleting Constraints of Circle Pairs Far Away

We made experiment on improvement of solution by deleting constraints of distant circle pairs as described in Sect. 3.2.2.

4.3.1 Circle Packing in a Rectangle with Fixed Width *W*

Table 2 shows comparison of packing density for circle packing in a rectangle. Figure 14 shows the output an example of the results. Packing density is increased for every SPC code. There is, however, one SPC code where packing density remains unchanged under reduced constraints, which means deleting redundant constraints has no effect on packing density.

Table 2 Comparison of packing density: circle packing in a rectangle

Nonlinear optimization (%)	Reduced constraints (%)	Reduction ratio (%)
80.0063	80.0190	99.98
81.7359	82.5008	99.07
78.8566	78.8571	100.00
79.3622	79.6832	99.60
70.0725	73.4027	95.46

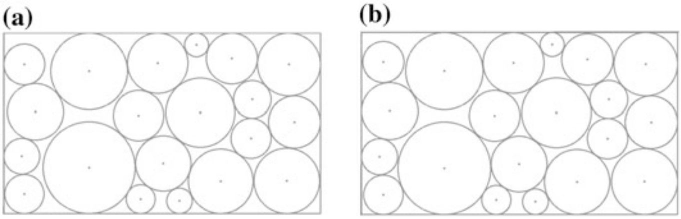


Fig. 14 Circle packing in a rectangle and deleting constraints **a** *packing density* 81.7359 %, **b** *packing density* 82.5008 %

Table 3 Comparison of packing density: circle packing in an equilateral triangle

Nonlinear optimization (%)	Reduced constraints (%)	Reduction ratio (%)
79.6339	79.6339	100.00
80.5609	80.5609	100.00
74.0812	74.8219	99.50
69.0736	70.6136	98.90
61.0309	65.9983	96.16

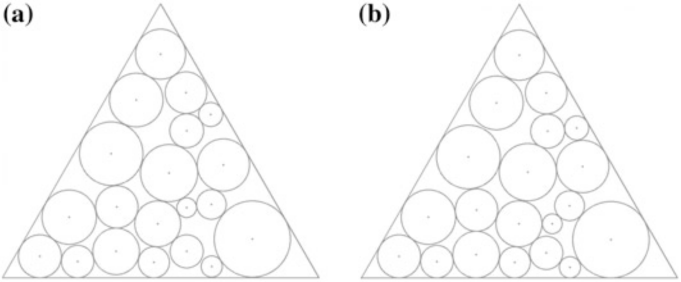


Fig. 15 Circle packing in an equilateral triangle and deleting constraints **a** *packing density* 74.0812 %, **b** *packing density* 74.8219 %)

4.3.2 Circle Packing in an Equilateral Triangle

Table 3 shows comparison of packing density for circle packing in an equilateral triangle. Figure 15 shows the output an example of the results. Similar result is observed as in the case of packing in a rectangle.

5 Conclusions

In this paper, we proposed Sequence-pair for circle packing (SPC), a method of representing relative location of circle pairs, which is an extended version of sequence-pair for rectangles. Further, we proposed searching various methods for determining the dense SPC code using Simulated annealing.

Computational experiments show that our algorithms give dense and feasible packing in a rectangle and an equilateral triangle. We compared several ways of approximation followed by nonlinear optimization on packing density and computational efficiency.

Remained problems left for further research are more efficient algorithms, packing various forms of objects including circles and rectangles, and three dimensional packing.

References

1. E.G. Birgin, J.M. Gentil, New and improved results for packing identical unitary radius circles within triangles, rectangles and strips. Comput. Oper. Res. **37**, 1318–1327 (2010)
2. H. Wang, W. Huang, Q. Zhang, D. Xu, An improved algorithm for the packing of unequal circles within a larger containing circle. Eur. J. Oper. Res. **141**, 440–453 (2002)
3. D. Zhang, A. Deng, An effective hybrid algorithm for the problem of packing circles into a larger containing circle. Comput. Oper. Res. **32**, 1941–1951 (2005)
4. H. Akeb, M. Hifi, R. M'Hallah, A beam search algorithm for the circular packing problem. Comput. Oper. Res. **36**, 1513–1528 (2009)
5. C.O. Lo′pez, J.E. Beasley, A Packing unequal circles using formulation space search. Comput. Oper. Res. **40**, 1276–1288 (2013)
6. T. Sawa, A. Nagao, T. Kambe, I. Shirakawa, K. Chihara, A method for rectangle packing problem. Inst. Electr. Inf. Commun. Eng. **97**(137), 159–166 (1997)
7. H. Murata, K. Fujiyoshi, S. Nakatake, Y. Kajitani, VLSI module placement based on rectangle-packing by the sequence-pair. IEEE Trans. CAD **15**(12), 1518–1524 (1996)
8. S. Morinaga, H. Ohta, M. Nakamori: A method for solving circle-packing problem by using extended seqence-pair, in *The 26th Workshop on Circuits and Systems*, pp. 489–494, 2013
9. S. Morinaga, H. Ohta, M. Nakamori: An algorithm for the circle-packing problem via extended sequence-pair with nonlinear optimization, lecture notes in engineering and computer science, in *Proceedings of The World Congress on Engineering and Computer Science 2013, WCECS 2013*, San Francisco, USA, pp. 1222–1227 23–25 Oct. 2013
10. E. Specht, packomania. www.packomania.com. Accessed 02 July 2013

Chapter 55
Integration of Knowledge-Based Approach in SHM Systems

Fazel Ansari, Inka Buethe, Yan Niu, Claus-Peter Fritzen and Madjid Fathi

Abstract This paper discusses the potentials for integration of knowledge-based techniques in Structural Health Monitoring (SHM). Knowledge-based techniques and methods reinforce health assessment and influence on predictive maintenance of structures. A concept of the knowledge-based approach is developed. In particular, toolboxes for a simple numerical 3-degree of freedom (dof)-model and for force reconstruction at Canton Tower are implemented, respectively. The case studies deepen the insight into identifying needs in the field of SHM to employ knowledge-based approaches, especially in the reasoning process. The proposed concept lays the ground for future research in the field of SHM for utilizing knowledge-based methods in correlation with SHM algorithms and analysis of feedbacks obtained from sensors, engineering expertise and users former experience. The foresight is to broaden the scope for applying Knowledge Management (KM) techniques and methods towards developing a decision-making component for supporting SHM systems, and in turn fostering the detection, localization, classification, assessment and prediction.

F. Ansari (✉) · M. Fathi
Institute of Knowledge-Based Systems and Knowledge Management,
University of Siegen, 57076 Siegen, Germany
e-mail: fazel.ansari@uni-siegen.de

M. Fathi
e-mail: fathi@informatik.uni-siegen.de

I. Buethe · Y. Niu · C.-P. Fritzen
The Group of Applied Mechanics, Institute of Mechanics and Control Engineering-Mechatronics, University of Siegen, Siegen, Germany
e-mail: inka.buethe@uni-siegen.de

Y. Niu
e-mail: yan.niu@uni-siegen.de

C.-P. Fritzen
e-mail: fritzen@imr.mb.uni-siegen.de

H. K. Kim et al. (eds.), *Transactions on Engineering Technologies*,
DOI: 10.1007/978-94-017-9115-1_55,

Keywords Force reconstruction · Integration · Knowledge-based systems · Knowledge management · Reasoning process · Structural health monitoring

1 Introduction

1.1 Structural Health Monitoring (SHM)

Today a variety of structures coming e.g. from civil, aeronautic, astronautic or, mechanical engineering sector need to be monitored. According to [1], approximately one third of all bridges in the US national inventory need either be repaired or replaced. The monitoring is not only expensive, but also – depending on the structure – hard to be realized at all times e.g. after an earthquake or a typhoon. The emerging research area Structural Health Monitoring (SHM) assesses the state of structural health continuously or periodically in an automated way via direct measurements and through appropriate data processing and interpretation in a diagnostic system. Advanced SHM systems also allow the prediction of the remaining useful lifetime of the structures. The automated SHM system involves the sensors and their integration, the signal conditioning unit and data storage as well as data evaluation unit and automated diagnosis (see [1] and [2]).

SHM features a variety of sensing and data evaluation techniques, based on different physical principles and mathematical approaches. In general, it can be separated into passive and active methods. Active methods always introduce some sort of excitation into the structure. The sensing of its interaction with the structure is used to gain information about the monitored object. Unlike active methods, passive methods do not require an additional excitation, but use ambient excitation like wind or traffic load. This way these methods investigate operational parameters and accordingly reveal the structures' state. Additional information is then needed to predict the remaining useful lifetime. SHM systems therefore can also be categorized into five subsequent levels: Detection, Localization, Classification, Assessment and Prediction [3].

Usage monitoring is the process of measuring responses of, and in some cases the inputs to, a structure while it is in operation [4]. The damage prognosis (DP) process attempts to forecast system performance by assessing the current damage state of the system (i.e. SHM), estimating the future loading environments, and predicting through simulation or statistical models as well as past experience, the remaining useful lifetime of the system, see [5] and [6].

An example of a technique used in SHM is the identification of external loads. In many practical cases, it is difficult or not possible to directly measure the external force applied to the structure. One solution is to reconstruct the external force from the measured structural responses, e.g. acceleration. This process is often an ill-posed inverse problem, in the sense that small noise in the measurements may cause large deviation in the reconstruction result. Force identification is the process of solving this type of an inverse problem and provides a stable estimate of the external force.

1.2 Knowledge-Based Systems (KBS)

Already the general overview on SHM, DP and usage monitoring reveals the potentials for developing and integrating knowledge-based approaches in the context of these kinds of systems. The knowledge-based approach for SHM requires effective deployment of knowledge sources, and therefore uses semantics to establish the relation between entities and to create a Knowledge Base (KB). Here, the common attribute is "Knowledge" for reasoning (i.e. reasoning under uncertainty). The knowledge-based approach is derived from the principles of Knowledge-Based Systems (KBS) [7, 8]. KBS consists of four core components: (1) Knowledge acquisition, (2) Knowledge representation, (3) Knowledge modeling, and (4) Dialogue system (User interface).

"Knowledge acquisition is the extraction of knowledge from sources of expertise and its transfer to the knowledge base" [9]. Knowledge acquisition methods are classified into three categories as: (1) Manual e.g. by means of interview, analysis of protocols, observation, case studies, brainstorming, etc. (2) Semi-automatic by direct support and influence of domain experts, (3) Automatic by minimizing or eliminating the role of experts or knowledge engineers. The automatic method to knowledge acquisition is ideal. However, the realization of an automatic solution depends on the advancement of the algorithms for transferring all required knowledge into the system. Therefore the semi-automatic method is the most used one.

In the context of KBS, the knowledge engineer is responsible to construct the system. The knowledge engineer contributes to domain experts who hold expertise in the problem domain (e.g. SHM). Using manual or semi-automatic methods, the knowledge engineer has a considerable role to be in contact with domain experts to submit questions, data and problems and to receive knowledge, concepts and solutions to be delivered to KB. KB is a warehouse which contains the required knowledge for formulating and solving problems with formalized structure [8, 9]. It provides the means for collection, organization, and retrieval of knowledge through establishing semantics between entities.

Knowledge representation includes two major forms, declarative or procedural [7, 8]. In declarative representations, knowledge stored as facts that must be interpreted, and it is accessible for a variety of purposes, while in procedural representations, knowledge stored as algorithms (program code), and it is usable only within specialized problem-solving contexts [7, 8]. In comparison, procedural representation is highly efficient in the correct context. Knowledge representation methods are e.g. predicate logic, rules, frames and scripts, or semantic networks.

Knowledge modeling consists of procedures for working with the knowledge stored in the KB, and mapping of the knowledge to fulfill certain tasks. It supports the acquisition and structuring of knowledge, formalization of knowledge for building KB, processing for solving a problem, e.g. using inference engine, visualizing of the knowledge and so on [7, 8].

Moreover, a dialogue system utilizes a variety of knowledge sources and models. It is the main communication interface between users and system. The dialogue system is realized in the form of a graphical user interface which encompasses a structure and mechanism in the back-end for interpreting the entries of the user, matching them with the representation structure, and accordingly to provide an understandable output to the user.

In the context of SHM, reasoning is a critical endeavor due to reliability and safety requirements of the structures. Some SHM studies already consider advanced decision making components, including a variety of factors and combining them e.g. via votes [10]. Nevertheless, many SHM applications use algorithms of damage detection isolatedly, which is not an optimal solution and could not consider all influential factors of decision-making in a right time and place. To this extent, knowledge-based approaches to SHM can improve the situation through utilizing and integrating various knowledge sources, combining SHM algorithms, and deploying knowledge acquisition, representation and modeling methods.

In the area of fault detection and isolation of automatic control systems already various approaches, taking into account knowledge-based systems, existing examples can be found in e.g. [11] and [12].

In this paper, we adopt the knowledge-based approach in the field of SHM by focusing on modeling rules. It reflects the first results on the conception of knowledge-based approach for a SHM system. Thus the primary objective is to develop the concept of reasoning in the context of SHM especially health assessment. In this way, the case study of force reconstruction is employed. The secondary objective is to foreground and highlight the promising potentials of this merging, using Knowledge Management (KM) approaches in future research.

1.3 Knowledge Management (KM)

Developing a KM solution requires effective deployment of knowledge technologies (i.e. semantic technologies, information systems or Mindware and Software Engineering) towards the evolution of adaptive and intelligent systems for industrial applications. The objective of integrated technologies and components for KM is to make information actionable and reusable. As explained, here, the common attribute is “Knowledge” for reasoning (i.e. reasoning under uncertainty), and decision-making by discovering hidden relations between knowledge attributes, and improving reliability of decisions. Such an approach requires the assurance of reliability of the data via the prioritization of different sensing and data evaluation techniques up to the reasoning and final decision-making.

Several accounts are detectable for defining KM. This is due to the generic nature of KM and its use in various fields such as, but not limited to, (strategic) management, computer science, psychology, education, quality management, information technology, and organizational learning.

Not only in academia but also in practice the variety of accounts on KM is visible, where e.g. managers interchangeably use Information Management (IM) and KM. The reason is that definitions of knowledge, itself, can differ, comparing application domains and sectors or even branches of a sector. Therefore, for instance, one can define an entity as information while at the same time it is used as knowledge by others. The definition of Groff and Jones explains KM more comprehensively [13]: "KM is taken as the tools, techniques and strategies to retain, analyze, organize, improve and share organizational expertise". This is also addressed by Wijnhoven [14], Eppler [15] and Maier [16]. Despite the diversity of accounts, KM is generally defined in computer science as an iterative, life cyclic, dynamic and systematic process which encompasses the creation, acquisition, extraction, storage, retrieval, discovery, application, review, sharing and transfer of the knowledge captured and stored in databases. In this context, several components should be integrated to manage e.g. databases (including non-homogenous types of documentations and data), soft/hard competences, human resources, experiences, quality and change of processes, and associated risks.

The term knowledge refers to a certain typology for distinction between tacit, explicit and latent knowledge. Tacit knowledge is a person-dependent knowledge (personal knowledge). This type of knowledge is not and cannot be expressed. Explicit knowledge "is or could be expressed without attenuation" [14]. Latent knowledge "could be expressed but it is difficult to express it without attenuation" [17]. In practice, knowledge is mostly seen as explicit or implicit. In this way, knowledge is classified into two major categories by identifying whether knowledge is represented, documented and codified or not. Particularly, undocumented or non-codified knowledge is considered as implicit, i.e. tacit or latent that needs to be extracted, documented or discovered using certain methodologies like experience management, observation, interview, etc. The given definition of KM only addresses the major aspects, and in turn KM needs to be redefined or adopted for each application domain e.g. SHM. Such definition should consider organic and functional relations between components and entire KM system. Considering various approaches, integrating KM in SHM deals with three major issues, but not limited to:

(1) Better accessibility to information which also includes access and reuse of documents, e.g. through deploying data management technologies or search engines.
(2) Reinforcing knowledge discovery from databases, e.g. through combining existing algorithms.
(3) Improving decision-making activities, e.g. through provision of evidences and adopting preferences.

Decision-making ultimately effects on quality of products, processes or services [18] and [19]. It is, in most of cases, a complex, time consuming, problematic activity, and requires simultaneous and systematic consideration of decision-parameters, risk factors, customer preferences, priorities and associated cost/benefit ratios [18] and [19]. n this context, knowledge-based approaches to decision-making can improve the situation through utilizing and integrating various knowledge sources, combining SHM algorithms, and deploying learning and intelligent methods for detection of semantics between instances and parameters of decision-making.

2 Knowledge-Based Approach in SHM Systems

In this section, two case studies, a simple numerical 3-dof-model used for Structural Damage Detection and force identification, are presented respectively. Besides, the conception for the integration of the knowledge-based approach for reasoning is proposed. The concept is developed based on the need analysis through discussion with the SHM's domain experts.

2.1 Case Study of Structural Damage Detection with a Numerical 3-dof-Model

A simple example, including the data of a 3-dof-numerical model (Fig. 1), visualizes the use of the Knowledge-Based Approach in SHM for Structural Damage Detection on a very simple level. The external database consists of the calculated system response as displacement of one of the masses after white noise excitation on the second mass.

The SHM database consists of data of the undamaged system, as well as of three different scenarios, which have been modelled. These include a temperature change, resulting in a slight stiffness change of all included springs, a sensor fault, modelled as a bias on the calculated displacement as well as a damage of the structure, modelled as a local stiffness reduction of the spring between the second and the third mass. All different scenarios are not visible in the time domain data, but with the help of different SHM algorithms, knowledge about the structures state can be extracted.

Simple algorithms have been included in the SHM toolbox, using statistical values, like mean, variance, using Autoregressive (AR) coefficients and eigenfrequencies extracted from the time domain data. For all three algorithms, the obtained features were combined with the help of Principal Component Analysis in a Damage Index [20]. Afterwards, these three damage indices are used in the Knowledge-based Decision Toolbox to extract knowledge from the information

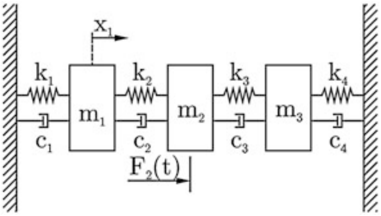
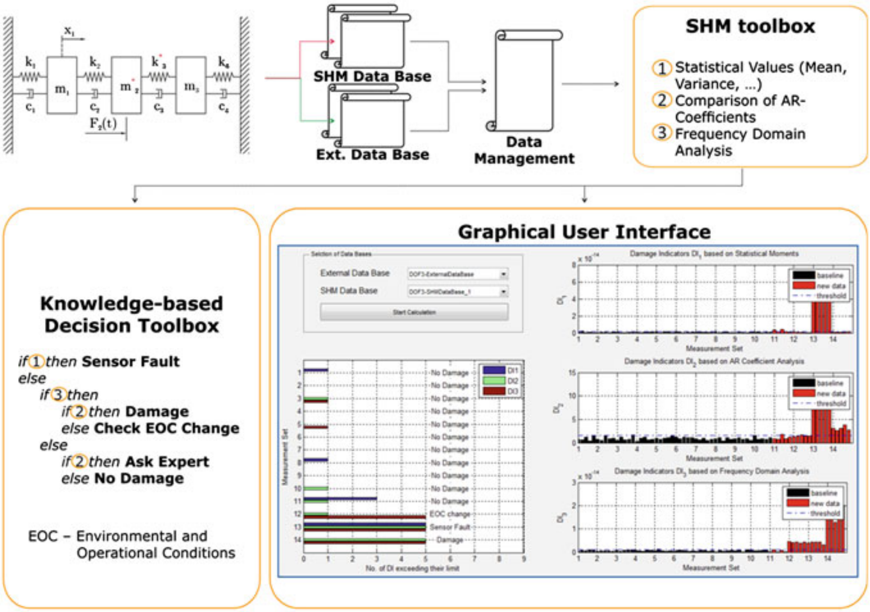


Fig. 1 Numerical 3-dof-model, used as structure for a simple case study


SHM Data Base
Ext. Data Base
Data Management
SHM toolbox
1 Statistical Values (Mean, Variance, ...)
2 Comparison of AR-Coefficients
3 Frequency Domain Analysis
Knowledge-based Decision Toolbox
Graphical User Interface
Selection of Data Bases
External Data Base
DOF3-ExternalDataBase
SHM Data Base
DOF3-SHMDataBase_1
Start Calculation
Damage Indicators DI_1 based on Statistical Moments
Damage Indicators DI_2 based on AR Coefficient Analysis
Damage Indicators DI_3 based on Frequency Domain Analysis
baseline
new data
threshold
Measurement Set
No. of DI exceeding their limit
DI1
DI2
DI3
No Damage
EOC change
Sensor Fault
Damage

```
if 1 then Sensor Fault
else
    if 3 then
        if 2 then Damage
        else Check EOC Change
    else
        if 2 then Ask Expert
        else No Damage
```

EOC – Environmental and Operational Conditions

Fig. 2 Integration of a knowledge-based approach in SHM of a simple 3-dof-structure (*EOC = environmental and operational conditions*)

included in these Damage Indices. A graphical User Interface makes this accessible to a wide audience, already giving some interpretation of the data, but also supplying all necessary information for an own interpretation of the data.

With the use of this knowledge-based approach all modelled scenarios can be identified, while none of the algorithms alone would be able to identify these, because the single algorithms respond differently to data of various system states like sensor fault or damage. A visualization of this example is given in Fig. 2.

2.2 Case Study of Force Identification

The knowledge of external loading conditions is crucial for both SHM and future loading estimation. For high-rise structures, e.g. offshore wind turbines and tall buildings, wind load is a major concern in SHM. Even though an anemometer can provide information on the wind speed and wind direction at a specified height, the wind load is not only stochastic in amplitude and direction, but distributed in space, which makes a direct measurement not economical or not feasible. A potential solution is to reconstruct the loading history information from the structural response measurements, i.e. displacement, strain, velocity or acceleration.

The 600 m tall Canton Tower is located in a typhoon active area, and a long-term SHM system has been designed and integrated into this tower [21]. These two aspects make the Canton Tower an ideal test-bed for the wind load reconstruction study using the field measurement data from the SHM measurement system. Figure 3 schematically shows a methodology adopted for the wind load reconstruction study [22]. In stage 1, an operational modal analysis (OMA) for the Canton Tower is performed using the previously recorded field measurements. According to the obtained OMA results, the finite element model (FEM) of the Canton Tower can be updated to better represent the dynamics of the real structure. Based on the updated FEM of the Canton Tower, a state-space model in modal coordinates can be constructed. In stage 2, given the field measurements from a new wind event, e.g. a typhoon, and the updated state-space model from stage 1, the wind load can be reconstructed by using an input estimation algorithm. In this study, the measurements from 20 accelerometers and 1 anemometer installed on the tower are used. These measurements were previously recorded and stored in the data warehouse of the SHM system for the Canton Tower.

Based on this methodology, an SHM toolbox for wind load reconstruction has been developed. As an example, the measurements recorded during the Typhoon “Kai-tak” in 2012 were taken out of the data warehouse and imported into the toolbox for analysis. Before passing the recorded data to the input estimation algorithm, a check of data quality is first performed by plotting the data from each sensor in the form of time history and power spectral density (PSD). In such a manner, the data sections with a discontinuity or abnormal outliers can be separated from the analysis. Figure 4 shows a screenshot of the toolbox in checking the data quality of accelerometer channel 20. It can be seen that the measured time history is continuous and no abnormal outlier is noticed. Then the sensor channels with good data quality can be selected, the FEM of the Canton Tower can be loaded and the wind load can be reconstructed with the help of the input estimation algorithm. Figure 5 shows the mean and standard deviation of the reconstructed wind load with respect to the nodes of a reduced-order finite element model (FEM) of the Canton Tower.

Such reconstructed wind load information may assist the health status assessment of the structure, e.g. material fatigue after the typhoon event, and the decision making process in maintenance scheduling.

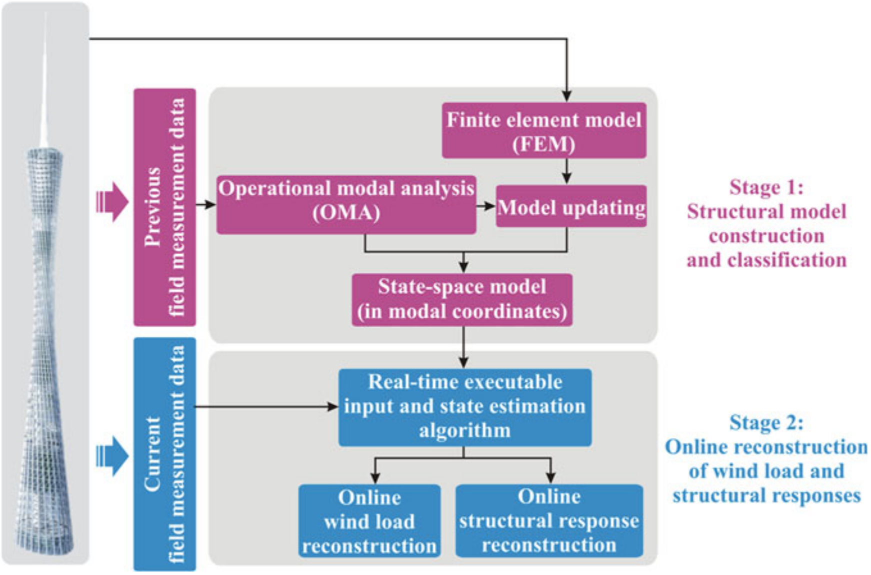
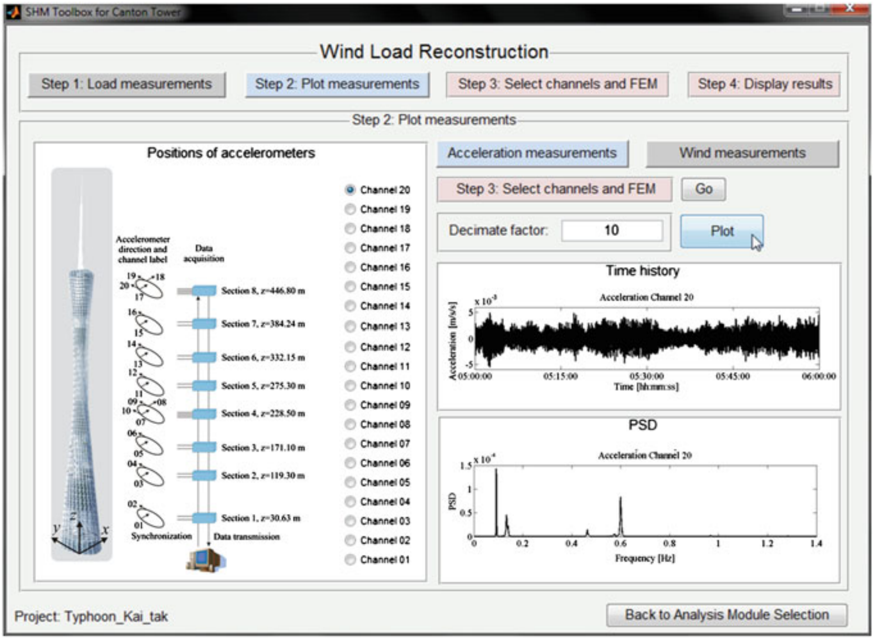

Previous field measurement data
Finite element model (FEM)
Operational modal analysis (OMA)
Model updating
State-space model (in modal coordinates)
Stage 1: Structural model construction and classification
Current field measurement data
Real-time executable input and state estimation algorithm
Online wind load reconstruction
Online structural response reconstruction
Stage 2: Online reconstruction of wind load and structural responses

Fig. 3 Online simultaneous wind load and structural response reconstruction methodology for high-rise structures [22]



Decimate factor: 10

Project: Typhoon_Kai_tak

Fig. 4 SHM toolbox for wind load reconstruction—plot measurements [23]

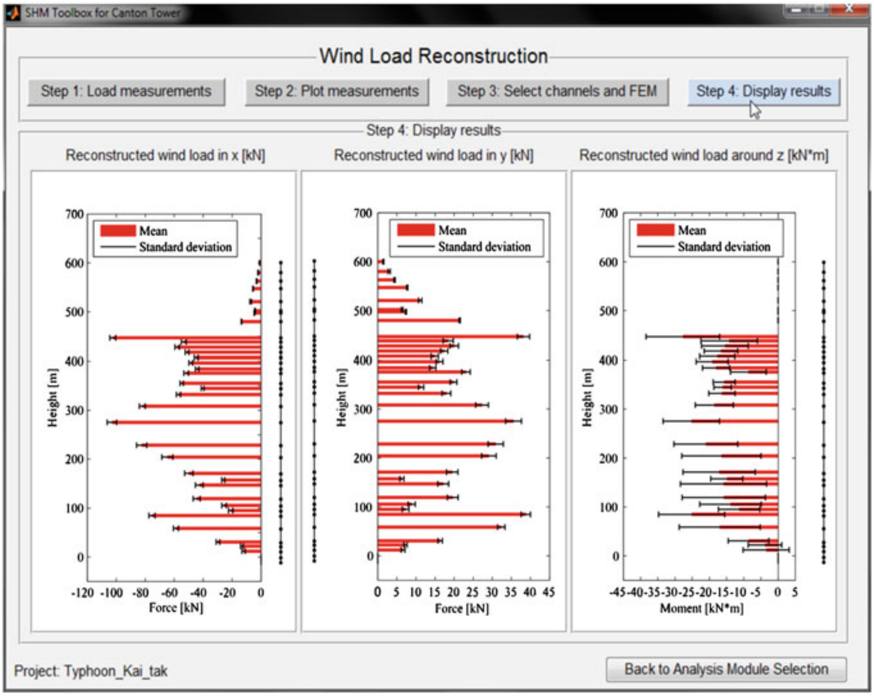


Fig. 5 SHM toolbox for wind load reconstruction—display results [23]

2.3 Conception of the Knowledge-Based Approach

Synergetic use of sensor data for health assessment and reconstruction of wind load information raise the great potentials for developing the knowledge-based approach for reasoning. As shown in Fig. 6, sensor data which are captured from a structure, e.g. Canton Tower, are used for health assessment.

Recorded data of the sensors are stored in the data warehouse. The sensor fault detection component is used to improve the quality of data through analysis of the sensor readings which are unmatched with expected values. The stored data are, therefore, filtered with the help of sensor information from the sensor fault detection component. This ensures that only reliable data are used in the wind load reconstruction process. The force reconstruction toolbox provides information for simulating the behavior of the tower subject to wind loads. It is used as the main channel of information.

The core part of the knowledge-based approach is the rule-based engine. It consists of the mechanism for modeling and updating the rules, inference machine and rule bank. The knowledge engineer requires a dialogue system (user interface) to access the KB, and to modify (ongoing) rules in the form *IF a(i) is A(i) AND b(i) is B(i) THEN C.* Of course, there will not be always two factors in the condition of

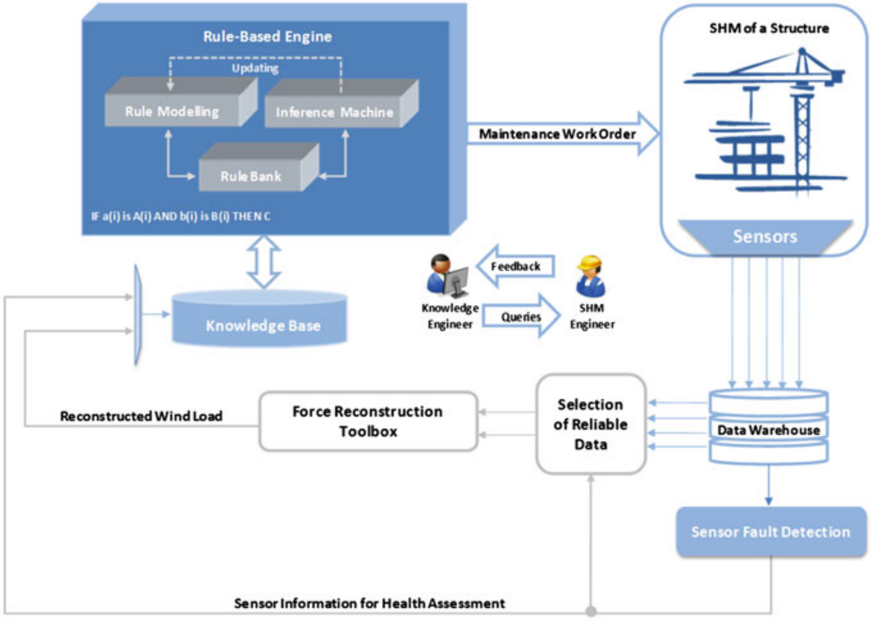


Fig. 6 Conceptual model for integration of knowledge-based approach in SHM

the rules. This should be identified in each case (see the sample rules). The rule modeling enhances the use of hybrid approaches through merging and combining linguistic and numerical variables. Thus the knowledge engineer is capable to use flexibly numerical values in the rule conclusion (*THEN C*) which can be, but not necessarily always, a function of input variables. The SHM engineer (i.e. domain expert) may assist based on his/her monitoring and controlling skills and expertise. Once the system is set up, automatic operation is possible. In addition, a learning mechanism can be used for self-adaptive tuning and manipulating of the rule representation *IF-THEN* (i.e. delete redundant rules, and extract relevant or dominant ones).

The inference machine controls the structure (rule interpreter) and provides a methodology for reasoning. It matches the responses given by the users and fires the rules. Thus, it needs to trace the rule bank to reach a conclusion on the queries of the user. In this context, the main methods are forward chaining and backward chaining [7, 8].

The currently developed toolboxes in SHM are often limited in the use of advanced rule-based engine for the support of the reasoning process. Therefore the proposed approach may lead to an improved use of knowledge-based analysis in health monitoring and assessment. In the context of SHM, the rules are mainly structured in a form of *IF-THEN,* and the conditions and consequences are defined by the domain experts. Using the case study of force reconstruction, sample rules are developed which only consider the damage resulted from loading. A sample rule of a component check is presented in the following:

Sample rule:

```
IF
(LastVisualInspection = Corrosion Medium AND
sensor_earthquake > threshold_eq_risk)
THEN (damage_potential = high)
```

The Rule combines input from the sensing of vibration (earthquake) with the results of visual inspection. In fact, some other mechanical effects (e.g. collision) may also be considered. Moreover, the entire reasoning process for health assessment includes cause-effect analysis, which leads to develop rules. The conditions and consequences can be adapted, based on the major influential factors in the health assessment of each type of structure and environment. The main advantage of such an approach is the adaptivity of modeling and updating the existing rules, based on conditional changes, and generating of new rules, based on new requirements. However, increasing the number of rules may raise a need to optimize the reasoning process and in turn use learning approaches for automatic adjustment and updating of parameters and preferences.

The reasoning is the pre-step for decision-making. It explains the relation of causes and expected effects and explains the consequent behavior. Therefore the end-result of the reasoning process is to support monitoring and maintenance planning of the structures e.g. Canton Tower. Nevertheless, SHM systems require a decision-making component for optimal prioritization and selection of alternatives (outputs of reasoning process) through incorporating preferences especially economic risk factors.

3 Prospective Knowledge-Based Decision Support in SHM-Systems

The concept for integration of knowledge-based approaches in SHM can be extended and further developed in the future. The potentials are employing knowledge-based approaches in decision-making, reinforcing of data management, and using a feedback component. In particular, the knowledge-based decision toolbox can be advanced using major principles of decision support instead of only employing logic rules in the reasoning process. The main principles are identifying decision models, internal and external preferences and disturbances as well as decision and feedback mechanisms. A decision is derived through management of data, including data quality control, analysis of data using SHM toolboxes like load reconstruction, calculation of damage indicators and finally by fulfilling certain predefined conditions defined by the domain expert, which might also be a physically based model comparison.

Decision models provide the structure for problem-solving and decision-making, considering risk, financial, environment, legal factors and customer preferences. In addition, the Graphical User Interface (GUI) is required so that the users, with different access levels, e.g. operator, engineer, manager, can communicate with the system. In this way, the communication of the domain expert with the system is managed and the system developer can improve the algorithms, based on their feedback.

The advanced toolbox should include sub-systems for incorporating decision models and automatic generation of decision alternatives. Hence, identification of dependencies between parameters and prioritizing them is essentially important. The decision models can be developed considering condition and environmental changes as well as risk and economic factors. Exploiting existing knowledge and generating new knowledge from past experiences of SHM engineers may also support decision-makers in continuous improvement of maintenance operations and cost controlling [24].

Using the described components, the system should in turn deploy a Learning Algorithm e.g. Artificial Neural Network (ANN) or Bayesian Network (BN) for automatic learning of the decision instances and related preferences. This is, in fact, strengthened through gathering feedbacks e.g. from sensors, servicemen and clients regarding the consequential or actual condition of the structure after selecting a certain decision alternative. The feedback component plays a major role as a watchdog of the system, especially to evaluate the effectiveness of selected decision alternatives and provide potential recommendations for improving SHM algorithms and preferences of decision-making.

4 Conclusion and Future Work

SHM practitioners deal with considerable amount of data gathered via sensors and inspection. Accumulation of data provides opportunities for elaborating the analyses and outlining a knowledge-base, consisting of certain semantics between entities. The semantics relate the entities in a meaningful form to be (re)-usable in reasoning and decision-making. Promising case studies of a simple 3-dof-model and force identification highlighted the potential for integration of knowledge-based approaches in SHM.

This paper commences the collaborative research for the integration of KBS, and related semantic technologies in SHM. It extends the scope of initial contribution appeared in [25]. The proposed concept foreground potential future research, especially dealing with aggregation or hybrid usage of SHM algorithms for discovering and extracting new knowledge in e.g. damage detection and predictive health assessment. Also the feedback component needs to be studied using knowledge visualization and data mining methods to discover improvement potentials for developing new sensors or materials.

The knowledge-based approach also makes the consideration of additional factors and constraints possible. As an exemplary factor for future research economical aspects should be mentioned. The major non-technological risk is associated with economic factors of SHM e.g. cost, present value, and return on investment. Such factors crucially influence the decision-making and maintenance programs as well as the advancement of SHM systems.

References

1. V. Giurgiutiu, *Structural Health Monitoring with Piezoelectric Wafer Active Sensors* (Elsevier, New York, 2007)
2. D. Balageas, C.P. Fritzen, A. Güemes, *Structural Health Monitoring* (ISTE Ltd., London, 2006)
3. K. Worden, J.M. Dulieu-Barton, An overview of intelligent fault detection in systems and structures. Struct. Health Monit. **3**, 85–98 (2004)
4. C.R. Farrar, N.A.J. Lieven, M.T. Bement, An introduction to damage prognosis, in *Damage prognosis—for aerospace, civil and mechanical systems*, ed. by D.J. Inman, C.R. Farrar, V.L. Junior, V.S. Junior (Wiley, Chichester, 2005), pp. 1–12
5. C.R. Farrar, K. Worden, An introduction to structural health monitoring. Philos. Trans. R. Soc. **356**, 303–315 (2007)
6. C.R. Farrar, N.A.J. Lieven, Damage prognosis: the future of structural health monitoring. Philos. Trans. R. Soc. **365**, 623–632 (2007)
7. S. Russell, P. Norvig, *Artificial Intelligence: A Modern Approach* (Prentice Hall, New Jersey, 2010)
8. C. Beierle, G. Kern-Isberner, *Methoden Wissensbasierter Systeme (In German, Methods of Knowledge Based Systems)*, 4th edn. (Springer, Berlin, 2008)
9. E. Turban, J.E. Aronson, T.-P. Liang, *Decision Support Systems and Intelligent Systems*, 7th edn. (Prentice Hall, New Jersey, 2005)
10. C. Haynes, M.D. Todd, E. Flynn, A. Croxford, Statistically-based damage detection in geometrically complex structures using ultrasonic interrogation. Struct. Control Health Monit. **12**(2), 41–52 (2012)
11. P.M. Frank, Fault diagnosis in dynamic systems using analytical and knowledge-based redundancy—a survey and some new results. Automatica **26**(3), 459–474 (1990)
12. R.J. Patton, P.M. Frank, *Fault diagnosis in dynamic systems, theory and applications* (Prentice Hall, New Jersey, 1989)
13. T.R. Groff, T.P. Jones, *Introduction to Knowledge Management: KM in Business* (Butterworth-Heinemann, Oxford, 2003)
14. F. Wijnhoven, in *Knowledge Management: More than a Buzzword*, Knowledge Integration: The practice of Knowledge Management in small and medium enterprises (Physica Verlag-Springer, Heidelberg, Germany, 2006), pp. 1–16
15. M. Eppler, *Managing Information Quality: Increasing the Value of Information in Knowledge-intensive Products and Processes* (Springer, Berlin, 2006)
16. R. Maier, *Knowledge Management Systems, Information and Communication Technologies für Knowledge Management* (Springer, Berlin, 2007)
17. M. Alavi, D.E. Leidner, Knowledge management and knowledge management systems, conceptual foundations and research issues. MIS Q. **25**, 107–136 (2001)
18. E. Turban, E. McLean, J. Wertherbe, *Information Technology for Management: Transforming Organizations in the Digital Economy* (Wiley, New York, 2004)

19. M.J. Druzdzel, R.R. Flynn, in *Decision Support Systems*, Encyclopedia of Library and Information Science (CRC Press, Boca Raton, FL, 2010)
20. L. /. U. E. I. Los Alamos National Laboratories, SHM tools—a matlab toolbox
21. Y.Q. Ni, Y. Xia, W.Y. Liao, J.M. Ko, Technology innovation in developing the structural health monitoring system for guangzhou new tv tower. Struct. Control Health Monit. **16**, 73–98 (2009)
22. Y. Niu, C.-P. Fritzen, Y.-Q. Ni, in *Online simultaneous reconstruction of wind load and structural responses for high-rise structures*, Proceedings of the 9th International Workshop on Structural Health Monitoring, 2013
23. X. Li, in *Graphical User Interface Development for a Structural Health Monitoring (SHM) Toolbox for the Canton Tower*, Master Thesis, Department of Electrical Engineering and Computer Science, University of Siegen, 2013
24. F. Ansari, P. Uhr, M. Fathi, Textual meta-analysis of maintenance management's knowledge assets. Int. J. Serv. Econ Manage. **6**(1), 14–37 (2014)
25. F. Ansari, Y. Niu, I. Buethe, C.P. Fritzen, M. Fathi, in *Integration of Knowledge-Based Approach in SHM Systems: A Case Study of Force Identification, Lecture Notes in Engineering and Computer Science*. Proceedings of The World Congress on Engineering and Computer Science, WCECS 2013, (San Francisco, USA, 2013), 23–25 Oct 2013, pp. 1185–1189

Chapter 56
Models Supporting Project Subcontractors Selection for Main Types of Contracts

Pawel Blaszczyk and Tomasz Blaszczyk

Abstract Depending on your view of the allocation of risk between the project owner and contractor use are three main types of contracts. In this paper, we analyze the problem optimal selection of subcontractors in the case of the application a fixed price, cost-plus and time and materials contracts. Models described in the article can be found applicable in the relations between the project owner and the contractor and between the contractor and subcontractors. As a methodological basis we use the multi-criterial decision model assigning each task to specific contractors (subcontractors) in the project with the function of distribution of penalties arising from delayed completion and potential benefits in the event of early project termination.

Keywords Contract management · Cost-plus contracts · Fixed price contracts · Project procurement · Subcontractors selection · Time and materials contracts

1 Introduction

The issue of contract management and procurement management is one of the fundamental problems in the practice of project management. Its importance is highlighted in the one of the most common project management methodologies—PMBoK [1] of the Project Management Institute, which gives it one of the nine areas of knowledge. As the process of “Plan contracting” indicates selection of a

P. Blaszczyk
University of Silesia, Katowice, Poland
e-mail: pawel.blaszczyk@us.edu.pl

T. Blaszczyk (✉)
University of Economics in Katowice, 1 Maja 50, 40-287 Katowice, Poland
e-mail: tomasz.blaszczyk@ue.katowice.pl

H. K. Kim et al. (eds.), *Transactions on Engineering Technologies*,
DOI: 10.1007/978-94-017-9115-1_56,

subcontractors cannot be planned without taking into account the relationship between the “project schedule”, “activity cost estimations”, “activity resource requirements”, and “Project Management Plan” with its components focused on “risk register” and “risk-related contractual agreements”. Therefore, we believe that the process of selecting a contractor must include the methodology used to create the schedule in accordance to the policy of distribution of risks between the project owner and the contractor. Classical scheduling methods of Critical Path Method (CPM) and Program Evaluation and Review Technique (PERT) do not explicitly take into account the factors associated with the uncertainty that surrounds the choice of the contractor and his participation in the implementation of the project scope. Therefore, in the paper we use assumptions described by Blaszczyk et al. [2] in terms of the bi-criterial model of project with time and cost buffers designed based on the concept of Critical Chain Project Management (CCPM). Concept of Critical Chain [3] is one of the newest project management methodologies. However, it is not impeccable (compare Herroelen and Leus [4], Rogalska et al. [5] Van de Vonder et al. [6]), is a balanced combination of CPM with the recognition of the overall uncertainty and the impact of the human factor in the planning and implementation of projects. Many later results (e.g. as described in review by Van de Vonder et al. [7, 8) is based on the concepts set forth in the buffers. In some industries (mainly IT) in later years has also found widespread the class of agile [9] methods like XP, Scrum, Lean Software Development. These methods assume the collaboration between client and the contractor including, among others the dynamic adjustment of the schedule and budget for the project to imprecisely specified range, which can also evolve as the work progresses. This approach can be effective, but we must keep in mind that in many business environments, it is not acceptable because of the uncertainty of the final price of the contract for construction of the project. Therefore, we would not consider them later in this paper. Instead, we will focus on the method resulting the concept of CCPM. We will also present considerations for the three types of contract, without indicating which one is more appropriate for a particular project owner. The choice of the type of contract should be determined for a detailed analysis of the type of project, its associated risks and project owners risk management policies. The chain and time buffers quantification methods were the results of successive authors. One of the detailed approaches was formally described by Tukel et al. [10]. The issues of buffering some project characteristics, other than duration, were considered by Leach [11], Gonzalez et al. [12], Blaszczyk and Nowak [13]. The model featured in this research takes into account the use of partition function the bonus fund for early implementation of the project. The problem of cost and time overestimation occurs in the present case twice: between the client and the general contractor of the project (related by the contract with the client) and after that between the subcontractors (performing partial ranges of the project) and a general contractor. Hence, if we assume that overestimations of cost and time act consistently in contracting relationships between client and contractor, and between contractor and sub-contractor, than estimations of cost and/or the schedule given to the project owner will be based on data twice overestimated. On the other hand, there are also lots of cases of under-estimations that are evaluated by the project owner in order to select contractors. The reason for this phenomenon is to seek bidders to obtain the best evaluation in the selection phase at the expense of an increased risk of non-compliance with contractual provisions. In order to increase project owners safety in similar to the consequences of actions that may result in extending the time for implementation the mechanism of contractual penalties applies. Such a mechanism causing that the bi-criterial time-cost trade-off can be represented as a single-criterial problem and can be applied in considered contract types. Any exceeding the agreed deadline for completion of the project or any part of it results in measurable, and the financial consequences set before. When this decision problem is analyzed by the general contractor, it is necessary to take into account the cost of penalties that the general contractor will be forced to pay for the customer, as well as any income from fines paid by its subcontractors. Also the type of the contract has influence how to select the subcontractors. Therefore, it appears appropriate to use for the design of the model of subcontractors selection both the information about contract type and the concept of time and the cost of buffers. In this case such models can be used as a tool to compensate for liabilities and cash flows.

2 Methodology

In this model we propose the extension of the approach described by Blaszczyk and Blaszczyk [14]. We consider a project which consist $x_1, \ldots, x_n$ tasks characterized by cost and time criteria. As the consequence of the contract between the project owner and the contractor there are contracted budget K_{max} and contracted duration T_{max} of the project. Moreover there are defined price I_p, success fee S_p and penalty fee P_p. The success fee and the contract penalty fee can be defined as follow:

$$S_p = r_s \cdot I_p/\text{day} \tag{1}$$

$$P_p = r_p \cdot I_p/\text{day} \tag{2}$$

where r_s and r_p are success rate and penalty rate respectively. We assume that we have q potential subcontractors for n tasks in the project. Let us consider the matrix

$$X = \begin{bmatrix} x_{11} & \ldots & x_{1q} \\ \vdots & \ddots & \vdots \\ x_{n1} & \ldots & x_{nq} \end{bmatrix}. \tag{3}$$

Elements of the matrix X equal 0 or 1. If x_{ij} equals 1 it means that subcontractor j submit a bid for task x_i. In the other case there is no such bid for task x_i. The matrix X we will call the subcontractors matrix.

2.1 The Fixed Price Contract

In our first model we assume that between project owner and the contractor there is contract with the fixed price. Let

$$K = [k_{ij}]_{i=1,\dots,n;\, j=1,\dots,q} \tag{4}$$

to be the matrix of costs of all subcontractors for all tasks and let

$$T = [t_{ij}]_{i=1,\dots,n;\, j=1,\dots,q} \tag{5}$$

be the matrix of amounts of work for subcontractors. Denote by k_i the cost and by t_i duration of the task x_i. Thus the total cost and the total duration of the project equals

$$K_c = \sum_{i=1}^{n} k_i \tag{6}$$

$$T_c = \max_{i=1,\dots,n} (ES_i + t_i) \tag{7}$$

where ES_i is the earliest start of task x_i. Contracts with all subcontractors can also include success fee s_i and penalty fee p_i. Also let

$$A = [a_{ij}]_{i=1,\dots,n;\, j=1,\dots,q} \tag{8}$$

to be the subcontractors assign matrix. Elements of the matrix A equal 0 or 1. If a_{ij} equals 1 it means that subcontractor j will be perform task x_i. Moreover, let

$$M = [m_{ij}]_{i=1,\dots,n;\, j=1,\dots,q} \tag{9}$$

denote the preference matrix. Elements of the matrix M equal 0 or 1. If m_{ij} equals 1 it means that the task x_i should be realized together with task X_j by the same subcontractors. Of course there are ones on the main diagonal. On the other hand let

$$D = [d_{ij}]_{i=1,\dots,n;\, j=1,\dots,q} \tag{10}$$

denote the restriction for tasks in project. Elements of the matrix D equal 0 or 1. If d_{ij} equals 1 it means that the task x_i could not be realized together with task x_j by the same subcontractors. In our case we have the following optimization problem. In order to simplified the calculation let us also introduce the following vectors

$$m = MI \tag{11}$$

$$d = DI \tag{12}$$

where $I = [i_{ij}]_{i,1,\ldots,n;j=1,\ldots,q}$ is an identity matrix. Under this assumptions we maximize the total benefits of the project. We have the following optimization problem

$$\begin{aligned}
&I_p + S_p - P_p - \sum_{i=1}^{n} a_{ij}x_{ij}(k_{ij} + s_{ij} - p_{ij}) \rightarrow \max \\
&a_{ij} \in \{0, 1\} \\
&\forall_{i=1,\ldots,n} \sum_{j=1}^{n} a_{ij} = 1 \\
&\forall_{i=1,\ldots,n} \sum_{j=1}^{n} a_{ij}x_{ij} = 1 \\
&T_c = \max_{i=1,\ldots,n} \{ES_i + t_i\} < T_{max} \\
&K_c = \sum_{i=1}^{n}\sum_{j=1}^{q} a_{ij}k_{ij} < K_{max} \\
&\forall_{j=1,\ldots,q} \sum_{i=1}^{n}\sum_{j=1}^{q} a_{ij}x_{ij}m_{ik} \in \{0, m_i\} \\
&\forall_{j=1,\ldots,q} \sum_{i=1}^{n}\sum_{j=1}^{q} a_{ij}x_{ij}d_{ik} \in \{0, d_i\}
\end{aligned} \tag{13}$$

where ES_i is the earliest start of task x_i, T_c denotes total duration of the project, and K_c denotes the total cost of the project T_{max}, K_{max} denotes maximum duration and cost for the project respectively. The T_{max}, K_{max} are results of the project requirements. It leads to find the optimal work assignments for every factor in each task. From the set of alternate optimal solutions we choose this one, for which the total duration of project is minimal.

2.2 The Cost-Plus Contract

In second model we assume that the project will be settled by the cost-plus formula on the basis of the quantity survey. Let cost and duration matrices be given by formulas (4) and (5) respectively. We treat cost from matrix (4) as the cost of actual implementation of each task for each subcontractors. Let

$$G = [g_{ij}]_{i=1,\ldots,n} \tag{14}$$

be the vector of profit margins for all subcontractors. The values g_i belongs to the interval [0, 1]. To protect against the uncontrolled growth of the cost of the task x_i in such type of contracts the so-called ceiling price is used. So let

$$C = [c_{ij}]_{i=1,\ldots,n} \tag{15}$$

be the vector of ceiling price for each tasks. Like in previous case denote by k_i the cost and by t_i duration of the task x_i. Thus the total cost of the project is given by

$$K_c = \sum_{i=1}^{n} k_i g_i \tag{16}$$

where g_i is the profit margin of subcontractor who will perform the task x_i. The total duration of project is given by (7). Let assign matrix, preference matrix and restriction matrix be given by (8)–(12) respectively. Under this assumptions we maximize the total benefits of the project. In our case we have the following optimization problem

$$\begin{aligned}
& I_p + S_p - P_p - \sum_{i=1}^{n} a_{ij} x_{ij} (k_{ij}(1 + g_i) + s_{ij} - p_{ij}) \to \max \\
& a_{ij} \in \{0, 1\} \\
& \forall_{i=1,\ldots,n} \sum_{j=1}^{n} a_{ij} = 1 \\
& \forall_{i=1,\ldots,n} \sum_{j=1}^{n} a_{ij} x_{ij} = 1 \\
& T_c = \max_{i=1,\ldots,n} \{ES_i + t_i\} < T_{max} \\
& K_c = \sum_{i=1}^{n} \sum_{j=1}^{q} a_{ij} k_{ij} < K_{max} \\
& \forall_{j=1,\ldots,q} \sum_{i=1}^{n} \sum_{j=1}^{q} a_{ij} x_{ij} m_{ik} \in \{0, m_i\} \\
& \forall_{j=1,\ldots,q} \sum_{i=1}^{n} \sum_{j=1}^{q} a_{ij} x_{ij} d_{ik} \in \{0, d_i\} \\
& \forall_{j=1,\ldots,q} \sum_{i=1}^{n} \sum_{j=1}^{q} a_{ij} x_{ij} k_{ij} (1 + g_i) < c_i
\end{aligned} \tag{17}$$

From the set of alternate optimal solutions we choose this one, for which the total duration of project is minimal.

2.3 Time and Materials Contract

In our third model we assume that the project will be settled by the time and used materials. Let cost and duration matrices be given by (4) and (5) respectively. In this model we treat cost from matrix (4) as the unit cost of a working hour. Moreover let

$$C = [c_{ij}]_{i=1,\ldots,n} \tag{18}$$

be the vector of cost of the materials for all subcontractors. Let k_i be the unit cost and t_i duration of the task x_i. Thus the total cost of the project is given by

$$K_c = \sum_{i=1}^{n} k_i t_i + c_i \tag{19}$$

where t_i is the duration and c_i is the cost of the materials for the task x_i. Let

$$E = [e_i]_{i=2,\ldots,n} \tag{20}$$

be the vector of maximal duration for each task. The total duration of project is given by (7). Like in previous model let assign matrix, preference matrix and restriction matrix be given by (8)–(12) respectively. Under this assumptions we maximize the total benefits of the project. In our case we have the following optimization problem

$$\begin{aligned}
& I_p + S_p - P_p - \sum_{i=1}^{n} a_{ij} x_{ij} (k_{ij} t_{ij} + c_{ij} + s_{ij} + b_i - p_{ij}) \to \max \\
& a_{ij} \in \{0, 1\} \\
& \forall_{i=1,\ldots,n} \sum_{j=1}^{n} a_{ij} = 1 \\
& \forall_{i=1,\ldots,n} \sum_{j=1}^{n} a_{ij} x_{ij} = 1 \\
& T_c = \max_{i=1,\ldots,n} \{ES_i + t_i\} < T_{max} \\
& K_c = \sum_{i=1}^{n} \sum_{j=1}^{q} a_{ij} k_{ij} < K_{max} \\
& \forall_{j=1,\ldots,q} \sum_{i=1}^{n} \sum_{j=1}^{q} a_{ij} x_{ij} m_{ik} \in \{0, m_i\} \\
& \forall_{j=1,\ldots,q} \sum_{i=1}^{n} \sum_{j=1}^{q} a_{ij} x_{ij} d_{ik} \in \{0, d_i\}
\end{aligned} \tag{21}$$

where the b_i is the bonus fee for the contractor for early completion of the task x_i. The value of the bonus depends on value of the task and completion time. To motivate the contractor the value of the bonus rate should be grater that the unit cost for the task. One of the possible ways of calculating the bonus is to use the formula

$$b_i = k_i(1 + \alpha) \tag{22}$$

where $\alpha > 0$ is the increase rate. This mechanism could be use on request in order to improve project performance. From the set of alternate optimal solutions we choose this one, for which the total duration of project is minimal.

3 Example

Let us consider the simplified example of typical software development project. In the following project we want to design and implement a software with three functionalities. The whole project was divided into 22 tasks $A_1, \ldots, A_{22}$. At the beginning we should define the problem (task A_1), describe the requirements (task A_2) and action plan (task A_3). After that each functionality should be designed (tasks A_4–A_6). Also some functionality integration should be done. After that each of them should be implemented (tasks A_8–A_{10}) and tested (tasks A_{11}–A_{13}). Also all of them should be tested together (task A_{17}). Next necessary improvements in each functionality should be done (tasks $A_{14} - A_{16}$). After that all of them should be integrated together. After that the whole program should be implemented into our customer environment (task A_{19}). After that some improvements may be necessary (task A_{20}). At the end the customers employees should be learned how to use this program (task A_{21}) and some marketing should be done (task A_{22}). Project Gantt chart network is presented on Fig. 1. These tasks can be performed by subcontractors or by ourselves. In the first step we collect bids from subcontractors and note their estimation of the time and costs required to complete this project. In this way we can construct the matrix of subcontractors X, time and cost matrices T and K respectively. Consider the first of the models. Let us assume that in our case we received bids from three potential subcontractors. Moreover part of the tasks we want to perform ourselves. The values of elements in matrix X are given in Table 1. In our contract fixed price equals $I_p = 250{,}000\$$ and time of duration $T_p = 300$ days. The maximal cost of the project equals $K_{max} = 200{,}000\$$, the maximal time of duration for whole project was fixed at $T_{max} = 270$ days. The times of duration (5) and cost (4) for tasks in project for all subcontractors are given in Tables 2, 3 respectively. In both of these matrices we add our estimations of times and costs for tasks in project. In this case we also have the following preferences. The functionality 1 should be designed (task A_4) and implemented (task A_5) by the same subcontractors. The same should be applied for functionality 2 and 3. Moreover the any two of functionalities should not to be designed or

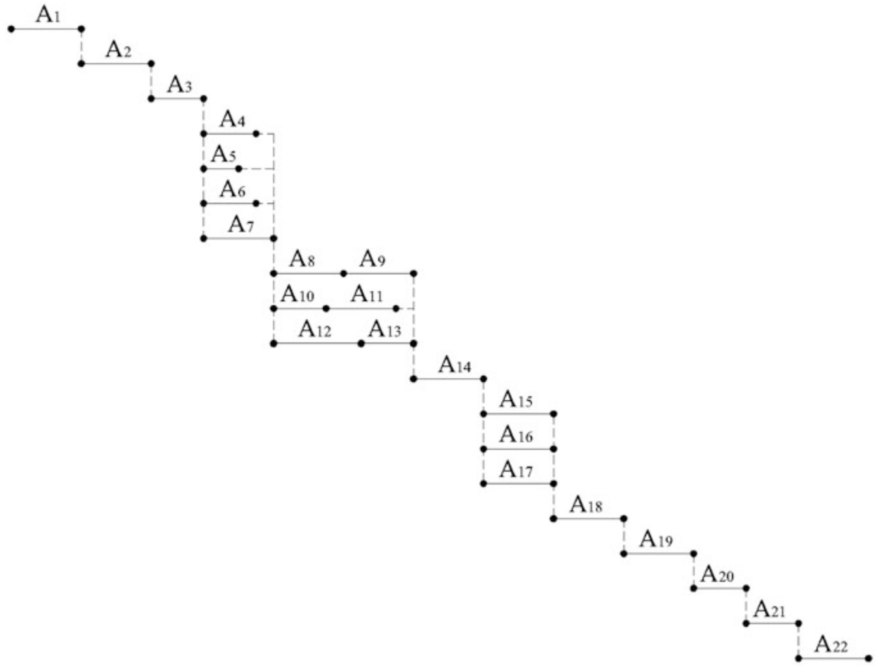


Fig. 1 Gantt chart

Table 1 The subcontractors influence matrix

Task	Subcontractor 1	Subcontractor 2	Subcontractor 3	Self
A_1	1	0	0	0
A_2	1	1	1	0
A_3	1	0	0	1
A_4	1	0	1	0
A_5	1	1	0	0
A_6	1	1	0	0
A_7	1	0	1	1
A_8	1	0	1	0
A_9	1	1	0	0
A_{10}	1	1	0	0
A_{11}	1	0	1	1
A_{12}	1	1	0	1
A_{13}	1	1	0	1
A_{14}	1	0	1	0
A_{15}	1	1	0	0
A_{16}	1	1	0	0
A_{17}	1	1	0	1
A_{18}	1	1	0	0
A_{19}	1	0	0	1
A_{20}	1	0	0	1
A_{21}	1	0	0	1
A_{22}	1	0	0	1

Table 2 Time matrix

Task	Subcontractor 1	Subcontractor 2	Subcontractor 3	Self
A_1	7	0	0	14
A_2	30	28	25	0
A_3	14	0	0	14
A_4	14	0	12	0
A_5	10	5	0	0
A_6	8	5	0	0
A_7	30	0	0	14
A_8	70	0	65	0
A_9	52	40	0	0
A_{10}	34	38	0	0
A_{11}	7	0	10	10
A_{12}	7	10	0	10
A_{13}	7	10	0	10
A_{14}	14	0	7	0
A_{15}	14	14	0	0
A_{16}	14	12	0	0
A_{17}	21	14	0	30
A_{18}	14	21	0	0
A_{19}	5	0	0	14
A_{20}	14	0	0	10
A_{21}	7	0	0	7
A_{22}	14	0	0	28

implemented by the same subcontractors. Also the tests for all functionalities (tasks A_{11}–A_{14}) should be done by another subcontractor. Moreover, the necessary corrections in each function should be performed by subcontractor, which implement that functionality. The values of assignment matrix A are given in Table 4. With such a task distribution we obtain the total cost of the project $K_c = 161{,}300$ and the total duration $T_c = 257$ days. Finally, the total profits of the project equals 88,700$. Now let us consider the second model. In such case the values of elements in X are given in Table 1. In this case the preferences are exactly the same like in previous model. Moreover in this case we defined the vector of the profit margins and vector of ceiling prices for all task in the project. This information are given in Tables 5, 6 respectively. Moreover in this model, we assume that each of the subcontractors reliably estimated the direct costs of the task. The times of duration and costs for tasks in project for all subcontractors are given in Tables 7, 8 respectively. In both of these matrices we add our estimations of times and costs for tasks in project. The values of assignment matrix A are given in Table 9. With such a task distribution we obtain the total cost of the project $K_c = 231{,}050$\$ and the total duration $T_c = 259$ days. Finally, the total profits of the project equals 1,850$. In time and materials contract maximal duration of

Table 3 Cost matrix

Task	Subcontractor 1	Subcontractor 2	Subcontractor 3	Self
A_1	2,000	0	0	1,000
A_2	10,000	8,000	12,500	0
A_3	3,000	0	0	1,500
A_4	10,000	0	8,000	0
A_5	8,000	9,000	0	0
A_6	6,000	4,000	0	0
A_7	1,500	0	0	1,500
A_8	2,5000	0	18,000	0
A_9	1,5000	14,500	0	0
A_{10}	1,2500	12,500	0	0
A_{11}	5,000	0	4,000	2,000
A_{12}	3,000	2,500	0	15,000
A_{13}	2,000	2,000	0	1,000
A_{14}	0	0	0	2,000
A_{15}	0	0	0	2,000
A_{16}	0	0	0	2,000
A_{17}	8,250	5,000	0	4,800
A_{18}	15,000	17,000	0	0
A_{19}	20,000	0	0	25,000
A_{20}	10,000	0	0	12,000
A_{21}	6,000	0	0	5,000
A_{22}	30,000	0	0	25,000

Table 4 Assignment matrix

Task	Subcontractor 1	Subcontractor 2	Subcontractor 3	Self
A_1	0	0	0	1
A_2	0	1	0	0
A_3	0	0	0	1
A_4	0	0	1	0
A_5	1	0	0	0
A_6	0	1	0	0
A_7	0	0	0	1
A_8	0	0	1	0
A_9	1	0	0	0
A_{10}	0	1	0	0
A_{11}	0	0	1	1
A_{12}	0	0	0	1
A_{13}	0	0	0	1
A_{14}	0	0	1	0
A_{15}	1	0	0	0
A_{16}	0	1	0	0
A_{17}	0	0	0	1
A_{18}	1	0	0	0
A_{19}	1	0	0	0
A_{20}	1	0	0	0
A_{21}	0	0	0	1
A_{22}	0	0	0	1

Table 5 Profit margins

Task	Subcontractor 1 (%)	Subcontractor 2 (%)	Subcontractor 3 (%)	Self (%)
A_1	30	35	40	20

Table 6 Celling prices

Task	Price	Task	Price	Task	Price	Task	Price	Task	Price
A_1	3,200	A_6	8,000	A_{11}	6,400	A_{15}	3,200	A_{19}	32,000
A_2	16,000	A_7	2,400	A_{12}	4,000	A_{16}	3,200	A_{20}	16,000
A_3	4,800	A_8	28,800	A_{13}	3,200	A_{17}	8,000	A_{21}	8,000
A_4	12,800	A_9	24,000	A_{14}	3,200	A_{18}	24,000	A_{22}	40,000
A_5	12,800	A_{10}	20,000						

Table 7 Time matrix

Task	Subcontractor 1	Subcontractor 2	Subcontractor 3	Self
A_1	7	0	0	14
A_2	30	28	25	0
A_3	14	0	0	14
A_4	14	0	12	0
A_5	10	5	0	0
A_6	8	5	0	0
A_7	30	0	0	14
A_8	70	0	65	0
A_9	52	40	0	0
A_{10}	34	38	0	0
A_{11}	7	0	10	10
A_{12}	7	10	0	10
A_{13}	7	10	0	10
A_{14}	14	0	7	0
A_{15}	14	14	0	0
A_{16}	14	12	0	0
A_{17}	21	14	0	30
A_{18}	14	21	0	0
A_{19}	5	0	0	14
A_{20}	14	0	0	10
A_{21}	7	0	0	7
A_{22}	14	0	0	28

project and maximal cost are the same like in previous models (Table 10). The times of duration and unit cost for tasks in project for all subcontractors are given in Tables 2, 11 respectively. In both of these matrices we add our estimations of times and costs for tasks in project. Also in this type of contract the preferences are

Table 8 Cost matrix

Task	Subcontractor 1	Subcontractor 2	Subcontractor 3	Self
A_1	2,000	2,000	2,000	2,000
A_2	10,000	10,000	12,500	12,500
A_3	3,000	3,000	3,000	3,000
A_4	8,000	8,000	8,000	8,000
A_5	8,000	8,000	8,000	8,000
A_6	5,000	5,000	5,000	5,000
A_7	1,500	1,500	1,500	1,500
A_8	18,000	18,000	18,000	18,000
A_9	15,000	15,000	15,000	15,000
A_{10}	12,500	12,500	12,500	12,500
A_{11}	4,000	4,000	4,000	4,000
A_{12}	2,500	2,500	2,500	2,500
A_{13}	2,000	2,000	2,000	1,000
A_{14}	2,000	2,000	2,000	2,000
A_{15}	2,000	2,000	2,000	2,000
A_{16}	2,000	2,000	2,000	2,000
A_{17}	5,000	5,000	5,000	5,000
A_{18}	15,000	15,000	15,000	15,000
A_{19}	20,000	20,000	20,000	20,000
A_{20}	10,000	10,000	10,000	10,000
A_{21}	5,000	5,000	5,000	5,000
A_{22}	25,000	25,000	25,000	25,000

Table 9 Assignment matrix

Task	Subcontractor 1	Subcontractor 2	Subcontractor 3	Self
A_1	0	0	0	1
A_2	0	0	1	0
A_3	0	0	0	1
A_4	0	0	1	0
A_5	0	1	0	0
A_6	1	0	0	0
A_7	0	0	0	1
A_8	0	0	1	0
A_9	0	1	0	0
A_{10}	1	0	1	0
A_{11}	0	1	0	1
A_{12}	1	0	0	1
A_{13}	0	0	0	1
A_{14}	0	0	1	0
A_{15}	0	1	0	0
A_{16}	1	0	0	0
A_{17}	0	0	0	1
A_{18}	1	0	0	0
A_{19}	0	0	0	1
A_{20}	0	0	0	1
A_{21}	0	0	0	1
A_{22}	0	0	0	1

Table 10 Maximal duration

Task	Duration	Task	Duration	Task	Duration	Task	Duration	Task	Duration
A_1	112	A_6	65	A_{11}	80	A_{15}	115	A_{19}	112
A_2	240	A_7	240	A_{12}	80	A_{16}	115	A_{20}	112
A_3	115	A_8	560	A_{13}	80	A_{17}	240	A_{21}	60
A_4	112	A_9	420	A_{14}	115	A_{18}	168	A_{22}	225
A_5	80	A_{10}	310						

Table 11 Unit cost matrix

Task	Subcontractor 1	Subcontractor 2	Subcontractor 3	Self
A_1	36	0	0	10
A_2	42	36	63	0
A_3	27	0	0	13
A_4	90	0	85	0
A_5	100	225	0	0
A_6	95	100	0	0
A_7	6	0	0	15
A_8	45	0	35	0
A_9	36	45	0	0
A_{10}	46	40	0	0
A_{11}	90	0	50	25
A_{12}	55	30	0	190
A_{13}	36	25	0	15
A_{14}	0	0	0	0
A_{15}	0	0	0	0
A_{16}	0	0	0	0
A_{17}	50	45	0	20
A_{18}	135	100	0	0
A_{19}	500	0	0	225
A_{20}	90	0	0	150
A_{21}	110	0	0	90
A_{22}	270	0	0	112

exactly the same like in previous model. Moreover in this case introduced the maximal duration of the tasks (in working hours). This information are given in Table 10. The values of assignment matrix A are given in Table 12. With such a task distribution we obtain the cost of the project $K_c = 224{,}985\$$ plus the cost of the materials. The total duration equals $T_c = 1{,}440$ working hours.

Table 12 Assignment matrix

Task	Subcontractor 1	Subcontractor 2	Subcontractor 3	Self
A_1	0	0	0	1
A_2	0	0	1	0
A_3	0	0	0	1
A_4	0	0	1	0
A_5	0	1	0	0
A_6	1	0	0	0
A_7	0	0	0	1
A_8	0	0	1	0
A_9	0	1	0	0
A_{10}	1	0	1	0
A_{11}	0	1	0	1
A_{12}	1	0	0	1
A_{13}	0	0	0	1
A_{14}	0	0	1	0
A_{15}	0	1	0	0
A_{16}	1	0	0	0
A_{17}	0	0	0	1
A_{18}	1	0	0	0
A_{19}	0	0	0	1
A_{20}	0	0	0	1
A_{21}	0	0	0	1
A_{22}	0	0	0	1

4 Conclusions

In this paper we consider three different contract types: the fixed price contract, time and materials contract and the cost-plus contract. For all of this contract types we present a theoretical approaches for selecting subcontractors to develop selected tasks in the project. Even though the fact that each of the models relates to a completely different type of contact they are similar to each other. In the presented models, it is possible that such a division of labor is part of the job was done by the contractor itself and part by the subcontractor. Moreover, the presented models takes into account both preferences and constraints contracting authority in relation to the number and type of tasks that should be or cannot be done by one subcontractor. The usefulness of all of these models has been presented with an embodiment of the software development project. For this simple example we present the principle of each of the models and the differences between them. However, the exploration of the possibility of applying both of this models in real-life conditions requires further studies, both theoretical and practical on the basis of the real-life decision-making problems. The problem of optimal choice of the contract type (the fixed price contract, the cost-plus contract or time and materials contract), according to the project environment and risk transfer policy, will be the subject of the future research.

Acknowledgements The project was funded by the Polish National Science Centre on the basis of the decision number DEC-2011/03/D/HS4/01649

References

1. PMI, A Guide to the Project Management Body of Knowledge (PMBOK Guide), 4th edn (2008)
2. P. Blaszczyk, T. Blaszczyk, M.B. Kania, The bi-criterial approach to project cost and schedule buffers sizing, Lecture Notes in Economics and Mathematical Systems, in *New state of MCDM in the 21st century*. (Springer, Berlin, 2011) pp. 105–114
3. E. Goldratt, *Critical Chain* (North River Press, Great Barrington, 1997)
4. W. Herroelen, R. Leus, On the merits and pitfalls of critical chain scheduling. J. Oper. Manag. **19**, 559–577 (2001)
5. M. Rogalska, W. Boejko, Z. Hejducki, Time/cost optimization using hybrid evolutionary algorithm in construction project scheduling. Autom. Constr. **18**, 24–31 (2008)
6. S. Van de Vonder, E. Demeulemeester, W. Herroelen, R. Leus, The use of buffers in project management: the trade-off between stability and makespan. Int. J. Prod. Econ. **97**, 227–240 (2005)
7. S. Van de Vonder, E. Demeulemeester, W. Herroelen, A classification of predictive-reactive project scheduling procedures. J. Sched. **10**, 195–207 (2007)
8. S. Van de Vonder, E. Demeulemeester, W. Herroelen, Proactive heuristic procedures for robust project scheduling: An experimental analysis. Eur. J. Oper. Res. **189**, 723–733 (2008)
9. Manifesto for Agile Software Development, http://agilemanifesto.org. Accessed 31 May 2013
10. O.I. Tukel, W.O. Rom, S.D. Eksioglu, An investigation of buffer sizing techniques in critical chain scheduling. Eur. J. Oper. Res. **172**, 401–416 (2006)
11. L. Leach, Schedule and cost buffer sizing: how account for the bias between project performance and your model. Proj. Manag. J. **34**, 34–47 (2003)
12. V. Gonzalez, L.F. Alarcon, K. Molenaar, Multiobjective design of work-in-process buffer for scheduling repetitive projects. Autom. Constr. **18**, 95–108 (2009)
13. T. Blaszczyk, B. Nowak, Project costs estimation on the basis of critical chain approach (in Polish), ed. by T. Trzaskalik, Modelowanie Preferencji a Ryzyko 08, Akademia Ekonomiczna w Katowicach (2008)
14. P. Blaszczyk. T. Blaszczyk, Project subcontractors selection in fixed price and cost-plus contracts, Lecture Notes in Engineering and Computer Science, in *Proceedings of The World Congress on Engineering and Computer Science 2013, WCECS 2013*, 23–25 Oct, 2013, San Francisco, USA, pp. 1131–1136