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

Transactions on Engineering Technologies

World Congress on Engineering and Computer Science 2015

Springer

Transactions on Engineering Technologies

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

Transactions on Engineering Technologies

World Congress on Engineering and Computer Science 2015

Springer

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

Mahyar A. Amouzegar
Provost and Senior Vice-President
University of New Orleans
New Orleans, LA
USA

Haeng Kon Kim
Department of Computer and Communication
Engineering College, Catholic University of Daegu
Daegu
South Korea

ISBN 978-981-10-2716-1 ISBN 978-981-10-2717-8 (eBook)
DOI 10.1007/978-981-10-2717-8

Library of Congress Control Number: 2016931995

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #22-06/08 Gateway East, Singapore 189721, Singapore

Preface

A large international conference on Advances in Engineering Technologies and Physical Science was held in San Francisco, California, USA, October 21–23, 2015, under the auspices of the World Congress on Engineering and Computer Science (WCECS 2015). The WCECS 2015 was organized by the International Association of Engineers (IAENG). IAENG, originally founded in 1968, is a non-profit international association for the engineers and the computer scientists. The WCECS Congress serves as an excellent platform for the members of the engineering community to meet and exchange ideas. The congress in its long history has found a right balance between theoretical and application development, which has attracted a diverse group of researchers, leading to its rapid expansion. The conference committees have been formed with over two hundred members including research center heads, deans, department heads/chairs, professors, and research scientists from over 30 countries. The full committee list is available at the congress' web site: www.iaeng.org/WCECS2015/committee.html. WCECS conference is truly an international meeting with a high level of participation from many countries. The response to WCECS 2015 conference call for papers was outstanding, with more than six hundred manuscript submissions. All papers went through a rigorous peer-review process and the overall acceptance rate was 50.89 %.

This volume contains 42 revised and extended research articles, written by prominent researchers, participating in the congress. Topics include engineering mathematics, electrical engineering, communications systems, computer science, chemical engineering, systems engineering, manufacture engineering, and industrial applications. This book offers the state of the art of tremendous advances in engineering technologies and physical science and applications; it also serves as an exceptional source of reference for researchers and graduate students working with/on engineering technologies and physical science and applications.

Hong Kong, Hong Kong — Sio-Iong Ao
Daegu, South Korea — Haeng Kon Kim
New Orleans, LA, USA — Mahyar A. Amouzegar

Contents

Chapter 1
Estimate the Impact of Different Heat Capacity Approximation Methods on the Numerical Results During Computer Simulation of Solidification

Robert Dyja, Elzbieta Gawronska, Andrzej Grosser, Piotr Jeruszka and Norbert Sczygiol

1.1 Introduction

Aluminum alloys are very interesting material widely used in industry. Modeling and computer simulation are one of the most effective methods of studying difficult problems in foundry and metallurgical manufacture. Numerical simulations are used for optimization of casting production. In many cases they are the only possible techniques for carrying out the experiments whose real statement is complicated. Computer modeling allows to define the major factors of a quality estimation of alloy castings. Simulations help to investigate interaction between solidifying casting and changes of its parameters or initial conditions. That process defines the quality of casting, and the problem of adequate modeling of foundry systems. The process mainly depends on the solution of heat equations [1].

Increasing capacity of computer memory makes it possible to consider growing problem sizes. At the same time, increased precision of simulations triggers even greater load. There are several ways to tackle this kind of problems. For instance, one

R. Dyja (✉) · E. Gawronska · A. Grosser · P. Jeruszka · N. Sczygiol
Czestochowa University of Technology, Dabrowskiego 69,
PL42201 Częstochowa, Poland
e-mail: robert.dyja@icis.pcz.pl

E. Gawronska
e-mail: elzbieta.gawronska@icis.pcz.pl

A. Grosser
e-mail: andrzej.grosser@icis.pcz.pl

P. Jeruszka
e-mail: piotr.jeruszka@icis.pcz.pl

N. Sczygiol
e-mail: norbert.sczygiol@icis.pcz.pl

S.-I. Ao et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-981-10-2717-8_1

can use parallel computing [2, 3], someone else may use accelerated architectures such as GPUs [4] or FPGAs [5], while another person can use special organization of computations [6–9].

Solidification may take place at a constant temperature or in the temperature range [10]. If solidification occurs at a constant temperature, it is then referred to as the so-called Stefan problem or the solidification problem with zero temperatures range. Pure metals or alloys of certain specific chemical compositions (e.g. having an eutectic composition) solidify at a constant temperature. However, most of the metal alloys solidify in certain temperature ranges (so-called temperature intervals of solidification). The temperature at which the alloy starts to solidify is called liquidus temperature (T_l), and the temperature at which solidification ends is called solidus temperature (T_s). In the case of alloys with eutectic transformation, in which the solute concentration exceeds its maximum solubility in the solid phase, the temperature of the solidification end is the eutectic temperature. Analytical (rarely) and numerical (commonly) methods are used in the modeling of solidification process. The finite elements method (FEM) is the most commonly used numerical method, but finite difference method (FDM), boundary element method (BEM), the Monte-Carlo and other methods are also used.

The most important heat effect, occurring during solidification, is the emission of (latent) heat of solidification (L). It is also the most difficult phenomenon to numerical modeling. The basic division of numerical methods of solidification modeling process relates to modeling of the latent heat emission. These methods can be divided into front-tracking methods and fixed-grid methods. Fixed-grid methods are also divided into temperature formulations (the latent heat of solidification is considered as the temperature-dependent term of heat source) and enthalpy formulations (the latent heat of solidification is considered as the temperature-dependent term of heat capacity) [11–16]. The enthalpy methods are divided into methods in which the effective heat capacity depends on the temperature and those in which the effective heat capacity depends on the enthalpy. In our article, we have focused on solving the solidification in the temperature range with the finite element method with the use of fixed-grid methods in enthalpy formulation. We have described the comparison of different ways of approximation of heat capacity in apparent heat capacity (AHC) formulation of solidification. This paper is the extention of previous work [17].

During the comparison, we have used Finite Element Method as numerical method of choice. The FEM was used in our own software, that is used to conduct simulations of solidification. This software is implemented in the C++ programming language with the use of essential libraries, i.e. PETSc (for linear algebra) [18] and TalyFEM (for procedures used in FEM) [19].

1.2 Description of the Entalphy Formulation

Solidification is described by a quasi-linear equation of heat conduction, considering a term of heat source $\dot{q}$ as a latent heat of solidification:

$$\nabla \cdot (\lambda \nabla T) + \dot{q} = c\rho \frac{\partial T}{\partial t} \tag{1.1}$$

By entering the following designation:

$$\dot{s} = \dot{q} - c\rho \frac{\partial T}{\partial t} \tag{1.2}$$

equation (1.1) can be written as

$$\nabla \cdot (\lambda \nabla T) + \dot{s} = 0 \tag{1.3}$$

where $\dot{s}$ denotes generalized heat source. By introducing enthalpy, defined as:

$$h = \int_{T_{ref}}^{T} c\rho(T)\, dt \tag{1.4}$$

where T_{ref} is the reference temperature, and calculating the derivative with respect to the temperature:

$$\frac{dH}{dT} = c\rho(T) = c^*(T) \tag{1.5}$$

where c^* is the effective heat capacity. Assuming the heat source is equal to zero, the Eq. (1.3) can be converted to the form:

$$\nabla \cdot (\lambda \nabla T) = c^* \frac{\partial T}{\partial t} \tag{1.6}$$

All above equations form the basis of the thermal description of solidification.

1.2.1 The Enthalpy and the Effective Heat Capacity

The enthalpy is the sum of explicit and latent heat [20]. For the metal solidifying in the temperature range (T_s–T_l) amounts to:

$$H = \int_{T_{ref}}^{T} c\rho_s(T)\, dT, \quad \text{for} \quad T < T_s,$$

$$
\begin{aligned}
H &= \int_{T_{ref}}^{T_s} c\rho_s(T)\, dT + \int_{T_s}^{T} (\rho_s(T)\frac{dL}{dT} + \\
&\quad + c\rho_f(T)\big)\, dT, \quad \text{for} \quad T_s \le T \le T_l, \\
H &= \int_{T_{ref}}^{T_s} c\rho_s(T)\, dT + \\
&\quad + \rho_s(T)L + \int_{T_s}^{T_l} c\rho_f(T)\, dT + \\
&\quad + \int_{T_l}^{T} c\rho_l(T)\, dT, \quad \text{for} \quad T > T_l
\end{aligned}
\tag{1.7}
$$

The integration of the expressions in Eq. 1.7 gives

$$
\begin{aligned}
c^* &= c\rho_s, \quad \text{for} \quad T < T_s, \\
c^* &= c\rho_f + \rho_s \frac{dL}{dT}, \quad \text{for} \quad T_s \le T \le T_l, \\
c^* &= c\rho_l, \quad \text{for} \quad T > T_l.
\end{aligned}
\tag{1.8}
$$

Assuming that the heat of solidification is exuded and spread evenly throughout the temperature range of solidification, the following can be written:

$$
\begin{aligned}
c^* &= c\rho_s, \quad \text{for} \quad T < T_s, \\
c^* &= c\rho_f + \rho_s \frac{L}{T_l - T_s}, \quad \text{for} \quad T_s \le T \le T_l, \\
c^* &= c\rho_l, \quad \text{for} \quad T > T_l.
\end{aligned}
\tag{1.9}
$$

On the basis of the Eq. 1.7 and the Eq. 1.9 one can make the following graphical comparison of the enthalpy and the effective thermal capacity distributions for alloy solidifying in the temperature range (see Fig. 1.1).

1.2.2 The Types of the Entalphy Formulations

There are three types of enthalpy formulations of solidification:

- basic enthalpy formulation (BEF)

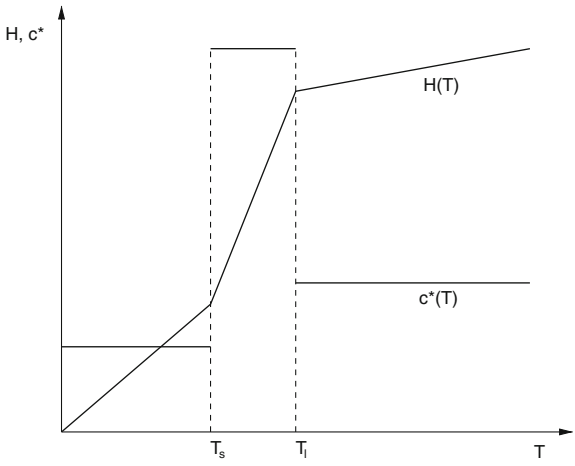


Fig. 1.1 Distribution of enthalpy and effective heat capacity depending on temperature

$$\nabla \cdot (\lambda \nabla T) = \frac{\partial H}{\partial t} \tag{1.10}$$

where

$$H(T) = \int_{T_{ref}}^{T} c\rho \, dT + (1 - f_s(T))\rho_s L \tag{1.11}$$

- apparent (or modified) heat capacity (AHC) formulation
  differentiate Eq. 1.11 with respect to temperature is obtained

$$\frac{dH}{dT} = c\rho - \rho_s L \frac{df_s}{dT} = c^*(T) \tag{1.12}$$

Since $H = H(T(x, t))$ then

$$\frac{\partial H}{\partial t} = \frac{dH}{dT}\frac{\partial T}{\partial t} = c^*(T)\frac{\partial T}{\partial t} \tag{1.13}$$

Substituting Eq. 1.13 to Eq. 1.10 is obtained

$$\nabla \cdot (\lambda \nabla T) = c^*(T)\frac{\partial T}{\partial t} \tag{1.14}$$

- source term formulation (STF)
  The total enthalpy is divided into two parts in accordance with:

$$H(T) = h(T) + (1 - f_s)\rho_s L \tag{1.15}$$

where

$$h(T) = \int_{T_{ref}}^{T} c\rho \, dT \tag{1.16}$$

Derivative Eq. 1.15 with respect to time is

$$\frac{\partial H}{\partial t} = \frac{\partial h}{\partial t} - \rho_s L \frac{\partial f_s}{\partial t} \tag{1.17}$$

Substituting Eq. 1.17 to Eq. 1.10 is obtained

$$\nabla \cdot (\lambda \nabla T) + \rho_s L \frac{\partial f_s}{\partial t} = \frac{\partial h}{\partial t} \tag{1.18}$$

1.3 Approximation of the Effective Heat Capacity

The effective heat capacity can be also calculated directly from the Eq. 1.5, but in this paper, we have presented the results of solidification simulations using the various methods of effective heat capacity approximation.

1. Morgan method—derivative of enthalpy is replaced by a backward differential quotient

$$c^* = \frac{H^n - H^{n-1}}{T^n - T^{n-1}} \tag{1.19}$$

   where $n-1$ and n are the time levels. In some cases, however, this substitution may lead to oscillations in the solution, especially near the boundaries of the temperature range of solidification.
2. Del Giudice method—in order to remove oscillations one should take into account the directional cosines of temperature gradient

$$c^* = \frac{\dfrac{\partial H}{\partial n}}{\dfrac{\partial T}{\partial n}} = \frac{\dfrac{\partial H}{\partial x_i}\alpha_{ni}}{\dfrac{\partial T}{\partial n}}$$

   where

$$\alpha_{ni} = \frac{\dfrac{\partial T}{\partial x_i}}{\dfrac{\partial T}{\partial n}}$$

and

$$\frac{\partial T}{\partial n} = \left(\frac{\partial T}{\partial n} \cdot \frac{\partial T}{\partial n}\right)^{\frac{1}{2}}$$

Hence

$$c^* = \frac{\frac{\partial H}{\partial x}\frac{\partial T}{\partial x} + \frac{\partial H}{\partial y}\frac{\partial T}{\partial y} + \frac{\partial H}{\partial z}\frac{\partial T}{\partial z}}{(\frac{\partial T}{\partial x})^2 + (\frac{\partial T}{\partial y})^2 + (\frac{\partial T}{\partial z})^2} = \frac{H_{,i}T_{,i}}{T_{,j}T_{,j}} \tag{1.20}$$

3. Lemmon method—the temperature gradient is normal to solidification surface

$$c^* = \sqrt{\frac{\left(\frac{\partial H}{\partial x}\right)^2 + \left(\frac{\partial H}{\partial y}\right)^2 + \left(\frac{\partial H}{\partial z}\right)^2}{\left(\frac{\partial T}{\partial x}\right)^2 + \left(\frac{\partial T}{\partial y}\right)^2 + \left(\frac{\partial T}{\partial z}\right)^2}} = \left(\frac{H_{,i}H_{,i}}{T_{,j}T_{,j}}\right)^{\frac{1}{2}} \tag{1.21}$$

4. Comini method—the apparent heat capacity is approximated by the expression

$$c^* = \frac{1}{n}\left(\frac{\frac{\partial H}{\partial x}}{\frac{\partial T}{\partial x}} + \frac{\frac{\partial H}{\partial y}}{\frac{\partial T}{\partial y}} + \frac{\frac{\partial H}{\partial z}}{\frac{\partial T}{\partial z}}\right) = \frac{1}{n}\frac{H_{,i}}{T_{,i}} \tag{1.22}$$

where n is the number of dimensions.

1.4 Numerical Model of Solidification

Solving the partial differential equations can pass from spatial discretization through time discretization to approximate solution. First, we use the finite element method.

The finite element method facilitates the modeling of many complex problems. Its wide application for founding comes from the fact that it permits an easy adaptation of many existing solutions and techniques of solidification modeling.

Computer calculations need to use discrete models, which means problems must be formulated by introducing time-space mesh. These methods convert given physical equations into matrix equations (algebraic equations). This system of algebraic equations usually contains many thousands of unknowns, that is why the efficiency of a method applied to solve them is crucial.

The discretization of the governing equation in space with the finite element method leads to the ordinary differential equation with time derivative, given as:

$$\mathbf{M}(T)\dot{\mathrm{T}} + \mathbf{K}(T)\mathrm{T} = \mathbf{b}(T) \tag{1.23}$$

where M is the capacity matrix, K is the conductivity matrix, T is the temperature vector and b is the right-hand side vector, whose values are calculated on the boundary conditions basis. The global form of these matrices is obtained by summing the coefficients for all the finite elements. The matrix components are defined for a single finite element as follows:

$$\mathbf{M} = \sum_e \int_{\Omega} c^* \mathbf{N}^T \mathbf{N} \, \mathrm{d}\Omega, \tag{1.24}$$

$$\mathbf{K} = \sum_e \int_{\Omega} \lambda \nabla^T \mathbf{N} \cdot \nabla \mathbf{N} \, \mathrm{d}\Omega, \tag{1.25}$$

$$\mathbf{b} = \sum_e \int_{\Gamma} \mathbf{N}_{\Gamma}^T \mathbf{q}^T \mathbf{n} \, \mathrm{d}\Gamma \,, \tag{1.26}$$

where $\mathbf{N}$ is a shape vector in the area Ω, $\mathbf{N}_{\Gamma}$ is a shape vector on the boundary Γ, $\mathbf{n}$ is an ordinary vector towards the boundary Γ, and $\mathbf{q}$ is a vector of nodal fluxes.

Next, we have applied the time discretization. To do this we used a modified Euler Backward time integration scheme (which belongs to a family of the one-step Θ integration schemes [21]):

$$(\mathrm{M}^n + \Delta t \mathrm{K}^n)\, \mathrm{T}^{n+1} = \mathrm{M}^n \mathrm{T}^n + \Delta t \mathrm{b}^{n+1}, \tag{1.27}$$

where superscript n refers to the number of time step.

1.5 Used Software of Engineering Simulation

Growth of computing power allows engineers to design and run engineering simulations on PC. Researchers can use typical engineering software (some kind of CAD etc.), but some of physical phenomena may not be implemented in such software. Authors decided to write their own solidification computing module, because it made computing each method of heat capacity approximation possible. The module was written in C++ programming language. We used the C++ for the module should be fast, scalable and compatible with chosen numerical utilities.

Both GMSH mesh generator and extended (to include our module) TalyFEM library have been used in numerical experiment [19]. GMSH is a widespread tool which allows finite element mesh of problem geometry (declared or created with graphical interface by user) to be generated. Furthermore, the boundary conditions (surfaces and/or volumes) can be declared with the pre-processor using the graphical user interface.

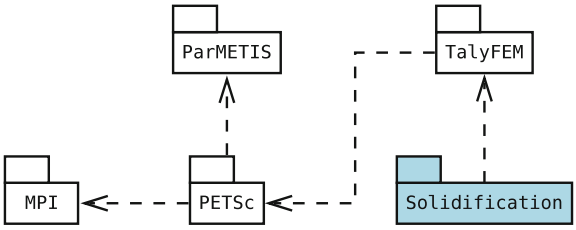


Fig. 1.2 Package diagram in TalyFEM project

TalyFEM is a scalable and extensible framework which uses FEM method to simulate some of physical phenomenon. TalyFEM contains many of PETSc data stuctures, including vectors, matrices or solvers [18]. PETSc library is easy to learn which allows programmers to use mentioned structures in building scalable scientific application. Scalable means using parallel techniques (provided by MPI) but a programmer does not have to write MPI communication routine—all communication about vectors and solvers is written in PETSc and TalyFEM [22]. ParMETIS (for nodes splitting) was also used in our GMSH loading file module (Fig. 1.2). TalyFEM also divides whole problems into sub-meshes (called domains)—each sub-problem can be solved on a separate process (processor or computer node in cluster; Fig. 1.3).

However, the framework does not allow to load meshes generated with GMSH pre-processor. We wrote a GMSH loader module using MPI to communicate the data between processes. Two problems occur:

- loading material properties (which nodes belongs to a cast or an alloy);
- loading neighbourhood nodes (on contact boundary condition) if both nodes are on separate processes;

which have been solved during implementation.

We also implemented solidification module with TalyFEM library. The library requires numerical model of problem solver to be implemented. A programmer does not have to write matrices of the assembly code or communication routines—they are provided with the library. Nevertheless, we considered parameters of boundary conditions (especially contact boundary condition), so we modified standard modules of filling matrices while we were creating the system of linear equations. Framework also solves the created equations (by mentioned communication routines) and writes them into output files (in TecPlot format).

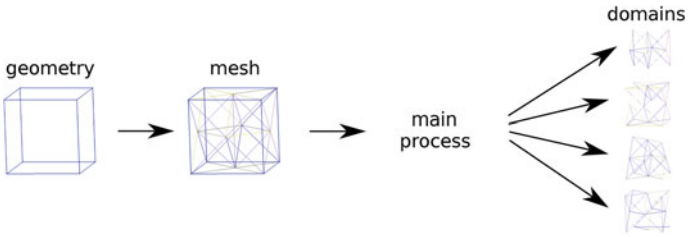


Fig. 1.3 Splitting whole problem to domains in TalyFEM GMSH loader module

1.6 Results of the Numerical Experiment

In the paper we considered a casting solidifying in a metal mold. The finite element mesh comprising 32 814 nodes and 158 417 elements was applied to the area of the casting and mold, as shown in Fig. 1.4. We introduced two boundary conditions: Newton and continuity condition, for which the environment temperature 400 K, the heat transfer coefficient with the environment 10 W/m^2K^{-1} and the heat transfer coefficient through the separation layer 1000 W/m^2K^{-1} are defined. The initial temperature of casting was 960 K, the initial temperature of mold was 600 K, the size of the time step was 0.05 s.

Material properties of the alloy (of which the casting is made of) and the mold are given in Tables 1.1 and 1.2, respectively.

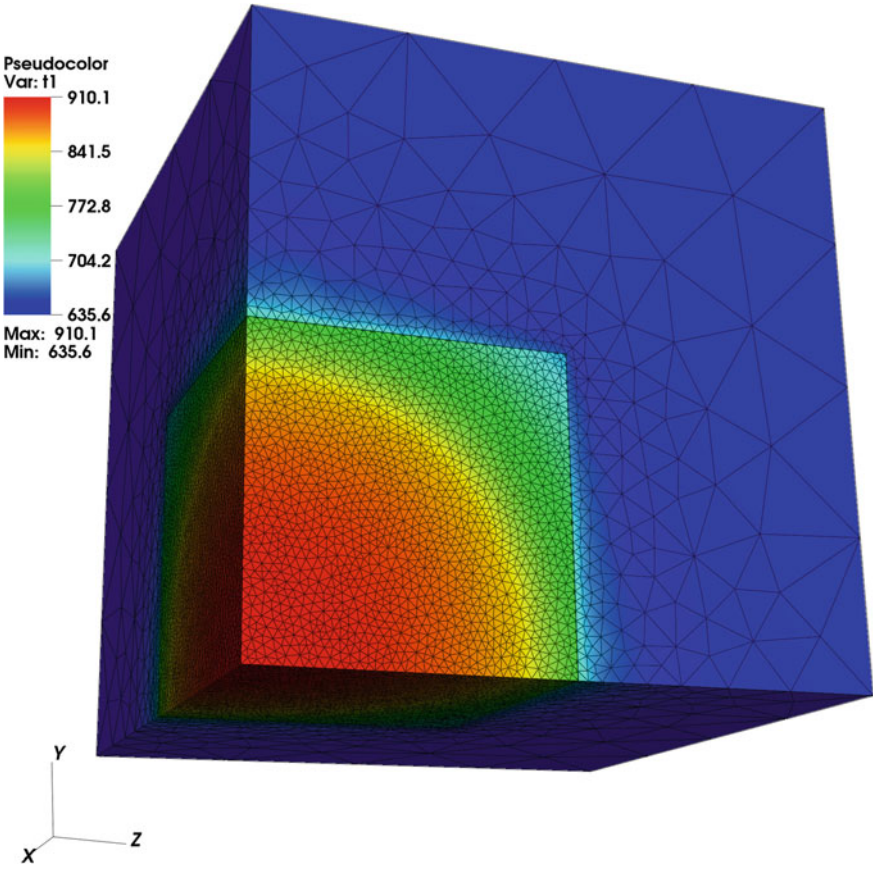


Fig. 1.4 Temperature field after 25 s of cooling. The finite element mesh is also visible

Table 1.1 Material properties of the Al-2 %Cu alloy (subscript *s* means solid phase and *l*—liquid phase)

Quantity name	Unit	Value
Density ρ_s	$\frac{\text{kg}}{\text{m}^3}$	2824
Density ρ_l	$\frac{\text{kg}}{\text{m}^3}$	2498
Specific heat c_s	$\frac{\text{J}}{\text{kgK}^{-1}}$	1077
Specific heat c_l	$\frac{\text{J}}{\text{kgK}^{-1}}$	1275
Thermal conductivity λ_s	$\frac{\text{W}}{\text{mK}^{-1}}$	262
Thermal conductivity λ_l	$\frac{\text{W}}{\text{mK}^{-1}}$	104
Solidus temperature T_s	K	853
Liquidus temperature T_l	K	926
Melt temperature of pure component T_M	K	933
Eutectic temperature T_E	K	821
Heat of solidification L	$\frac{\text{J}}{\text{kg}}$	390 000
Partition coefficient of solute k	–	0.125

Table 1.2 Material properties of the mold

Quantity name	Unit	Value
Density	$\frac{\text{kg}}{\text{m}^3}$	7500
The specific heat	$\frac{\text{J}}{\text{kgK}^{-1}}$	620
Thermal conductivity coefficient	$\frac{\text{W}}{\text{mK}^{-1}}$	40

Figure 1.4 shows the distribution of temperatures in the casting after 25 s for the Morgan heat capacity approximation.

The graphs in Figs. 1.5 and 1.7 show the lack of differences in the obtained results. We can see that cooling curves and solid fractions of all methods overlap. It is caused by the fact that although different heat capacity approximations use different formulas, the resulting approximations are very close in values of effective heat capacity, as can be seen in Fig. 1.6.

However, there is a visible difference between the heat capacity approximation formulas in calculation times. The Fig. 1.8 shows the difference in assembly time for different methods. It is easy to notice that the Morgan method requires the least time, while the other formulas are close together in time requirement.

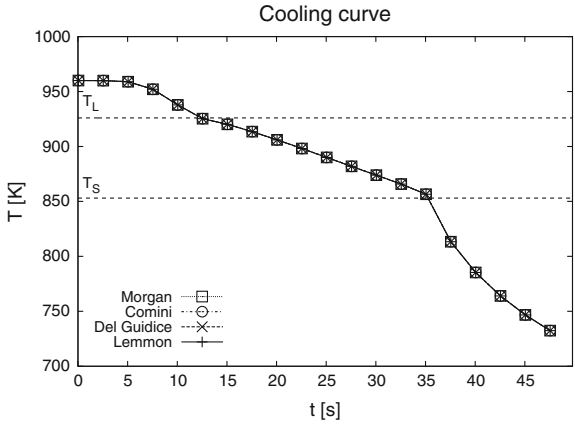


Fig. 1.5 A cooling curve of a point located in origin of coordinate system

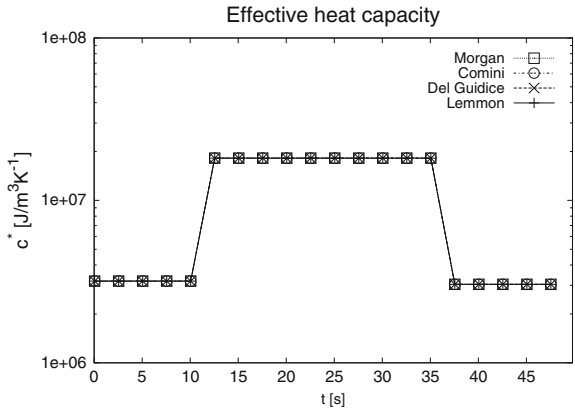


Fig. 1.6 The change of heat capacity approximate during solidification process

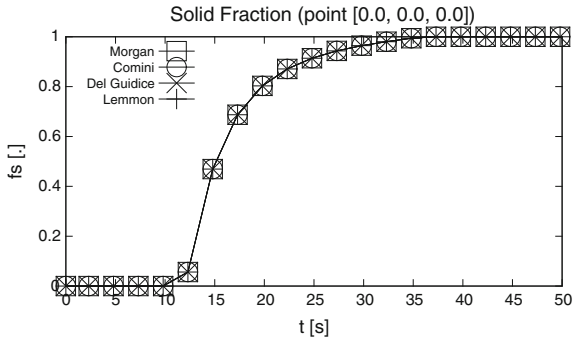


Fig. 1.7 Curve of the solid phase fraction in point (0, 0, 0)

The results from the Fig. 1.8 were obtained for 750 time-steps and mesh from Fig. 1.1. On the 750th time-step (after 37.5 s) the minimum solid fraction was still 0.95 (see Fig. 1.7). This ensures that during the whole calculation time in at least some fraction of finite elements the heat capacity approximation formulas were used.

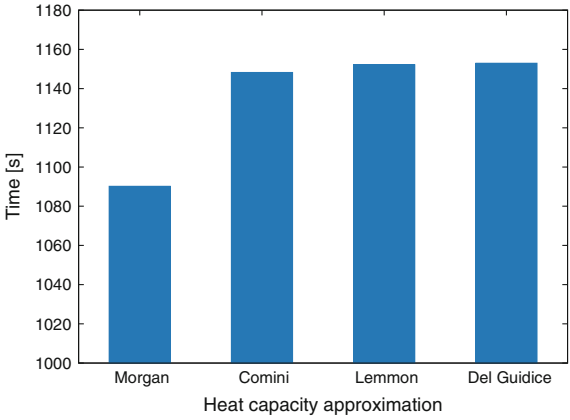


Fig. 1.8 The difference between the heat capacity approximation formulas in time needed for the main matrix assembly

1.7 Conclusions

By analyzing the numerical results obtained from calculations carried out with the help of our own computer program using the finite element method and the apparent heat capacity method we can draw the following remarks and conclusions:

1. The capacity formulation gives an equation very similar to the equation of heat conduction; heat of solidification is hidden in the effective thermal capacity.
2. The use of any of the heat capacity approximation methods does not affect the obtained result, if the solution is stable.
3. When using the Morgan method of heat capacity approximation, one should be careful not to apply too small time step, because then the obtained results might be incorrect.
4. Heat capacity approximation formulas other than Morgan are susceptible to give wrong results if temperature values in nodes of one finite element differ by very small values. The Comini method is especially sensitive to this.
5. Morgan method requires the least time for calculations.

References

1. Stefanescu DM (2002) Science and engineering of casting solidification. Kluwer Academic, New York
2. Wyrzykowski R, Szustak L, Rojek K (2014) Parallelization of 2d mpdata eulag algorithm on hybrid architectures with gpu accelerators. Parallel Comput V40(8):425–447
3. Kim JW, Sandberg RD (2012) Efficient parallel computing with a compact finite difference scheme. Comput Fluids 58:70–87
4. Michalski G, Sczygiol N (2014) Using CUDA architecture for the computer simulation of the casting solidification process. In: Proceedings of the international multiconference of engineers and computer scientists. Lecture notes in engineering and computer science. Hong Kong, 03/2014, pp 933–937

5. Yang N, Li DW, Zhang J, Xi YG (2012) Model predictive controller design and implementation on FPGA with application to motor servo system. Control Eng Pract V20(11):1229–1235
6. Gawronska E, Sczygiol N (2015) Numerically stable computer simulation of solidification: association between eigenvalues of amplification matrix and size of time step. In: Transactions on engineering technologies. Springer, Netherlands, pp 17–30
7. Gawronska E, Sczygiol N (2014) Relationship between eugenvalues and size of time step in computer simulation of thermomechanics phenomena. In: Proceedings of the international multiconference of engineers and computer scientists. Lecture notes in engineering and computer science. Hong Kong, 03/2014, pp. 881–885
8. Ghoneim A, Ojo OA (2011) Numerical modeling and simulation of a diffusion-controlled liquid-solid phase change in polycrystalline solids. Comput Mater Sci V50(3):1102–1113
9. Gawronska E, Sczygiol N (2010) Application of mixed time partitioning methods to raise the efficiency of solidification modeling. In: 12th international symposium on symbolic and numeric algorithms for scientific computing (SYNASC 2010), pp. 99–103
10. Stefanescu F, Neagu G, Mihai A, Stan I, Nicoara M, Raduta A, Opris C (2012) Controlled temperature distribution and heat transfer process in the unidirectional solidification of aluminium alloys. Adv Mater Struct IV 188:314–317
11. Date AW (1994) A novel enthalpy formulation for multidimensional solidification and melting of a pure substance. Sadhana-Acad Proc Eng Sci 19:833–850
12. Duan Q, Tan FL, Leong KC (2002) A numerical study of solidification of n-hexadecane based on the enthalpy formulation. J Mater Process Technol V120(1/3):249–258
13. Ganguly S, Chakraborty S (2006) A generalized formulation of latent heat functions in enthalpy—based mathematical models for multicomponent alloy solidification systems. Metall Mater Trans B-Process Metall Mater Process Sci V37(1):143–145
14. Prabha AJ, Raju S, Jeyaganesh B, Rai AK et al (2011) Thermodynamics of $\alpha'' \rightarrow \alpha$ phase transformation and heat capacity measurements in Ti-15at%Nb alloy. Phys B V406(22):4200–4209
15. Famouri M, Jannatabadi M, Ardakani HTF (2013) Simultaneous estimations of temperature-dependent thermal conductivity and heat capacity using a time efficient novel strategy based on megann. Appl Soft Comput V13(1):201–210
16. Dyja R, Sczygiol N, Domanski Z, Ao S, Chan A, Katagiri H, Xu L (2014) The effect of cavity formation on the casting heat dissipation rate. In: IAENG transactions on engineering sciences, pp 341–347
17. Dyja R, Gawronska E, Grosser A, Jeruszka P, Sczygiol N (2015) Comparison of different heat capacity approximation in solidification modeling. In: Proceedings of the world congress on engineering and computer science 2015. Lecture notes in engineering and computer science. San Francisco, USA, pp 875-879. Accessed 21–23 Oct 2015
18. Balay S et al (2014) PETSc Users Manual, Argonne National Laboratory
19. Kodali HK, Ganapathysubramanian B (2012) A computational framework to investigate charge transport in heterogeneous organic photovoltaic devices. Comput Methods Appl Mech Eng N247:113–129
20. Sczygiol N, Szwarc G (2001) Application of enthalpy formulation for numerical simulation of castings solidification. Comput Assist Mech Eng Sci N8:99–120
21. Wood LW (1990) Practical time-stepping schemes. Clarendon Press, Oxford
22. MPI: A Message Passing Interface. http://www.mpi-forum.org, mpi30-report.pdf, 15.03.2016 r

Chapter 2
Analysis of Systemic Risk: A Dynamic Vine Copula-Based ARMA-EGARCH Model

Kuan-Heng Chen and Khaldoun Khashanah

2.1 Introduction

The definition of systemic risk from the report to G20 Finance Ministers and Governors agreed upon among the International Monetary Fund (IMF), Bank for International Settlements (BIS), and Financial Stability Board (FSB) [1] is that "(i) caused by an impairment of all or parts of the financial system and (ii) has the potential to have serious negative consequences for the real economy". Furthermore, "G-20 members consider an institution, market or instrument as systemic if its failure or malfunction causes widespread distress, either as a direct impact or as a trigger for broader contagion." A common factor from the various definitions of systemic risk is that a trigger event causes a chain of adverse economic consequences, referred to as a "domino effect". Given the definition of systemic risk quoted above, measuring systemic risk is done by estimating the probability of failure of an institution that is the cause of a distress for the financial system. Therefore, we only consider the Value-at-Risk (VaR), the potential loss in value of an asset or portfolio for a given time period and probability, as the risk measurement. In addition, the VaR ratio of a sector to the system (S&P 500 Index), which interprets that the sector risk provides to the entire system, is present.

Girardi and Ergün [2] modified the CoVaR methodology that is proposed by Adrian and Brunnermeier [3] which used the dynamic conditional correlation GARCH, while Hakwa et al. [4] modified the methodology based on copula

K.-H. Chen (✉) · K. Khashanah
Department of Financial Engineering, Stevens Institute of Technology,
Hoboken, NJ 07030, USA
e-mail: kchen3@stevens.edu

K. Khashanah
e-mail: kkhashan@stevens.edu

S.-I. Ao et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-981-10-2717-8_2

modeling. We present dynamic vine Copula-based ARMA-EGARCH (1, 1) VaR measure into a high dimensional analysis in systemic risk.

Sklar [5] introduced the copula to describe the dependence structure between variables. Patton [6] defined the conditional version of Sklar's theorem, which extends the copula applications to the time series analysis. Otani and Imai [7] presented a basket credit default swaps (CDSs) pricing model with nested Archimedean copulas. However, multivariate Archimedean copulas are limited in that there are only one or two parameters to capture the dependence structure. Joe [8] introduced a construction of multivariate distribution based on pair-copula construction (PCC), while Aas et al. were the first to recognize that the pair-copula construction (PCC) principle can be used with arbitrary pair-copulas, referred to as the graphical structure of R-vines [9]. Furthermore, Dissmann et al. [10] developed an automated algorithm of jointly searching for an appropriate R-vines tree structures, the pair-copula families and their parameters. Accordingly, a high dimensional joint distribution can be decomposed to bivariate and conditional bivariate copulas arranged together according to the graphical structure of a regular vine. Besides, Rockinger and Jondeau [11] was the first to introduce the copula-based GARCH modeling. Afterwards, Lee and Long [12] concluded that copula-based GARCH models outperform the dynamic conditional correlation model, the varying correlation model and the BEKK model. In addition, Fang et al. [13] investigated that using Akaike Information Criterion (AIC) as a tool for choosing copula from a couple of candidates is more efficient and accurate than the multiplier goodness-of-fit test method.

During 2008, the subprime mortgage crisis was a systemic collapse triggered by the financial industry. The purpose of this paper is to present an application of the estimation of systemic risk in terms of the VaR/ES ratio by using the dynamic vine copula-based ARMA-EGARCH (1, 1) model. To compute systemic risk for our system, we use S&P 500 sector indices and S&P 500 index to be our components and system, respectively. Since the parameters change over time, we calibrate the parameters every ten steps to capture a change of the structure. This scalable prototype of US financial system with limited dimensionality can be easily tailored to any underlying sector, country or financial market.

This paper has four sections. The first section briefly introduces existing research regarding systemic risk. The second section describes the definition of the VaR/ES ratio, and outlines the methodology of vine Copula-based EGARCH (1, 1) modeling. The third section describes the data and explains the results of VaR/ES ratio. The fourth section concludes our findings.

2.2 Methodology

2.2.1 Risk Methodology

The definition of Value-at-Risk (VaR) is that the maximum loss at most is (1-α) probability given by a period [14]. People usually determine α as 95 %, 99 %, or 99.9 % to be their confidence level. In this study, we use the Copula-based ARMA-EGARCH (1, 1) methodology to obtain the dynamic VaR from each sector. We denote $VaR^{i\to j}_{t,1-\alpha}$ ratio, the sector i's risk contribution to the system j (S&P 500 index) at the confidence level α, by

$$VaR^{i\to j}_{t,1-\alpha}\,Ratio = \frac{VaR^{i}_{t,1-\alpha}}{VaR^{j}_{t,1-\alpha}}$$

The higher $VaR^{i\to j}_{t,1-\alpha}$ ratio indicates that the sector is the risk provider to the system. In addition, the methodology can be easily extended from the VaR ratio to the expected shortfall (ES) ratio.

2.2.2 Univariate ARMA-EGARCH Model

Engle is the first researcher to introduce the ARCH model, which deals with volatility clustering, usually referred to as conditional heteroskedasticity. Bollerslev [15] extended the ARCH model to the generalized ARCH (GARCH) model. Chen and Khashanah [16] implemented ARMA (p, q)-GARCH (1, 1) with the Student's t distributed innovations for the marginal to account for the time-varying volatility, whereas the Student's t distributed innovations cannot explain the skewness. In addition, to overcome the leverage effect in financial time series, we use the exponential GARCH (EGARCH) model in handling asymmetric effects between positive and negative asset returns proposed by Nelson [17]. According to the augmented Dickey–Fuller (ADF) test, all the series are stationary. Therefore, ARMA (p, q)-EGARCH (1, 1) with the skewed Student's t distributed innovation can then be written as [18]

$$\begin{aligned} r_t &= \mu_t + \sum_{i=1}^{p} \vartheta_i r_{t-i} + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} + \varepsilon_t, \\ \varepsilon_t &= \sigma_t z_t, \\ \ln(\sigma_t^2) &= \kappa_t + \alpha_t z_{t-1} + \xi_t(|z_{t-1}| - \mathrm{E}\{|z_{t-1}|\}) + \beta_t \ln(\sigma_{t-1}^2) \end{aligned}$$

where r_t is the log return, μ_t is the drift term, ε_t is the error term, ξ_t capture the size effect, and the standardized innovation term z_t is the skewed Student's t distribution. The skewed student's t density function can be expressed as [19]

$$p(z|\eta,f)=\frac{2}{\eta+\eta^{-1}}\left\{f(\frac{z}{\eta})I_{[0,\infty)}(z)+f(\eta z)I_{(-\infty,0]}(z)\right\}$$

where f is a univariate pdf that is symmetric around 0, I_S is the indicator function on S, η is the asymmetric parameter, and $\eta=1$ for the symmetric Student's t distribution. In addition, the correlated random variables can be flexible and easily estimated under an overwhelming feature of Copula-based ARMA-EGARCH model.

2.2.3 Sklar's Theory

Sklar's Theorem [5] states that given random variables $X_1, X_2, \ldots, X_n$ with continuous distribution functions $F_1, F_2, \ldots, F_n$ and joint distribution function H, and there exists a unique copula C such that for all $x=(x_1, x_2, \ldots, x_n)\in\Re^n$

$$H(x)=C(F_1(x_1), F_2(x_2), \ldots, F_n(x_n))$$

If the joint distribution function is n-times differentiable, then taking the nth cross-partial derivative of the equation:

$$\begin{aligned}f(x_1, x_2, \ldots, x_n)&=\frac{\partial^n}{\partial x_1\ldots\partial x_n}H(x)\\&=\frac{\partial^n}{\partial u_1\ldots\partial u_n}C(F_1(x_1), \ldots, F_n(x_n))\cdot\prod_{i=1}^{n}f_i(x_i)\\&=c(F_1(x_1), \ldots, F_n(x_n))\cdot\prod_{i=1}^{n}f_i(x_i)\end{aligned}$$

where u_i is the probability integral transform of x_i.

For the purpose of estimating the VaR or ES based on time series data, Patton [6] defined the conditional version of Sklar's theorem. Let $F_{1,t}$ and $F_{2,t}$ be the continuous conditional distributions of $x_1|\mathfrak{F}_{t-1}$ and $x_2|\mathfrak{F}_{t-1}$, given the conditioning set $\mathfrak{F}_{t-1}$, and let H_t be the joint conditional bivariate distribution of $(X_1, X_2|\mathfrak{F}_{t-1})$. Then, there exists a unique conditional copula C_t such that

$$H_t(x_1, x_2|\mathfrak{F}_{t-1})=C_t(F_{1,t}(x_1|\mathfrak{F}_{t-1}), F_{2,t}(x_2|\mathfrak{F}_{t-1})|\mathfrak{F}_{t-1})$$

2.2.4 Parametric Copulas

Joe [9] and Nelsen [20] gave comprehensive copula definitions for each family.

(1) The bivariate Gaussian copula is defined as:

$$C(u_1, u_2; \rho) = \Phi_\rho(\Phi^{-1}(u_1), \Phi^{-1}(u_2))$$

where Φ_ρ is the bivariate joint normal distribution with linear correlation coefficient ρ and Φ is the standard normal marginal distribution.

(2) The bivariate student's t copula is defined by the following:

$$C(u_1, u_2; \rho, \nu) = t_{\rho,\nu}(t_\nu^{-1}(u_1), t_\nu^{-1}(u_2))$$

where ρ is the linear correlation coefficient and ν is the degree of freedom.

(3) The Clayton generator is given by $\varphi(u) = u^{-\theta} - 1$ with $\theta \in (0, \infty)$, its copula is defined by

$$C(u_1, u_2; \theta) = (u_1^{-\theta} + u_2^{-\theta} - 1)^{-1/\theta}$$

(4) The Gumbel generator is given by $\varphi(u) = (-\ln u)^\theta$ with $\theta \in [1, \infty)$, and the bivariate Gumbel copula is given by

$$C(u_1, u_2; \theta) = \exp(-\left\{(-\ln u_1)^\theta + (-\ln u_2)^\theta\right\}^{1/\theta})$$

(5) The Frank generator is given by $\varphi(u) = \ln[(e^{-\theta u} - 1)/(e^{-\theta} - 1)]$ with $\theta \in (-\infty, 0) \cup (0, \infty)$, and the bivariate Frank copula is defined by

$$C(u_1, u_2; \theta) = -\frac{1}{\theta}\ln(1 + \frac{\left(e^{-\theta u_1} - 1\right)\left(e^{-\theta u_2} - 1\right)}{e^{-\theta} - 1})$$

(6) The Joe generator is $\varphi(u) = u^{-\theta} - 1$, and the Joe copula is given by

$$C(u_1, u_2; \theta) = 1 - (\overline{u_1}^\theta + \overline{u_2}^\theta - \overline{u_1}^\theta \overline{u_2}^\theta)^{1/\theta}, \text{with } \theta \in [1, \infty)$$

(7) The BB1 (Clayton-Gumbel) copula with $\theta \in (0, \infty) \cap \delta \in [1, \infty)$ is

$$C(u_1, u_2; \theta, \delta) = \left(1 + \left[\left(u_1^{-\theta} - 1\right)^{\delta} + \left(u_2^{-\theta} - 1\right)^{\delta}\right]^{1/\delta}\right)^{-1/\theta}$$

(8) The BB6 (Joe-Gumbel) copula with $\theta \in [1, \infty) \cap \delta \in [0, \infty)$ is

$$C(u_1, u_2; \theta, \delta) = 1 - \left(1 - \exp\left\{-\left[\left(-\ln\left(1 - \overline{u}_1^{\theta}\right)\right)^{\delta} + \left(-\ln\left(1 - \overline{u}_2^{\theta}\right)\right)^{\delta}\right]^{1/\delta}\right\}\right)^{1/\theta}$$

(9) The BB7 (Joe-Clayton) copula with $\theta \in [1, \infty) \cap \delta \in [0, \infty)$ is

$$C(u_1, u_2; \theta, \delta) = 1 - \left(1 - \left[\left(1 - \overline{u}_1^{\theta}\right)^{-\delta} + \left(1 - \overline{u}_2^{\theta}\right)^{-\delta} - 1\right]^{-1/\delta}\right)^{1/\theta}$$

(10) The BB8 (Frank-Joe) copula with $\theta \in [1, \infty) \cap \delta \in (0, 1]$ is

$$C(u_1, u_1; \theta, \delta) = \frac{1}{\delta}\left(1 - \left[1 - \frac{1}{1 - (1 - \delta)^{\theta}}\left(1 - (1 - \delta u_1)^{\theta}\right)\left(1 - (1 - \delta u_2)^{\theta}\right)\right]^{1/\theta}\right)$$

2.2.5 Vine Copulas

Even though it is simple to generate multivariate Archimedean copulas, they are limited in that there are only one or two parameters to capture the dependence structure. Vine copula method allows a joint distribution to be built from bivariate and conditional bivariate copulas arranged together according to the graphical structure of a regular vine, which is a more flexible measure to capture the dependence structure among assets. It is well known that any multivariate density function can be decomposed as

$$f(x_1, \ldots, x_n) = f(x_n | x_1, \ldots, x_{n-1}) \ldots f(x_3 | x_1, x_2) f(x_2 | x_1) f(x_1)$$

Moreover, the conditional densities can be written as copula functions. For instance, the first and second conditional density can be decomposed as

$$
\begin{aligned}
f(x_2|x_1) &= c_{1,2}(F_1(x_1), F_2(x_2)) \cdot f_2(x_2), \\
f(x_3|x_1, x_2) &= c_{2,3|1}(F_{2|1}(x_2|x_1), F_{3|1}(x_3|x_1)) \cdot f_3(x_3|x_1) \\
&= c_{2,3|1}(F_{2|1}(x_2|x_1), F_{3|1}(x_3|x_1)) \cdot c_{1,3}(F_1(x_1), F_3(x_3)) \cdot f_3(x_3)
\end{aligned}
$$

After rearranging the terms, the joint density can be written as

$$
\begin{aligned}
f(x_1, x_2, x_3) = c_{2,3|1}(F_{2|1}(x_2|x_1), F_{3|1}(x_3|x_1)) \cdot c_{1,2}(F_1(x_1), F_2(x_2)) \\
\cdot c_{1,3}(F_1(x_1), F_3(x_3)) \cdot f_1(x_1) \cdot f_2(x_2) \cdot f_3(x_3)
\end{aligned}
$$

The summary of vine copulas is given by Kurowicka and Joe [21], and the general n-dimensional canonical vine copula, in which one variable plays a pivotal role, can be written as

$$
f(x_1, \ldots, x_n) = \prod_{k=1}^{n} f(x_k) \times \prod_{j=1}^{n-1} \prod_{i=1}^{n-j} c_{j,j+i|1,\ldots,j-1}
$$

Similarly, D-vines are constructed by choosing a specific order for the variables, and the general n-dimensional D-vine copula can be written as

$$
f(x_1, \ldots, x_n) = \prod_{k=1}^{n} f(x_k) \times \prod_{j=1}^{n-1} \prod_{i=1}^{n-j} c_{i,i+j|i+1,\ldots,i+j-1}
$$

Dissmann et al. [10] proposed that the automated algorithm involves searching for an appropriate R-vine tree structure, the pair-copula families, and the parameter values of the chosen pair-copula families based on AIC, which is summarized in Table 2.1.

Table 2.1 Sequential method to select an R-Vine model

Algorithm. Sequential method to select an R-Vine model
1. Calculate the empirical Kendall's tau for all possible variable pairs.
2. Select the tree that maximizes the sum of absolute values of Kendall's taus.
3. Select a copula for each pair and fit the corresponding parameters based on AIC.
4. Transform the observations using the copula and parameters from Step 3. To obtain the transformed values.
5. Use transformed observations to calculate empirical Kendall's taus for all possible pairs.
6. Proceed with Step 2. Repeat until the R-Vine is fully specified.

2.2.6 Tail Dependence

Tail dependence looks at the concordance and discordance in the tail, or extreme values of u_1 and u_2. It concentrates on the upper and lower quadrant tails of the joint distribution function. Given two random variables $u_1 \sim F_1$ and $u_2 \sim F_2$ with copula C, the coefficients of tail dependency are given by [22, 20]

$$\lambda_L \equiv \lim_{u \to 0^+} P[F_1(u_1) < u | F_2(u_2) < u] = \lim_{u \to 0^+} \frac{C(u,u)}{u},$$
$$\lambda_U \equiv \lim_{u \to 1^-} P[F_1(u_1) > u | F_2(u_2) > u] = \lim_{u \to 1^-} \frac{C(u,u)}{u}$$

where C is said to have lower (upper) tail dependency *iff* $\lambda_L \neq 0$ ($\lambda_U \neq 0$). The interpretation of the tail dependency is that it measures the probability of two random variables both taking extreme values as shown in Table 2.2 [22, 20].

2.2.7 Estimation Method

Generally, the two-step separation procedure is called the Inference functions for the margin method (IFM) [9]. It implies that the joint log-likelihood is simply the sum of univariate log-likelihoods and the copula log-likelihood is shown as below.

$$\log f(x) = \sum_{i=1}^{n} \log f_i(x_i) + \log c(F_1(x_1), \ldots, F_n(x_n))$$

Table 2.2 The coefficients of tail dependency

Family	Lower tail dependence	Upper tail dependence
Gaussian	–	–
Student's t	$2t_{\nu+1}\left(-\sqrt{\frac{(\nu+1)(1-\theta)}{1+\theta}}\right)$	$2t_{\nu+1}\left(-\sqrt{\frac{(\nu+1)(1-\theta)}{1+\theta}}\right)$
Clayton	$2^{-1/\theta}$	–
Gumbel	–	$2-2^{1/\theta}$
Frank	–	–
Joe	–	$2-2^{1/\theta}$
BB1	$2^{-1/\theta\delta}$	$2-2^{1/\delta}$
BB6	–	$2-2^{1/\theta\delta}$
BB7	$2^{-1/\delta}$	$2-2^{1/\theta}$
BB8	–	$2-2^{1/\theta}$ if $\delta=1$, otherwise 0

Note:—represents that there is no tail dependency

Therefore, it is convenient to use this two-step procedure to estimate the parameters by maximum log-likelihood, where marginal distributions and copulas are estimated separately.

2.3 Data and Empirical Findings

2.3.1 Data Representation

We use indices prices instead of other financial instruments or financial accounting numbers. One of the main reasons is that an index price could reflect a timely financial environment in contrast to financial accounting numbers that are published quarterly. Furthermore, indices can easily be constructed and tell us which sector contributes more risk to the entire market. Standard and Poor separates the 500 members in the S&P 500 index into 10 different sector indices based on the Global Industrial Classification Standard (GICS). All data is acquired from Bloomberg, sampled at daily frequency from January 1, 1995 to June 5, 2009. We separate the sample into two parts, the in-sample estimation period is from January 1, 1995 to December 31, 2007 (3271 observations) and the out-of-sample forecast validation period is from January 1, 2008 to June 5, 2009 (360 observations). Since the parameters change over time, we recalibrate the parameters every ten steps in the out-sample dataset. In other words, we calibrate the parameters 36 times. The following tables only show the results for the first calibration.

The summary statistics of these indices are listed in Table 2.3 as well as the statistical hypothesis testing. The statistical hypothesis testing for the unit-root based on Augmented Dickey-Fuller (ADF) test, and the results show that the values of 1 in ADF test rejects the null hypothesis of a unit root in a univariate time series. The results of Jarque-Bera (J-B) test reject that the distributions of returns are normality, and the results of Engle's ARCH test show that the indices' returns present conditional heteroscedasticity at the 5 % significance level. In addition, we assign an identification number to each sector.

2.3.2 Results for the Marginal Models

We estimate the parameters of p and q by minimizing Akaike information criterion (AIC) values for possible values ranging from zero to five. Table 2.4 lists the parameters which are estimated by minimized AIC values, and statistical hypothesis tests for the residuals are based on the Jarque-Bera (J-B) test and the Engle's ARCH

Table 2.3 Summary statistics of the in-sample dataset and statistical hypothesis testings

ID	Sector	Mean (%)	Sigma (%)	Skew	Kurt	ADF test	J-B test	ARCH test	Min/Max
1	S5FINL Index Financials	0.04	1.41	0.073	6.078	1	1	1	−8.04 %/8.39 %
2	S5INFT Index Technology	0.04	1.99	0.183	6.775	1	1	1	−10.01 %/ 16.08 %
3	S5COND Index Consumer Discretionary	0.03	1.24	−0.147	8.231	1	1	1	−10.33 %/8.47 %
4	S5ENRS Index Energy	0.06	1.39	−0.089	4.648	1	1	1	−7.21 %/7.94 %
5	S5HLTH Index Health Care	0.04	1.21	−0.180	7.097	1	1	1	−9.17 %/7.66 %
6	S5INDU Index Industrials	0.04	1.18	−0.227	7.410	1	1	1	−9.60 %/7.21 %
7	S5UTIL Index Utilities	0.02	1.12	−0.409	9.608	1	1	1	−9.00 %/8.48 %
8	S5CONS Index Consumer Staples	0.03	0.97	−0.233	9.905	1	1	1	−9.30 %/7.59 %
9	S5MATR Index Materials	0.03	1.31	0.036	5.929	1	1	1	−9.12 %/6.98 %
10	S5TELS Index Telecommunication Services	0.02	1.44	−0.100	6.674	1	1	1	−10.32 %/8.03 %
11	S&P 500 Index	0.04	1.07	−0.136	6.438	1	1	1	−7.11 %/5.57 %

Table 2.4 Estimation of in-sample parameters and statistical hypothesis testings for each marginal for the first calibration

	1	2	3	4	5	6	7	8	9	10	11
p	5	5	4	4	3	4	4	5	3	4	4
q	4	2	2	3	5	4	2	4	3	3	4
φ_1	0.4510	1.8223	1.8961	1.6080	−0.8015	0.8777	1.5697	0.1674	−0.7647	0.7181	−0.1260
φ_2	0.4420	−1.0559	−1.0770	−1.5377	0.2459	0.1199	−1.0570	0.6270	−0.3414	0.7243	1.3587
φ_3	0.5891	0.1041	0.1001	0.6024	0.1156	0.8492	0.0708	−0.6045	−0.0104	−0.6112	0.2708
φ_4	−0.8651	−0.0666	−0.0445	−0.0671	*	−0.9740	−0.0332	−0.4271	*	0.0066	−0.5033
φ_5	0.0305	0.0110	*	*	*	*	*	−0.0677	*	*	*
θ_1	−0.4268	−1.8192	−1.8833	−1.6313	0.8216	−0.8771	−1.5463	−0.1991	0.7991	−0.7441	0.1028
θ_2	−0.4789	0.9966	1.0042	1.5122	−0.2823	−0.1379	0.9896	−0.6501	0.3417	−0.7061	−1.4048
θ_3	−0.6191	*	*	−0.5585	−0.2107	−0.8330	*	0.6227	0.0391	0.5958	−0.2610
θ_4	0.8570	*	*	*	−0.0439	0.9744	*	0.4464	*	*	0.5640
θ_5	*	*	*	*	−0.0197	*	*	*	*	*	*
μ	0.0004	0.0005	0.0002	0.0006	0.0004	0.0004	0.0003	0.0004	0.0004	0.0004	0.0006
κ	−0.0974	−0.0597	−0.1056	−0.1281	−0.0825	−0.1404	−0.1566	−0.1136	−0.1083	−0.0655	−0.1502
α	−0.0723	−0.0539	−0.0907	−0.0383	−0.0479	−0.0891	−0.0325	−0.0552	−0.0480	−0.0315	−0.1062
β	0.9890	0.9928	0.9884	0.9853	0.9909	0.9847	0.9832	0.9882	0.9878	0.9925	0.9841
ξ	0.1335	0.1253	0.1159	0.1115	0.1184	0.1064	0.1730	0.1256	0.1264	0.1133	0.1016
η	0.9681	0.9080	0.9179	0.9289	0.9475	0.9357	0.8963	0.9677	1.0031	1.0001	0.8934
ν	9.9165	14.5997	10.4272	15.8222	7.9741	10.5557	9.2773	8.0843	9.3139	8.4905	9.8285
LLH	9897.7	8810.4	10304.2	9606.4	10281.9	10397	10658.2	11036.6	9903.6	9721.9	10734
AIC	−19763	−17593	−20582	−19185	−20534	−20764	−21290	−22041	−19781	−19415	−21438
J-B test	1	1	1	1	1	1	1	1	1	1	1
ARCH test	0	0	0	0	0	0	0	0	0	0	0

test. The results show that using the skewed Student's t innovation distribution for the residual terms is appropriately fitted to the return data because the degree of freedom is usually smaller than 15 and Jarque-Bera test rejects the null hypothesis of normality. In addition, the asymmetric parameter is not equal to one. Using EGARCH (1, 1) model is appropriate because the result of the Engle's ARCH test of residuals shows no conditional heteroscedasticity, and parameter β is usually larger than 0.9, which indicates the conditional volatility is time-dependent.

2.3.3 Results for the Copula Models

After the estimation of each marginal, we consider the set of standardized residuals from the ARMA-EGARCH (1, 1) model and transform them into a set of uniform variables. All the transformed residuals of the Kolmogorov-Smirnov test are 0, indicating that the distribution of transformed residuals and the uniform distribution are from the same continuous distribution, which null hypothesis cannot be rejected at the 5 % significance level. Using the Student's t copula as our benchmark with a degree of freedom of 10.1289, AIC values in Table 2.5 show that vine copula-based model outperforms Student's t copula-based model in high-dimensional modeling.

The catalogue of pair-copula families includes elliptical copulas such as Gaussian and Student's t, single parameter Archimedean copulas such as Clayton, Frank, and Gumbel, alongside two parameter families such as BB1, BB6, BB7, and BB8. All various copulas implemented are in the VineCopula library in R [23].

2.3.4 Results for the Copula VaR and Copula VaR Ratio

We empirically examine which sector dominates more risk contributions on systemic risk with 10,000 Monte Carlo simulations using vine Copula-based ARMA-EGARCH (1, 1) modeling. The results of residuals, fitted by ARMA-EGARCH (1, 1) with the skewed student's t innovations, are shown in Fig. 2.1. Using the vine copula, the tree structure in tree level one is shown in Fig. 2.2. The results of the worst 5 % return loss, shown in Fig. 2.3, are not surprising that the financial industry has the largest risk during 2008 financial crisis. As

Table 2.5 Estimation for the copula models from in-sample data for the first calibration

	Number of parameters	Log-likelihood	AIC
Gaussian copula	55	15259	−30408.4
Student's t copula	56	16074	−32038
Vine copula	96	16242	−32292.6

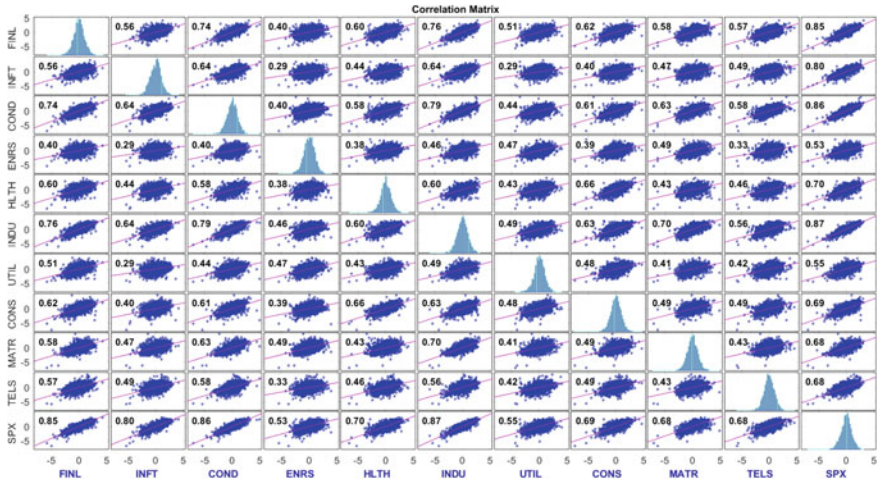


Fig. 2.1 The scatter plot of residuals for the first calibration

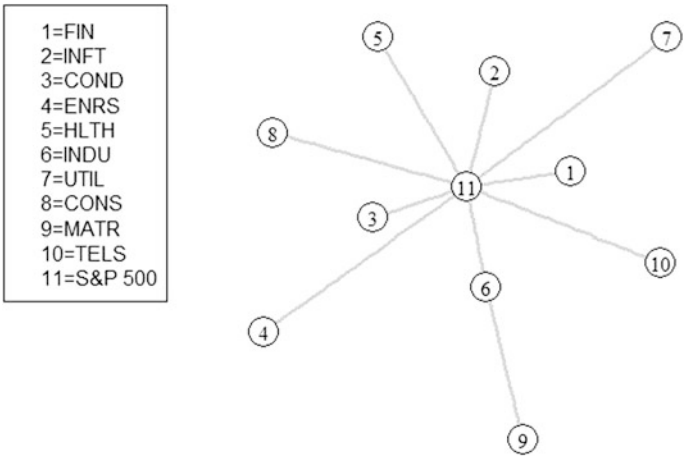


Fig. 2.2 The plot of vine copula structure in tree one for the first calibration

seen in Fig. 2.4 and Fig. 2.5, we realize that the financial sector contributed more risk during the subprime crisis from 2008 to 2009, while the consumer staples sector is the major risk receiver. Hence, this measure is a simplified and efficient methodology to analyze systemic risk.

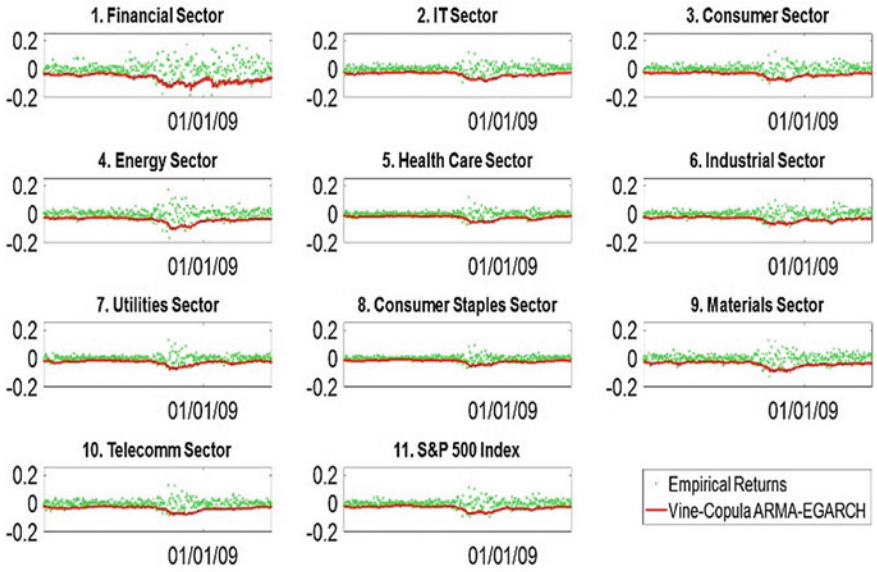


Fig. 2.3 The one-day ahead worst 5 % return loss for each sector index

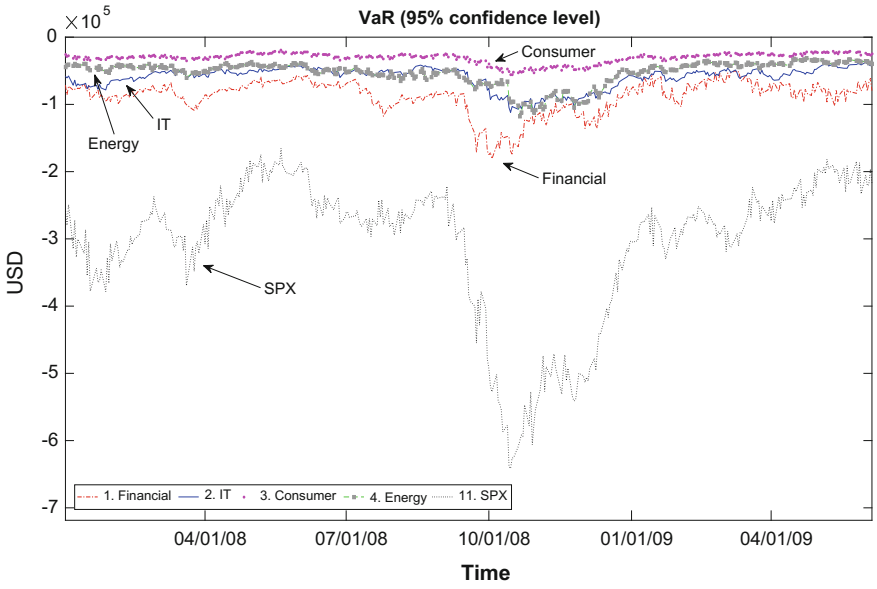


Fig. 2.4 The one-day ahead VaR at the 95 % confidence level

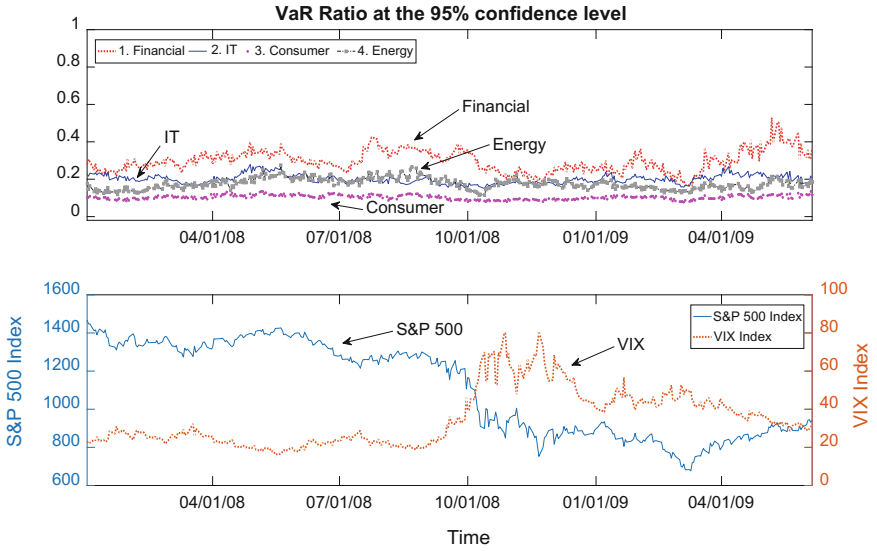


Fig. 2.5 The one-day ahead VaR ratio at the 95 % confidence level

2.4 Conclusion

The evidence in our paper shows that not only does vine Copula-based ARMA-EGARCH (1, 1) outperform the Gaussian and the Student's t copula-based ARMA-EGARCH (1, 1) based on AIC values, but also is the skewed student's t innovations more appropriate than using the student's t innovations.

In addition, using vine Copula-based ARMA-EGARCH model to forecast Copula VaR and Copula VaR ratio, we develop a real-time and useful way to evaluate systemic risk. Moreover, the VaR ratio provides the information of the risk contribution from each sector. This approach is very general and can be tailored to any underlying country and financial market easily.

References

1. Board, Financial Stability. International Monetary Fund and Bank for International Settlements (2009) Guidance to Assess the Systemic Importance of Financial Institutions, Markets and Instruments: Initial Considerations. (Background Paper. Report to the G-20 Finance Ministers and Central Bank Governors (October))
2. Girardi G, Tolga Ergün A (2013) Systemic risk measurement: multivariate GARCH estimation of CoVaR. J Bank Finance 37(8):3169–3180
3. Adrian T, Brunnermeier MK (2011) CoVaR. No. w17454. National bureau of economic research

4. Hakwa B, Jäger-Ambrożewicz M, Rüdiger B (2012) Measuring and analysing marginal systemic risk contribution using CoVaR: a Copula approach. arXiv preprint arXiv:1210.4713
5. Sklar M (1959) Fonctions de répartition à n dimensions et leurs marges. Université Paris 8
6. Patton AJ (2002) Applications of copula theory in financial econometrics. PhD Dissertation, University of California, San Diego
7. Otani Yuko, Imai Junichi (2013) Pricing portfolio credit derivatives with stochastic recovery and systematic factor. IAENG Int J Appl Math 43(4):176–184
8. Joe H (1996) Families of m-variate distributions with given margins and m (m-1)/2 bivariate dependence parameters. Lecture notes-monograph series 120–141
9. Joe H (1997) Multivariate models and multivariate dependence concepts. CRC Press
10. Dissmann J, Brechmann EC, Czado C, Kurowicka D (2013) Selecting and estimating regular vine copulae and application to financial returns. Comput Stat Data Anal 59:52–69
11. Rockinger M, Jondeau E (2001) Conditional dependency of financial series: an application of copulas
12. Lee Tae-Hwy, Long Xiangdong (2009) Copula-based multivariate GARCH model with uncorrelated dependent errors. J Econometrics 150(2):207–218
13. Fang Y, Madsen L, Liu L (2014) Comparison of two methods to check copula fitting. IAENG Int J Appl Math 44(1):53–61
14. Riskmetrics: technical document. Morgan Guaranty Trust Company of New York 1996
15. Bollerslev T (1986) Generalized autoregressive conditional heteroskedasticity. J Econometrics 31(3):307–327
16. Chen K-H, Khashanah Khaldoun (2015) Measuring systemic risk: vine copula-GARCH model. In: Proceedings of the world congress on engineering and computer science 2015. Lecture Notes in Engineering and Computer Science. San Francisco, USA, pp 884–889. Accessed 21–23 Oct 2015
17. Nelson DB (1991) Conditional heteroskedasticity in asset returns: a new approach. Econometrica: J Econometric Soc 347–370
18. Ghalanos A (2014) rugarch: Univariate GARCH models. R package version 1–3
19. Ferreira JTAS, Steel MFJ (2012) A constructive representation of univariate skewed distributions. J Am Stat Assoc
20. Nelsen RB (1999) An introduction to copulas, vol 139. Springer Science & Business Media
21. Kurowicka D, Joe H (2011) Dependence modeling-handbook on vine copulae
22. Cherubini U, Luciano E, Vecchiato W (2004) Copula methods in finance. John Wiley & Sons
23. Schepsmeier Ulf, Stoeber J, Brechmann EC, Graeler B (2012) VineCopula: statistical inference of vine copulas, 2012. http://CRAN.R-project.org/package=VineCopula. R package version: 1-1
24. Bedford T, Cooke RM (2002) Vines: a new graphical model for dependent random variables. Ann Stat 1031–1068

Chapter 3
MOESP_AOKI_VAR: Algorithm for Space State Identification of Non-stationary Multivariable Noisy Linear Systems

Johanna B. Tobar and Celso P. Bottura

3.1 Introduction

An initial study of different kind of systems for identification, based on the state space is performed. Additionally, a structure to be used in the problem resolution of computational modelling for non-stationary noisy linear systems is proposed. Through this study, non-stationary systems are treated as a group of invariant models with respect to the time. It is also considered that the matrix system A_K, B_K, C_K, D_K, presents small changes with respect to the time. This is translated into continuous and slow changes within the matrices and it allows for the generation of a recursive algorithm, which is the main objective of this study. A linear system is considered as the superposition of its deterministic and stochastic part. A MOESP_VAR algorithm is used for modelling the deterministic part, whilst the stochastic part is modelled by the use of AOKI_VAR algorithm. Finally, the algorithm is tested by using a benchmark [1].

3.2 Foundation

3.2.1 Stationary Deterministic Linear System

The representation of the deterministic linear system in state spaces has the following form:

J.B. Tobar (✉)
University of the Armed Forces, Sangolqui, Ecuador
e-mail: jbtobar@espe.edu.ec

C.P. Bottura
State University of Campinas, Campinas, Brazil
e-mail: cpbottura@fee.unicamp.br

S.-I. Ao et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-981-10-2717-8_3

$$\begin{cases} x_{k+1} = Ax_k + Bu_k \\ y_k = Cx_k + Du_k \end{cases} \tag{3.1}$$

where, $x\_k \in R^n$ is the state vector, $u\_k \in R^m$ is the input vector and $y\_k \in R^l$ is the output vector. The A, B, C and D matrices are considered constant for every k instant [2, 3].

3.2.2 Stationary Noisy Linear System

In order to quantify the uncertainty and external perturbations of the system, v_k *and* w_k are added in the the state-space equations. These terms are considered as inputs where non-control exists:

$$\begin{cases} x_{k+1} = Ax_k + Bu_k \\ y_k = Cx_k + Du_k \end{cases} \tag{3.2}$$

The perturbation vectors $v_k \in R^n$ and $w_k \in R^l$ are random variables with zero means. The sequence of $\left(v_k,\ k = 0, \pm 1, \pm 2, ...\right)$ *and* $\left(w_k,\ k = 0, \pm 1, \pm 2, ...\right)$ are considered stochastic processes of Gaussian white noise.
Additionally, the stochastic process can be represented by defining the error vector as $e_k = \begin{bmatrix} w_k \\ v_k \end{bmatrix}$, with $E\left[e_k\right] = 0 \quad \forall k$, and its innovation representation is as following:

$$\begin{cases} x_{k+1} = Ax_k + Ke_k \\ y_k = Cx_k + e_k \end{cases} \tag{3.3}$$

where, e_k is a white noise sequence and its covariance matrix is given by $\Delta = E\left(e_k e_k^T\right)$.

When referring to the covariance domain, the Markov parameters of the system can be represented as:

$$\Lambda_i = \begin{cases} C\Pi_0 C^T + R & i = 0 \\ G^T \left(A^T\right)^{-i-1} C^T & i < 0 \\ CA^{i-1}G & i \geq 1 \end{cases} \tag{3.4}$$

where G is also presented, $G = A\Pi_0 C^T + S$, yielding as result the following:

$$\begin{cases} R = \Lambda_0 - CPC^T \\ Q = P - APA^T \\ S = M - APC^T \end{cases} \tag{3.5}$$

The stochastic realization problem consists of finding one or more models in the state-space through process statistical data such us covariance. For further information read Caceres (2005), Tamariz (2005), Barreto [4].

3.2.3 Non-stationary Deterministic Linear System

A non-stationary deterministic linear system is represented by the following state-space equations:

$$\begin{cases} x_{j,k+1} = A_{j,k}x_{j,k} + B_{j,k}u_{j,k} \\ y_{j,k} = C_{j,k}x_{j,k} + D_{j,k}u_{j,k} \end{cases} \tag{3.6}$$

being $j \in [j_0, j_0 + n - 1]$ and $k \in [k_0, k_0 + T - 1]$, where j_0 is the first interval of experiment, k_0 is the first instant of the experiment, n is the total number of simple experiments and $T \geq n$. Equation (3.6) can also be expressed as:

$$y_H = O_k X_H + T_k U_H \tag{3.7}$$

To get detailed information about the process to obtain the (3.7), consult the matrices y_H, O_k, X_H, T_k y U_H on [5, 6].

3.2.4 Non-stationary Noisy Linear System

A non-stationary noisy linear system is expressed as follows:

$$\begin{cases} x_{j,k+1} = A_{j,k}x_{j,k} + B_{j,k}u_{j,k} + v_{j,k} \\ y_{j,k} = C_{j,k}x_{j,k} + D_{j,k}u_{j,k} + w_{j,k} \end{cases} \tag{3.8}$$

being $j \in [j_0, j_0 + n - 1]$ and $k \in [k_0, k_0 + T - 1]$, where j_0 is the first interval of the experiment, k_0 is the first instant of time of the experiment, n is the total number of simple experiments, $T \geq n$ and $v_{j,k} \in R^n$ and $w_{j,k} \in R^l$ are random variables of null arithmetic mean, and the sequences $(v_k, k = 0, \pm 1, \pm 2, \ldots)$ and $(w_k, k = 0, \pm 1, \pm 2, ...)$ are non-stationary stochastic processes that are generated by the non-stationary stochastic system represented by the state-space equation:

$$\begin{cases} x_{j,k+1} = A_{j,k}x_{j,k} + K_{j,k}e_{j,k} \\ y_{j,k} = C_{j,k}x_{j,k} + e_{j,k} \end{cases} \tag{3.9}$$

where, e_k is the white noise stochastic process.

3.2.5 Time Variant Identification

The identification algorithm works on the assumption that time-varying systems can be treated as a set of time-invariant models for a given time interval. Thus, the identification of time-varying systems consists of a set of n time-invariant models which describes the system for the defined experiment [7, 8].

The following expression relates the input variables to the state vector in a variant linear system in the Kth moment:

$$x_k = A_{(k-1)}x_0 + \sum_{l=0}^{k-1} A^{(k-l-1)}B_l u_l \tag{3.10}$$

where, $A_{(n)}$ and $A^{(n)}$ represent the transition matrices that satisfy:

$$\begin{cases} A_{(0)} = A_0 \\ A^{(0)} = I \\ A_{(n)} = A_n A_{(n-1)} = A_n \dots A_2 A_1 A_0 \\ A^{(n)} = A_n A^{(n-1)} = A_n \dots A_2 A_1 \end{cases} \tag{3.11}$$

A set of indices are stated to interpret the identification of problem parameters; additionally, the set of indices j, k on $u_{j,k}$ indicates the sample input for instant kth and for system experimentation range jth (3.8). Furthermore, $j \in \left[j_0, j_0 + n - 1\right]$ and $k \in \left[k_0, k_0 + T - 1\right]$, where j_0 is the first range of experimentation, k_0 is the first time instant, n is the total number of experiments or tests and T is the time required for a single experiment.

The problem to be solved is represented by the state space model:

$$\begin{bmatrix} x_{j,k+1} \\ y_{j,k} \end{bmatrix} = \begin{bmatrix} A_k & B_k \\ C_k & D_k \end{bmatrix} \begin{bmatrix} x_{j,k} \\ u_{j,k} \end{bmatrix} \tag{3.12}$$

based on the following output data sequences:

$$Y_{j,k} = \begin{bmatrix} y_{j_0,k_0} & y_{j_0,k_0+1} & \cdots & y_{j_0,k_0+T-1} \\ y_{j_0+1,k_0} & y_{j_0+1,k_0+1} & \cdots & y_{j_0+1,k_0+T-1} \\ \vdots & \vdots & \cdots & \vdots \\ y_{j_0+n-1,k_0} & y_{j_0+n-1,k_0+1} & \cdots & y_{j_0+n-1,k_0+T-1} \end{bmatrix} \tag{3.13}$$

The problem is also based on the input sequences $U_{j,k}$ for the same series of experiments and same time interval. The matrix $Y_{j,k}$ represents the set of $(n-1)$ intervals of experimentation. It allows to develop general expressions that govern the system and it also relates the inputs and outputs at a start time k_0 and establishes a correct experimentation interval j, for a discrete time variant system which is represented in state space by:

$$\begin{cases} x_{j,k+1} = A_k x_{j,k} + B_k u_{j,k} \\ y_{j,k} = C_k x_{j,k} + D_k u_{j,k} \end{cases} \tag{3.14}$$

for the next instant, the expression is:

$$\begin{cases} x_{j,k+2} = A_{k+1} x_{j,k+1} + B_{k+1} u_{j,k+1} \\ y_{j,k+1} = C_{k+1} x_{j,k+1} + D_{k+1} u_{j,k+1} \end{cases} \tag{3.15}$$

plugging in Eq. (3.15) into (3.14) yields:

$$\begin{cases} x_{j,k+2} = A_{k+1} A_k x_{j,k} + A_{k+1} B_k u_{j,k} + B_{k+1} u_{j,k+1} \\ yy_{j,k+1} = C_{k+1} A_k x_{j,k} + C_{k+1} B_k u_{j,k} + D_{k+1} u_{j,k+1} \end{cases} \tag{3.16}$$

and this process is repeated successively. Solving Eq. (3.14) for any moment of time $k_0 \geq 0$ the solution can be written as:

$$y_{j,l} = \begin{cases} C_{l,j} + D_l u_{j,l} & l = 0 \\ C_l A_{(l-1)} x_{j,l} + & \\ + \sum_{i=0}^{l-1} C_l A^{(l-i-1)} B_i u_{j,i} + D_l u_{j,l} & l > k_0 \end{cases} \tag{3.17}$$

Equation (3.14) can be rewritten in a shorter form by the extended model:

$$Y_N = O_k X_H + T_k U_H \tag{3.18}$$

3.2.5.1 Determination of the Extended Model

For an experiment j, the matrix U_H for the stationary case is equal to the matrix U_H from the time-varying case. Data from a single experiment is assumed. To solve the non-stationary discrete case is necessary to assume that data is available from a single experiment.

3.2.5.2 Recursive Algorithm for the Deterministic Part

A recursive algorithm is applied for time-varying systems that assumes small variations in a predefined range of operation in the system matrices. A recursive scheme is implemented to systems that vary slowly with time [9].

QR Factorization Update

1. Let's assume that the following measures were already processed by the algorithm:
   $\left[u_{j0,k0}\, u_{j0,k0+1} \ldots u_{j0,k0+T-1}\right]^T$
   and
   $\left[y_{j0,k0} \quad y_{j0,k0+1} \quad \cdots \quad y_{j0,k0+T-1}\right]^T$

2. Being QR factorization:

$$\begin{bmatrix} R_{j0,11} & 0 \\ R_{j0,21} & R_{j0,22} \end{bmatrix} \begin{bmatrix} Q_{j0,1} \\ Q_{j0,2} \end{bmatrix} \tag{3.19}$$

   where, j_0 represents the set of input-output values processed in the most recent experiment.
3. Finally, suppose that during the time interval $\left[j_0, j_0 + n - 1\right]$ the model in state space is invariant and equal to:

$$\begin{cases} x_{k+1} = A_{j0} x_k + B_{j0} u_k \\ y_k = C_{j0} x_k + D_{j0} u_k \end{cases} \tag{3.20}$$

   It can also be represented in a shorter form by relating the different data matrices as:

$$Y_H = \mathcal{O}_k X_H + \mathcal{T}_{j0,k} U_H \tag{3.21}$$

   Recursive algorithm for the stochastic: See Johanna and Bottura (2010)

3.2.6 Deterministic-Stochastic Modelling of the State Space of Non-stationary System

One way of representing non-stationary discrete multivariate noisy linear systems, with exogenous time variable inputs in the state space is:

$$\begin{cases} x_{j,k+1} = A_{j,k} x_{j,k} + B_{j,k} u_{j,k} + v_{j,k} \\ y_{j,k} = C_{j,k} x_{j,k} + D_{j,k} u_{j,k} + w_{j,k} \end{cases} \tag{3.22}$$

with

$$E\left[\begin{pmatrix} v_{j,k} \\ w_{j,k} \end{pmatrix} \begin{pmatrix} v_{j,s}^T & w_{j,s}^T \end{pmatrix}\right] = \begin{bmatrix} Q & S \\ S^T & R \end{bmatrix} \quad \begin{matrix} k = s \\ 0 \quad k \neq s \end{matrix} \tag{3.23}$$

A theorem for the decomposition by superposition of noisy time variant linear systems is proposed as follows:

Theorem 1 *By superposition, a variant time noisy linear model* S*, in an innovative form given by:*

$$S : \begin{cases} x_{j,k+1} = A_{j,k}x_{j,k} + B_{j,k}u_{j,k} + K_{j,k}e_{j,k} \\ y_{j,k} = C_{j,k}x_{j,k} + D_{j,k}u_{j,k} + e_{j,k} \end{cases} \tag{3.24}$$

can be decomposed into the following two subsystems:

$$S_d : \begin{cases} x^d_{j,k+1} = A_{j,k}x^d_{j,k} + B_{j,k}u_{j,k} \\ y^d_{j,k} = C_{j,k}x^d_{j,k} + D_{j,k}u_{j,k} \end{cases} \tag{3.25}$$

and

$$S_e : \begin{cases} x^e_{j,k+1} = A_{j,k}x^e_{j,k} + v_{j,k} \\ y^e_{j,k} = C_{j,k}x^e_{j,k} + w_{j,k} \end{cases} \tag{3.26}$$

where, superscript d and e refers to the deterministic and the stochastic subsystems S_d and S_e respectively and $y_{j,k} = y^d_{j,k} + y^e_{j,k}$. The noisy signal state is: $x_{j,k} = \begin{bmatrix} x^d_{j,k} \\ x^e_{j,k} \end{bmatrix}$ where,

$A = \begin{bmatrix} A^d & 0 \\ 0 & A^e \end{bmatrix}$

$B = \begin{bmatrix} B^d \\ 0 \end{bmatrix}$ $C = \begin{bmatrix} C^d & C^e \end{bmatrix}$

$D = D^d$, $K = K^e$

Proof Applying the principle of superposition, the linear systems can be written as follows:
$y_{j,k} = y^d_{j,k} + y^e_{j,k} = C^d_{j,k}x^d_{j,k} + C^e_{j,k}x^e_{j,k} + D^d_{j,k}u_{j,k} + K^e_{j,k}e_{j,k} + e_{j,k}$ with

$$y^d_{j,k} = C_{j,k}x^d_{j,k} + D_{j,k}u_{j,k}$$

and

$$y^e_{j,k} = C_{j,k}x^e_{j,k} + e_{j,k}$$

where

$$y_{j,k} = y^d_{j,k} + y^e_{j,k} = \begin{bmatrix} C^d_{j,k} & C^e_{j,k} \end{bmatrix} \begin{bmatrix} x^d_{j,k} \\ x^e_{j,k} \end{bmatrix} + D^d_{j,k}u_k + K^e_{j,k}e_{j,k} + e_{j,k}$$

The state vector of the noisy signal is:

$$x_k = \begin{bmatrix} x^d_k \\ x^e_k \end{bmatrix},$$

so

$$x_{j,k+1} = \begin{bmatrix} x^d_{k+1} \\ x^e_{k+1} \end{bmatrix} = \begin{bmatrix} A^d_{j,k} x^d_{j,k} + D^d_{j,k} u_{j,k} \\ A^e_{j,k} x^e_{j,k} + K^e_{j,k} e_{j,k} \end{bmatrix}$$
$$= \begin{bmatrix} A^d_{j,k} & 0 \\ 0 & A^e_{j,k} \end{bmatrix} \begin{bmatrix} x^d_{j,k} \\ x^e_{j,k} \end{bmatrix} + \begin{bmatrix} D^d_{j,k} \\ 0 \end{bmatrix} u_{j,k} + e_{j,k}$$

where the system equation is obtained from (3.24).

3.3 MOESP_AOKI_VAR Algorithm

The proposed algorithm MOESP_AOKI_VAR collects the MOESP_VAR proposed in [5, 10] and the MOESP_AOKI proposed in [5, 11] and the AOKI_VAR in [1, 12] respectively. Therefore, the algorithm MOESP_AOKI_VAR is as follows:

1. Obtain the Hankel matrices Y_H and U_H
2. Perform QR factorization:

$$\left[\frac{U_H}{Y_H}\right] = \left[\frac{R_{11} \mid 0}{R_{21} \mid R_{22}}\right] \left[\frac{Q_1}{Q_2}\right] \tag{3.27}$$

   where R_{11} and R_{22} are invertible square matrices.
3. Compute SVD of R_{22} as:

$$R_{22} = \left[U_H \middle| U_H^{\perp} \right] \left[\frac{\sum_n \mid 0}{0 \mid \sum_2}\right] \left[\frac{V_n^T}{\left(V_n\right)^T}\right] \tag{3.28}$$

4. Solve the equation system:

$$U_H^{(1)} A_T = U_H^{(2)} \tag{3.29}$$

5. Update: The recursive algorithm shown in Sect. 3.2.5.2 is applied
   Thus, obtaining the matrices $A^d_{j,k}, B^d_{j,k}, C^d_{j,k}, D^d_{j,k}$, of $y^d_{j,k}$ for each k time instants and j intervals with respect to the time.
6. Determine the signal generated by the matrices $H^A, H^M, H^C, H, Y_-, Y_+$

$$Y_- = \begin{bmatrix} \bar{y_1} & \bar{y_2} & \bar{y_3} & \cdots & \bar{y_{N-1}} \\ 0 & \bar{y_1} & \bar{y_2} & \cdots & \bar{y_{N-2}} \\ 0 & 0 & \bar{y_1} & \cdots & \bar{y_{N-3}} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \ldots & \bar{y_{N-k-1}} & \bar{y_{N-k}} \end{bmatrix}$$

$$H = \frac{Y_+ Y_-^T}{N} = \begin{bmatrix} \Lambda_1 & \Lambda_2 & \cdots & \Lambda_k \\ \Lambda_2 & \Lambda_3 & \cdots & \Lambda_{k+1} \\ \vdots & \vdots & \ddots & \vdots \\ \Lambda_j & \Lambda_{j+1} & \cdots & \Lambda_{j+k} \end{bmatrix}$$

$$H^A = \begin{bmatrix} \Lambda_2 & \Lambda_3 & \cdots & \Lambda_{k+1} \\ \Lambda_3 & \Lambda_4 & \cdots & \Lambda_{k+2} \\ \vdots & \vdots & \ddots & \vdots \\ \Lambda_{j+1} & \Lambda_{j+2} & \cdots & \Lambda_{j+k+1} \end{bmatrix}$$

$$H^M = \begin{bmatrix} \Lambda_1 \\ \Lambda_2 \\ \vdots \\ \Lambda_j \end{bmatrix}$$

$$H^C = \begin{bmatrix} \Lambda_1 & \Lambda_2 & \cdots & \Lambda_k \end{bmatrix}$$

7. Obtain the SVD for the covariance Hankel matrix

$$H = U \Sigma^{1/2} \Sigma^{1/2} V^T$$

8. Calculate the matrices $A^e_{j,k}, C^e_{j,k}, K^e_{j,k}$
9. Validate.

3.4 Experimentation and Results

The proposed algorithm is initially defined for T intervals of experimentation thus getting the system matrices identification. The presented MOESP_AOKI_VAR is assessed T times in order to determine T sets of matrices corresponding to each experiment.

If $\triangle$ represents small increments, L_j is an integer number for each I_j experimentation interval stated by:

$$I_j = [k_j - L_j \triangle, \quad k_j + L_j \triangle] \tag{3.30}$$

to validate the proposed algorithm a benchmark is implemented [13]. The identification at time instant k_j (which is the middle point of each interval I_j) is determined as $k_{j+1} = k_j + v\triangle$, where v is an integer number, ∇ represents an increment with respect to the simulation time and it is given by $\triangle = M \triangle_j$, where $\triangle_j$ is the j-th sampling period and M is an integer number. $L_j = 500 \; \forall \, j$ is defined for this study.

The benchmark system is:
The deterministic part is given by the following matrices

$$A_k = \begin{bmatrix} -0.3 & a_k \\ 1 & -1 \end{bmatrix} \tag{3.31}$$

where,

$$a_k = -\frac{1}{3} - \frac{1}{10} sen(\frac{2\pi k}{400})$$

and the remaining matrices are considered constants:

$$B_k = \begin{bmatrix} -2 & 1 \\ 1 & 1 \end{bmatrix}; \qquad C_k = \begin{bmatrix} 1 & 3 \\ 1 & 2 \end{bmatrix} \tag{3.32}$$

$$D_k = 0$$

The system input is randomically changing for each iteration of the algorithm. The proposed algorithm presents the following results for $k = 1$:

$$A_{j,k} = \begin{bmatrix} -0.3000 & 0 \\ 0 & -0.4040 \end{bmatrix}$$

$$B_{j,k} = \begin{bmatrix} -2 & 1 \\ 1 & 1 \end{bmatrix}$$

$$C_{j,k} = \begin{bmatrix} 1 & 3 \\ 1 & 2 \end{bmatrix}$$

$$C_{j,k}B_{j,k} = \begin{bmatrix} 1 & 4 \\ 0 & 3 \end{bmatrix}$$

$$C_{j,k}A_{j,k}B_{j,k} = \begin{bmatrix} -0.6121 & -1.5121 \\ -0.2081 & -1.1081 \end{bmatrix}$$

The deterministic-stochastic identification of the system under noise is realized. The obtained results after running the first part of the algorithm MOESP_AOKI_VAR are shown:

$$A^d_{j,k} = \begin{bmatrix} -0.3000 & 0 \\ -0.0000 & -0.4077 \end{bmatrix}$$

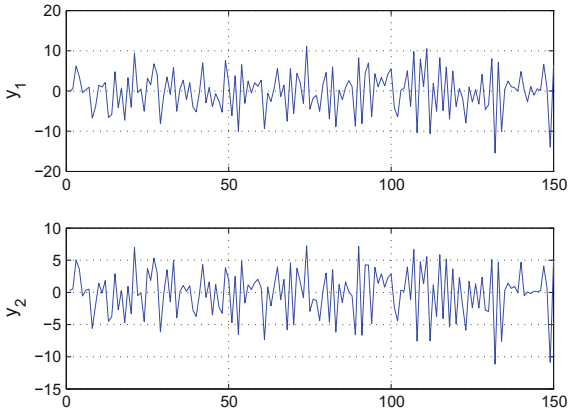


Fig. 3.1 Output of the deterministic subsystem

$$B^d_{j,k} = \begin{bmatrix} 4.1559 & 3.7636 \\ -3.1924 & 1.5997 \end{bmatrix}$$

$$C^d_{j,k} = \begin{bmatrix} 0.7705 & 0.6867 \\ 0.5137 & 0.6677 \end{bmatrix}$$

$$C^d_{j,k} B^d_{j,k} = \begin{bmatrix} 1.0098 & 3.9984 \\ 0.0031 & 3.0014 \end{bmatrix}$$

$$C^d_{j,k} A^d_{j,k} B^d_{j,k} = \begin{bmatrix} -0.6478 & -1.5118 \\ -0.2308 & -1.1086 \end{bmatrix}$$

Figure 3.1 shows the output of the deterministic model

The results after running the second part of the algorithm MOESP_AOKI_VAR are:

$$\Delta = [1.6092]$$

$$A^e_{j,k} = \begin{bmatrix} -0.2092 & -0.7319 \\ 0.7319 & 0.1218 \end{bmatrix}$$

$$K_{j,k} = \begin{bmatrix} -0.2582 \\ 1.1257 \end{bmatrix}$$

$$C^e_{j,k} = \begin{bmatrix} -2.7437 & 1.0369 \end{bmatrix}$$

where Δ is the covariance matrix of the noise.

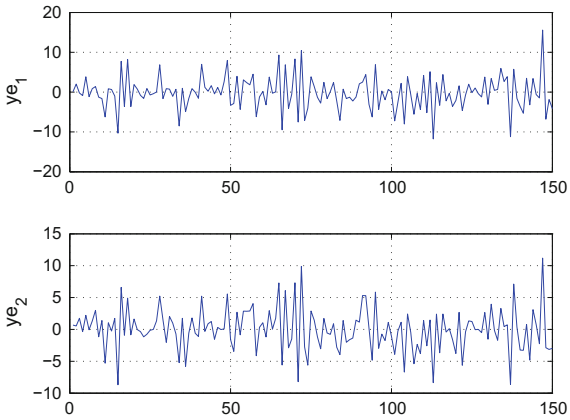


Fig. 3.2 Output of the stochastic subsystem

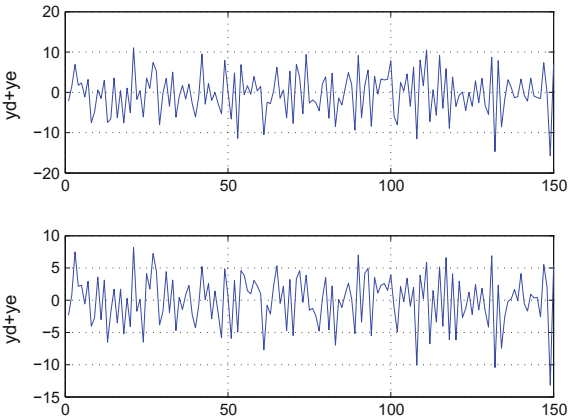


Fig. 3.3 Overlapping signals

Figure 3.2 presents the stochastic modeled signal by using the second part of the algorithm MOESP_AOKI_VAR.

Figure 3.3 displays the overlapped signal of the deterministic and stochastic output signals $y^d_{j,k}$ and $y^e_{j,k}$. Finally, a verification and validation process of the proposed combined algorithm **MOESP_AOKI_VAR** is performed and Fig. 3.4 shows the results. The algorithm **MOESP_AOKI_VAR** describes satisfactorily the noisy signal.

Finally Figs. 3.5 and 3.6 show an example of noise innovated identification and calculation of their error respectively, for further illustration.

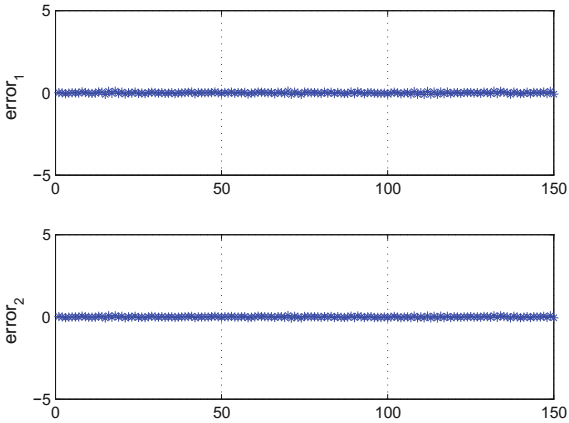


Fig. 3.4 Error result

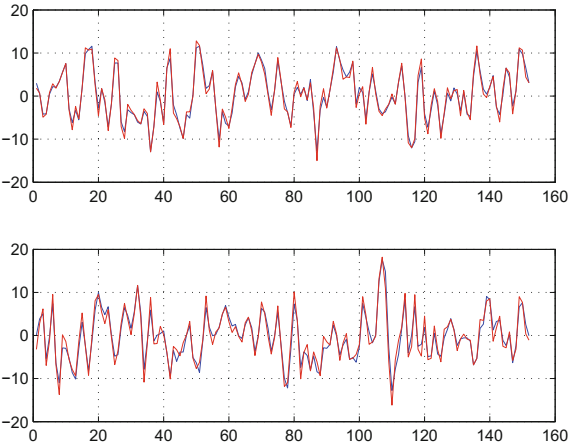


Fig. 3.5 Output of noise innovated

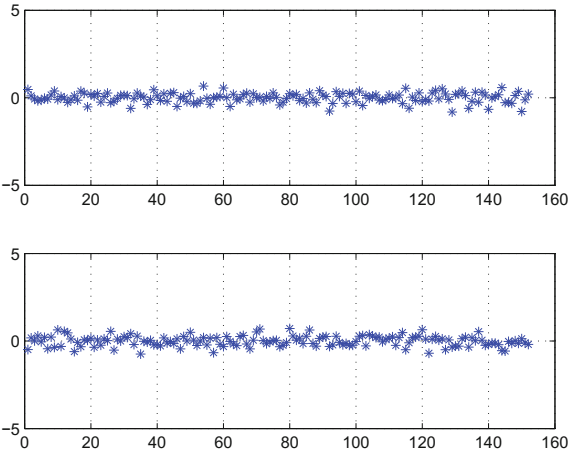


Fig. 3.6 Error identification noise innovated

3.5 Conclusions

A data computational procedure is formulated for the identification of discrete stochastic multivariate linear systems time variant. The considered hypothesis is that the variation in time of the dynamic system is sufficiently slow to guarantee the efficiency of the subspace identification MOESP for a non-stationary noisy linear system. The algorithm MOESP_AOKI_VAR has its foundation in AOKI and MOESP algorithms and it is given by the superposition of them. The obtained results describe that the proposed method guarantee an appropriate behaviour when modeling this kind of systems.

References

1. Johanna BT, Bottura CP (2015) Data computational modelling of multivariable non-stationary noisy linear systems by MOESP_AOKI_VAR algorithm. In: Proceedings of the world congress on engineering and computer science 2015, WCECS 2015. Lecture notes in engineering and computer science. San Francisco, USA, pp 896–900. Accessed 21–23 Oct 2015
2. Ohsumi A, Kawano T (2002) Subspace identification for a class of time-varying continuous-time stochastic systems via distribution based approach. In: Proceedings 15th IFAC world congress. Barcelona, July 2002
3. Dewilde P, Van der Veen AJ (1998) Time-varying systems and computations. Kluwer Academic Publishers
4. Barreto G (2002) Modelagem Computacional Distribuda e Paralela de Sistemas e de Sries Temporais Multivariveis no Espao de Estado. Tese de Doutorado, Faculdade de Engenharia Eltrica e de Computao da Universidade Estadual de Campinas, UNICAMP
5. Caceres AFT (2005) Identificao e Controle Estocsticos Decentralizados de Sistemas Interconectados Multivariveis no Espao de Estado. Tese de Doutorado a ser apresentada na FEEC da UNICAMP, em 2005
6. Shokoohi S, Silverman LM (1987) Identification and model reduction of time-varying discrete-time systems. Automatica 23:509–522
7. Tamariz ADR, Bottura CP, Barreto G (2003) Algoritmo Iterativo do tipo MOESP para Identificao de sistemas discretos variantes no tempo - Parte I: Formula, DINCON
8. Tamariz ADR, Bottura CP, Barreto G (2003) Algoritmo Iterativo do tipo MOESP para Identificao de sistemas discretos variantes no tempo - Parte II: Implementa e Experimentao. DINCON
9. Verhaegen M, Deprettere E (1991) A fast recursive MIMO state space model identification algorithm. In: Proceedings of the 30th conference on decision and control. Brighton, England, Dec 1991
10. Tamariz ADR (2005) Modelagem Computacional de Dados e Controle Inteligente no Espao de Estados. Tese Doutorado, LCSI/FEEC/UNICAMP. Campinas SP, Brasil
11. Aoki M (1987) State space modeling of time series. Springer Verlag
12. Tobar Johanna, Bottura Celso, Giesbrechtr Mateus (2010) Computational modeling of multivariable non-stationary time series in the state space by the AOKI_VAR algorithm. IAENG Int J Comput Sci 37(4):411–416
13. Tamariz ADR, Bottura CP, Barreto G (2005) Iterative MOESP type algorithm for discrete time variant system identification. In: Proceedings of the 13th mediterranean conference on control and automation, (MED'2005). Limassol, Cyprus. Accessed 27–29 Jun 2005

14. Verhaegen M, Dewilde P (1992) Subspace model identification—part 1: the outputterror state-space model identification class of algorithms. Int J Control 56(5):1187–1210
15. Verhaegen M, Xiaode Y (1995) A class of subspace model identification algorithms to identify periodically and arbitrarily time-varying systems. Automatica 31(2):201–216

Chapter 4
Comparative the Performance of Control Charts Based on Copulas

Sasigarn Kuvattana and Saowanit Sukparungsee

4.1 Introduction

Statistical Process Control (SPC) is an important tool to monitor and control the quality or productivity of the process. A Control chart plays a vital role in SPC and very popular instruments in several fields [1] such as manufacturing, economics and finance, health care and epidemiology, computer intrusion, environmental statistics and etc. The quality controls are often used more than a single quality characteristic then the multivariate statistical process control can overcome this problem. Multivariate Statistical Process Control (MSPC) charts regarded as the multivariate extensions of univariate charts [2]. Most of multivariate detection procedures are based on a multi-normality assumption and independence but many processes are often non-normality and correlated. Many multivariate control charts have the lack of the related joint distribution and copula can specify this property.

A copula function can be studied about dependence or association between random variables [3] tool for nonlinearity, asymmetricality and tail dependence in various areas, it is based on a representation from Sklar's theorem [4] and it can estimate joint distribution of nonlinear. There are several papers apply copula in control chart, i.e., Fatahi et al. [5, 6] proposed a copula-based bivariate zero inflated Poisson distribution to monitor rare events. Dokouhaki and Noorossana [7] studied a copula Markov Cumulative Sum chart (CUSUM) for monitoring two correlated binomial data sets with an autocorrelation. Hryniewicz [8] discussed the robustness

S. Kuvattana
Applied Statistics, Maejo University, Phrae Campus, Rong Kwang 54140
Phrae, Thailand
e-mail: sasigarn2010@gmail.com

S. Sukparungsee (✉)
Applied Statistics, King Mongkut's University of Technology North Bangkok,
Bang Sue, Bangkok 10800, Thailand
e-mail: saowanit.s@sci.kmutnb.ac.th

S.-I. Ao et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-981-10-2717-8_4

of Shewhart control charts to the case of dependent data that is described by autocorrelated and normal data. Verdier [9] presented a new approach for non-normal multivariate cases, including Cauchy, Student's t and normal distributions. Kuvattana et al. [1, 3, 10] proposed multivariate control charts for bivariate copulas, i.e., Multivariate Shewhart, Multivariate Exponentially Weighted Moving Average (MEWMA) and Multivariate Cumulative Sum (MCUSUM) control charts. In general, the performance characteristics for SPC charts is measured by the Average Run Length (*ARL*). The *ARL* is classified into two states as ARL_0 and ARL_1, which acceptable ARL_0 should be sufficient large when the process is in-control and ARL_1 should be small when the process is out-of-control.

According to literatures, times is an important part to monitor some attributes or variable measures that involved in consecutive events. The distribution of the time for an event is well-known as an exponential distribution. This paper compares Multivariate Exponentially Weighted Moving Average (MEWMA), Double Multivariate Exponentially Weighted Moving Average (DMEWMA) and Multivariate Cumulative Sum (MCUSUM) control charts when observations are exponential distribution with the mean shifts and uses multivariate copulas to specify dependence between random variables.

4.2 Research Methodology

This section explain many procedures and its overall consists of the following.

4.2.1 Multivariate Exponentially Weighted Moving Average Control Chart

Multivariate Exponentially Weighted Moving Average (MEWMA) control chart was first proposed by Lowry et al. [11] that is a logical extension of the univariate Exponentially Weighted Moving Average (EWMA) control chart. The EWMA statistic assigns progressively less weight to past observations than to the current observation [12].

Suppose $X_1, X_2, \ldots$ are observations from p-variate normal distribution $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, define recursively, for $i = 1, 2, \ldots,$

$$\mathbf{z}_i = \Lambda \mathbf{x}_i + (\mathbf{I} - \Lambda)\mathbf{z}_{i-1} \tag{4.1}$$

where $\mathbf{z}_0$ is the vector of variable values from the historical data, and Λ is a diagonal matrix with entries $\lambda_1, \ldots, \lambda_p$. The quantity to be plotted is

$$T_i^2 = \mathbf{z}_i^{/} \sum_{\mathbf{z}_i}^{-1} \mathbf{z}_i \tag{4.2}$$

where

$$\sum_i = \frac{\lambda}{2-\lambda}\left[1-(1-\lambda)^{2i}\right]\sum \tag{4.3}$$

when $\lambda_1 = \ldots = \lambda_p = \lambda \in (0,1)$, as assumed in the present paper. The control chart signals a shift in the mean vector when $T_i^2 > h$, where h is the control limit chosen to achieve a desired in-control.

The performance of the MEWMA in detecting changes in the mean is generally measured by the Average Run Length (ARL) as a function of the difference $\boldsymbol{\mu} - \boldsymbol{\mu_0}$ between the target mean $\boldsymbol{\mu_0}$ and its real value $\boldsymbol{\mu}$. Furthermore, the ARL depends on the degree of dependence between the variables, measured by the covariance matrix $\boldsymbol{\Sigma}$, and the scalar charting weight λ associated to the past observations. Note that if $\lambda = 1$ in (4.1), the MEWMA control chart statistic reduces to $T_i^2 = \mathbf{x}_i^{/} \sum_{\mathbf{Z}_i}^{-1} \mathbf{x}_i$, the statistic used in the Hotelling T^2 control chart [2].

4.2.2 Double Multivariate Exponentially Weighted Moving Average Control Chart

Double Multivariate Exponentially Weighted Moving Average (DMEWMA) control chart is extension from Double Exponentially Weighted Moving Average (DEWMA) [13]. Suppose that $\mathbf{x}_1, \mathbf{x}_2, \ldots$ are $p \times 1$ random vectors each representing p variables. The MEWMA statistic $\mathbf{y}_i$ calculated from $\mathbf{x}_i$ is written by

$$\mathbf{y}_i = \Lambda \mathbf{x}_i + (\mathbf{I} - \Lambda)\mathbf{y}_{i-1}. \tag{4.4}$$

The DMEWMA statistic $\mathbf{z}_i$ is defined as:

$$\mathbf{z}_i = \Lambda \mathbf{y}_i + (\mathbf{I} - \Lambda)\mathbf{z}_{i-1} \tag{4.5}$$

with $\Lambda = \lambda \mathbf{I}, 0 < \lambda < 1$ and $\mathbf{y}_0 = \mathbf{z}_0 = \boldsymbol{\mu}_0$. The DMEWMA control chart statistics T_{di}^2 is

$$T_{di}^2 = \mathbf{z}_i^{/} \sum_{\mathbf{Z}_i}^{-1} \mathbf{z}_i, \tag{4.6}$$

where h is the control limit. If $T_{di}^2 > h$ then the signal gives an out-of-control, otherwise, the process is in-control.

4.2.3 Multivariate Cumulative Sum Control Chart

Multivariate Cumulative Sum (MCUSUM) control chart is the multivariate extension of the univariate Cumulative Sum (CUSUM) control chart. The MCUSUM

control chart was initially proposed by Crosier [14]. The MCUSUM control chart may be expressed as follows:

$$C_t = [(\mathbf{S}_{t-1} + \mathbf{x}_t - \mathbf{a})' \sum^{-1} (\mathbf{S}_{t-1} + \mathbf{x}_t - \mathbf{a})]^{1/2} \tag{4.7}$$

where covariance matrix $\sum$ and $\mathbf{S}_t$ are the cumulative sums;

$$\mathbf{S}_t = \begin{cases} \mathbf{0}, & \text{if } C_t \le k \\ (\mathbf{S}_{t-1} + \mathbf{X}_t - \mathbf{a})\left(1 - \frac{k}{C_t}\right), & \text{if } C_t > k \end{cases} \tag{4.8}$$

the reference value $k > 0$ and $\mathbf{a}$ is the aim point or target value for the mean vector [15, 16]. The control chart statistics for MCUSUM chart is

$$Y_t = [\mathbf{S}_t' \sum^{-1} \mathbf{S}_t]^{1/2} \tag{4.9}$$

where h is the control limit, if $Y_t > h$ then the signal gives an out-of-control.

4.2.4 Copulas Concepts

For a bivariate case, if $H(x, y)$ is the joint distribution of the random vector (X, Y), with continuous marginal distributions $F(x) = H(x, \infty)$, $G(y) = H(\infty, y)$ of X and Y, respectively. Then, there are exists a unique copula $C(u, v)$ such that $H(x, y) = C(F(x), G(y); \theta)$ where $C(u, v)$ is determined by

$$C(u, v) = H(F^{-1}(u), G^{-1}(v))$$

where $F^{-1}(.)$, $G^{-1}(.)$ are quantiles functions of $F(.)$, $G(.)$, respectively, which are defined as

$$F^{-1}: [0, 1] \to \mathbb{R}$$

$$F^{-1}(u) = \inf\{x \in \mathbb{R} : F(x) \ge u\}$$

For the purposed of the statistical method, it is desirable to parameterize the copula function. Let θ denote the association parameter of the bivariate distribution, and there exists a copula C [17]. This article attends to the Normal, Clayton and Frank copulas because these copulas are well-known and very useful [18].

4.2.4.1 Normal Copula

The Normal copula is an elliptical copula and a bivariate Normal distribution with zero mean, unit variances and 2×2 correlation matrix $\mathbf{\Sigma}$, the Normal copula is defined as:

$$C(u, v; \mathbf{\Sigma}) = \Phi_N(\Phi^{-1}(u), \Phi^{-1}(v); \mathbf{\Sigma}) \tag{4.10}$$

where $\Phi_N(u, v)$ is the cumulative probability distribution function of the bivariate Normal distribution, $\Phi^{-1}(u)$ and $\Phi^{-1}(v)$ are the inverse of the cumulative probability function of the univariate Normal distribution.

4.2.4.2 Archimedean Copulas

Archimedean copulas are an associative class of copulas. Let Φ be a class of continuous strictly decreasing functions $\phi: [0, 1] \to [0, \infty]$ such that $\phi(1) = 0$, $\phi^{'}(t) < 0$ and $\phi^{''}(t) > 0$ for all $0 < t < 1$ [17–21]. Archimedean copulas generated as follows:

Clayton Copula

$$C(u, v; \theta) = \left[max(u^{-\theta} + v^{-\theta} - 1, 0)\right]^{-1/\theta}, \phi(t) = (t^{-\theta} - 1)/\theta \tag{4.11}$$

Frank Copula

$$C(u, v; \theta) = -\frac{1}{\theta} ln(1 + \frac{(e^{-\theta u} - 1)(e^{-\theta v} - 1)}{e^{-\theta} - 1}), \phi(t) = -ln(\frac{e^{-\theta t} - 1}{e^{-\theta} - 1}) \tag{4.12}$$

4.2.5 Dependence and Kendall's tau

According to the earlier literatures, the copulas can be used in the study of dependence or association between random variables and the values of Kendall's tau are easy to calculate so this measure is used for observation dependencies. Let X and Y be continuous random variables which copula is C, then Kendall's tau is given by $\tau_c = 4\int_{\mathbf{I}^2} C(u, \nu)\, dC(u, \nu) - 1$ where τ_c is Kendall's tau of copula C and the unit square $\mathbf{I}^2$ is the product $\mathbf{I} \times \mathbf{I}$ where $\mathbf{I} = [0, 1]$ and the expected value of the function $C(u, \nu)$ of uniform (0,1) random variables U and V whose joint distribution function is C, i.e., $\tau_c = 4E[C(U, V)] - 1$ [20] (Table 4.1).

Table 4.1 Kendall's tau of copulas and generation

Copula type	Function of copula	τ	θ
Normal	$C(u, v) = \Phi_N(\Phi^{-1}(u), \Phi^{-1}(v))$	$arcsin(\theta)/(\pi/2)$	$[-1, 1]$
Clayton	$C(u, v) = \left[max(u^{-\theta} + v^{-\theta} - 1, 0)\right]^{-1/\theta}$	$\theta/(\theta + 2)$	$[-1, \infty)\backslash 0$
Frank	$C(u, v) = -\frac{1}{\theta} ln(1 + \frac{(e^{-\theta u} - 1)(e^{-\theta v} - 1)}{e^{-\theta} - 1})$	$1 + 4\left(\frac{1}{\theta}\int_0^{\theta} \frac{t}{e^t - 1}\, dt - 1\right)/\theta$	$(-\infty, \infty)$

Table 4.2 Comparison of *ARL* of control charts with Kendall's tau values equal to 0.5 in the case of one exponential parameters shift

Mean shifts		*ARL*								
		DMEWMA ($\lambda = 0.05$)			MCUSUM			MEWMA ($\lambda = 0.05$)		
μ_1	μ_2	Normal	Frank	Clayton	Normal	Frank	Clayton	Normal	Frank	Clayton
1	**1**	**369.800**	**370.232**	**370.030**	**370.017**	**370.016**	**370.047**	**370.156**	**369.857**	**370.060**
1	1.25	224.068	224.369	223.780	199.409	195.186	194.026	**194.572**	**194.206**	195.067
1	1.5	141.603	142.946	144.851	111.201	109.707	108.168	**94.668**	98.999	100.708
1	1.75	101.343	102.373	103.814	65.302	66.200	65.716	70.831	63.638	**55.007**
1	2	77.341	78.577	79.499	41.357	42.433	42.626	**29.412**	31.764	32.832
1	2.25	62.742	63.791	65.401	28.160	29.353	29.525	25.344	23.445	**21.371**
1	2.5	52.923	53.561	54.620	20.479	21.439	21.687	**12.261**	13.544	14.116
1	2.75	45.212	46.159	47.247	15.468	16.237	16.639	11.432	10.935	**9.997**
1	3	40.007	40.762	41.483	12.203	12.807	13.216	**6.218**	6.921	7.264
1	**1**	**369.800**	**370.232**	**370.030**	**370.017**	**370.016**	**370.047**	**370.156**	**369.857**	**370.060**
1.25	1	223.339	222.797	224.030	198.894	195.577	193.795	**194.331**	195.262	194.924
1.5	1	141.681	143.327	143.815	109.680	110.218	108.418	**96.096**	99.059	101.038
1.75	1	100.827	101.907	103.276	65.097	66.199	65.872	70.730	63.281	**55.231**
2	1	77.807	78.469	80.310	41.243	42.851	42.613	**29.537**	31.386	32.924
2.25	1	62.526	63.798	64.582	28.014	29.511	29.429	25.308	23.209	**20.933**
2.5	1	52.173	53.601	54.496	20.317	21.580	21.641	**12.305**	13.428	14.202
2.75	1	45.480	46.040	47.079	15.527	16.421	16.606	11.388	10.859	**10.039**
3	1	39.379	40.412	41.392	12.191	12.834	13.100	**6.216**	6.936	7.399

Table 4.3 Comparison of *ARL* of control charts with Kendall's tau values equal to 0.5 in the case of two exponential parameters shifts

Mean shifts		*ARL*								
		DMEWMA ($\lambda = 0.05$)			MCUSUM			MEWMA ($\lambda = 0.05$)		
μ_1	μ_2	Normal	Frank	Clayton	Normal	Frank	Clayton	Normal	Frank	Clayton
1	**1**	**369.800**	**370.232**	**370.030**	**370.017**	**370.016**	**370.047**	**370.156**	**369.857**	**370.060**
1.25	1.25	162.779	160.870	161.016	138.760	133.198	128.598	127.813	125.496	**124.135**
1.5	1.5	94.444	93.933	93.548	66.428	64.077	61.195	**51.746**	52.322	51.781
1.75	1.75	65.617	64.926	64.900	36.480	35.862	34.465	35.592	30.371	**25.431**
2	2	50.005	49.594	49.769	22.482	22.500	21.785	**13.908**	14.535	14.588
2.25	2.25	40.7318	40.098	39.889	15.217	15.416	14.971	11.472	10.447	**8.863**
2.5	2.5	33.855	33.581	33.666	10.988	11.235	10.809	**5.336**	5.752	5.923
2.75	2.75	29.045	28.853	28.915	8.200	8.477	8.145	4.897	4.635	**4.125**
3	3	25.530	25.550	25.453	6.232	6.441	6.309	**2.479**	2.808	2.912

Table 4.4 Comparison of *ARL* of control charts with Kendall's tau values equal to 0.8 in the case of one exponential parameters shift

Mean shifts		*ARL*								
		DMEWMA ($\lambda = 0.05$)			MCUSUM			MEWMA ($\lambda = 0.05$)		
μ_1	μ_2	Normal	Frank	Clayton	Normal	Frank	Clayton	Normal	Frank	Clayton
1	**1**	**369.739**	**370.036**	**369.904**	**369.860**	**370.129**	**370.041**	**370.053**	**369.885**	**369.931**
1	1.25	209.982	210.934	215.128	180.303	195.096	195.156	**172.323**	196.494	200.217
1	1.5	119.358	123.360	129.240	85.849	105.377	107.211	**68.434**	91.958	100.640
1	1.75	79.129	83.020	88.060	44.090	59.581	62.768	**30.661**	43.806	50.501
1	2	58.588	61.836	65.551	25.383	36.230	38.588	**15.645**	23.565	28.456
1	2.25	45.604	49.055	52.858	16.397	23.634	25.877	**9.117**	14.100	17.046
1	2.5	37.833	40.597	43.896	11.377	16.566	18.506	**5.504**	8.758	10.940
1	2.75	31.735	34.513	37.517	8.351	12.187	13.778	**3.519**	5.853	7.367
1	3	27.758	29.987	32.575	6.387	9.336	10.622	**2.423**	3.870	5.122
1	**1**	**369.739**	**370.036**	**369.904**	**369.860**	**370.129**	**370.041**	**370.053**	**369.885**	**369.931**
1.25	1	208.451	208.892	215.027	179.977	196.190	195.710	**173.422**	195.705	200.712
1.5	1	119.674	123.309	129.386	86.487	105.999	108.657	**68.604**	90.928	100.840
1.75	1	79.245	83.409	88.377	44.249	60.205	63.247	**30.715**	44.335	50.995
2	1	58.691	61.923	66.249	25.372	36.091	38.869	**15.837**	23.802	28.400
2.25	1	45.832	49.051	52.578	16.320	23.573	25.919	**9.068**	13.884	17.134
2.5	1	37.919	40.632	43.832	11.420	16.441	18.501	**5.525**	8.865	10.858
2.75	1	31.890	34.646	37.293	8.335	12.148	13.708	**3.517**	5.785	7.270
3	1	27.599	30.117	32.627	6.360	9.335	10.643	**2.492**	3.955	5.076

Table 4.5 Comparison of *ARL* of control charts with Kendall's tau values equal to 0.8 in the case of two exponential parameters shifts

Mean shifts		*ARL*								
		DMEWMA ($\lambda = 0.05$)			MCUSUM			MEWMA($\lambda = 0.05$)		
μ_1	μ_2	Normal	Frank	Clayton	Normal	Frank	Clayton	Normal	Frank	Clayton
1	**1**	**369.739**	**370.036**	**369.904**	**369.860**	**370.129**	**370.041**	**370.053**	**369.885**	**369.931**
1.25	1.25	176.171	167.007	165.839	148.697	149.487	145.866	**137.850**	146.644	142.534
1.5	1.5	104.900	97.444	96.584	76.813	80.001	76.576	**59.246**	67.359	67.537
1.75	1.75	73.051	68.053	66.915	43.990	48.708	46.467	**29.924**	35.872	35.944
2	2	55.320	51.738	51.142	27.655	31.957	30.540	**16.999**	21.105	21.785
2.25	2.25	44.524	41.841	41.435	18.737	22.423	21.532	**10.372**	13.804	14.056
2.5	2.5	37.606	34.900	34.772	13.508	16.533	15.978	**6.765**	9.169	9.716
2.75	2.75	32.143	30.273	30.049	10.247	12.726	12.340	**4.566**	6.640	6.877
3	3	28.228	26.757	26.389	7.911	10.005	9.762	**3.145**	4.846	5.193

4.3 Numerical Results

A Monte Carlo simulation was conducted using the R statistical software [22, 23] to compare the *ARL* in each control charts. The number of simulation runs was 50,000 repeating and sample size is 1,000 with the empirical simulation studies. Observations were from exponential distribution with parameter (μ) equal to 1 for in-control process $(\mu_0 = 1)$ and the shifts of the process level (δ) by $\mu = \mu_0 + \delta$. The means of process shifts are equal to 1.25, 1.5, 1.75, 2, 2.25, 2.5, 2.75 and 3 for out-of-control process. Copula estimations are restricted to the cases of positive dependence. For all copula models, the setting θ corresponds to Kendall's tau and the level of dependence is measured by Kendall's tau values $(-1 \leq \tau \leq 1)$. The authors define 0.5 and 0.8 to moderate and strong dependence, respectively.

The simulation results are presented in Tables 4.2, 4.3, 4.4 and 4.5, we specify $\lambda = 0.05$ for MEWMA and DMEWMA control charts. The different values of exponential parameters denote μ_1 for the variables X and μ_2 for the variables Y. For in-control process, control charts were chosen by setting the desired $ARL_0 = 370$ for each copula and control charts. The results in Tables 4.2, 4.3, 4.4 and 4.5 show that the ARL_1 values of the MEWMA control chart are less than the other charts. In Tables 4.4 and 4.5, in the case of one and two exponential parameter shifts, the Normal copula on the MEWMA control chart are minimum when compared with other copulas. Consequently, the MEWMA control chart performs the best procedure in order to detect shifts in parameter.

4.4 Conclusion and Future Work

In the manufacturing process, dependence between random variables of multivariate control chart can be used copula concept, which is not designed under the statistical assumption that is described by independent and identically distributed random variable. The authors compare MEWMA, DMEWMA and MCUSUM control charts for bivariate copulas when observations are exponential distribution using the Monte Carlo simulation approach. The numerical results found that MEWMA performs the best control chart and the Normal copula of Multivariate Exponentially Weighted Moving Average (MEWMA) control chart can be used for almost all shifts. For future work, this work can extension more than two variables and use for the other control charts that practitioners may apply several fields.

Acknowledgments The author would like to express my gratitude to Department of Applied Statistics, Faculty of Sciences, King Mongkut's University of Technology, North Bangkok, Thailand for supporting the supercomputer.

References

1. Kuvattana S, Sukparungsee S, Busababodhin P, Areepong Y (2015) Efficiency of bivariate copulas on the CUSUM chart. In: Proceedings of the international multiconference of engineers and computer scientists 2015, IMECS 2015. Lecture notes in engineering and computer science. Hong Kong, pp 829–832. Accessed 18–20 Mar 2015
2. Montgomery DC (2013) Statistical quality control: a modern introduction, 7th edn. John Wiley & Sons, Singapore
3. Kuvattana S, Sukparungsee S, Busababodhin P, Areepong Y (2015) A Comparison of efficiency between multivariate Shewhart and multivariate CUSUM control chart for bivariate copulas. In: Proceedings of international conference on applied statistics, ICAS 2015. Statistics for global evolution vision in the 21st century. Pattaya, Thailand, pp 219–223. Accessed 15–17 Jul 2015
4. Sklar A (1973) Random variables, joint distributions, and copulas. Kybernetica 9:449–460
5. Fatahi AA, Dokouhaki P, Moghaddam BF (2011) A bivariate control chart based on copula function. In: IEEE international conference on quality and reliability (ICQR). Bangkok, Thailand, pp 292–296. Accessed 14–17 Sept 2011
6. Fatahi AA, Noorossana R, Dokouhaki P, Moghaddam BF (2012) Copula-based bivariate ZIP control chart for monitoring rare events. Commun Stat Theory Methods 41:2699–2716. doi:10.1080/03610926.2011.556296
7. Dokouhaki P, Noorossana R (2013) A copula markov CUSUM chart for monitoring the bivariate auto-correlated binary observation. Qual Reliab Eng Int 29(6):911–919. doi:10.1002/qre.1450
8. Hryniewicz O (2012) On the robustness of the Shewhart control chart to different types of dependencies in data. Front Stat Qual Control 10:20–33. doi:10.1007/978-3-7908-2846-7_2
9. Verdier G (2013) Application of copulas to multivariate control charts. J Stat Plann Infer 143:2151–2159. doi:10.1016/j.jspi.2013.05.005
10. Kuvattana S, Sukparungsee S, Busababodhin P, Areepong Y (2015) Performance comparison of bivariate copulas on the CUSUM and EWMA control charts. In: Proceedings of the world congress on engineers and computer science 2015, WCECS 2015. Lecture notes in engineering and computer science. San Francisco, USA, pp 853–857. Accessed 21–23 Oct 2015
11. Lowry CA, Woodall WH, Champ CW, Rigdon SE (1992) A multivariate exponentially weighted moving average control chart. Technometrics 34(1):46–53. doi:10.1080/00401706.1992.10485232
12. Midi H, Shabbak A (2011) Robust multivariate control charts to detect small shifts in mean. Math Probl Eng 2011:1–19. doi:10.1155/2011/923463
13. Shamma SE, Shamma AK (1992) Development and evaluation of control charts using double exponentially weighted moving averages. Int J Qual Reliab Manage 9(6):18–25. doi:10.1108/02656719210018570
14. Crosier RB (1988) Multivariate generalizations of cumulative sum quality control schemes. Technometrics 30(3):291–303. doi:10.1080/00401706.1988.10488402
15. Khoo BC, Atta MA, Phua HN (2009) A Study on the performances of MEWMA and MCUSUM charts for skewed distributions. In: Proceedings of 10th Islamic countries conference on statistical science. Cairo, Egypt, pp 817–822. Accessed 20–23 Dec 2009
16. Alves CC, Samohyl RW, Henning E (2010) Application of multivariate cumulative sum control charts (MCUSUM) for monitoring a machining process. In: 16th international conference on industrial engineering and operations management, 2009. Sao Carlos, Brazil, pp 1–7. Accessed 12–15 Oct 2010
17. Trivedi PK, Zimmer DM (2005) Copula Modeling: an introduction for practitioners. Foundations and trends in econometrics
18. Genest C, McKay RJ (1986) The joy of copulas: bivariate distributions with uniform marginals. Am Stat 40(4):280–283. doi:10.1080/00031305.1986.10475414

19. Jaworski P, Durante F, Härdle W, Rychlik T (2009) Copula theory and its applications. In: Proceedings of the Workshop Held in Warsaw. Lecture notes in statistics. Springer-Verlag, Berlin, Heidelberg. Warsaw, Poland. Accessed 25–26 Sept 2009
20. Nelsen RB (2006) An introduction to copulas, 2nd edn. Springer, New York
21. Genest C, Rivest L-P (1993) Statistical inference procedures for bivariate Archimedean copulas. J Am Stat Assoc 88(423):1034–1043. doi:10.1080/01621459.1993.10476372
22. Yan J (2007) Enjoy the joy of copulas: with a package copula. J Stat Softw 21(4):1–21. doi:10.18637/jss.v021.i04
23. Hofert M, Mächler M, McNeil AJ (2012) Likelihood inference for Archimedean copulas in high dimensions under known margins. J Multivar Anal 110:133–150. doi:10.1016/j.jmva.2012.02.019

Chapter 5
Mamdani Fuzzy Decision Model for GIS-Based Landslide Hazard Mapping

Monalee A. dela Cerna and Elmer A. Maravillas

5.1 Introduction

Natural disasters can be prevented by mitigating the destruction using information and technologies that are readily available. With the advent of technology, software is invented to map risks especially landslide risk associated with improper environmental practices. Small-scale Mining (SSM) activities are less checked or monitored, and no existing profiling and mapping of the occurrence are made. The continued investigation into the occurrence and implication of geohazards, about the mining exploration, has led to particular mitigation studies and establishing relevant prevention mechanisms.

Landslides have an important role in the evolution of landscapes but also signify serious threat in many parts of the world. Recent studies show that at the onset of 20th century many casualties were reported due to rainstorms, massive floods, landslides, and earthquakes [1]. Landslides are triggered by natural causes or by human activity. One of the fundamental causes of landslides is mining [2]. As reported by IRINnews.org [3], SSM is one of the primary triggers of the landslide in the Philippines a country highly vulnerable to natural disaster. In recent years, there have been spates of mining-related accidents in Mindanao, particularly in Davao region. In 2011 to 2012, 1,124 deaths, 162 missing and 23 injured happened due to landslide brought by continuous heavy rains [2].

M.A. dela Cerna (✉)
IAENG, Faculty in the Information Technology Division, Surigao State College of Technology, 8400 Surigao City, Philippines
e-mail: madc712004@yahoo.com

E.A. Maravillas
Chair, Graduate School in the College of Computer Studies, Cebu Institute of Technology University, 6000 Cebu City, Philippines
e-mail: elmer.maravillas@gmail.com

S.-I. Ao et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-981-10-2717-8_5

Moreover, similar incident happened in Surigao del Norte. An unreported incident of fatalities caused by a landslide due to SSM unregulated activities occurred. Furthermore, concerns were raised by the Mines and Geosciences Bureau (MGB), Surigao Metropolitan Water District (SMWD) and the local government offices regarding persistent illegal small-scale gold mining activities in landslide-prone areas of Surigao and protected watershed areas where large deposits of gold can be found. Despite the known hazards of SSM especially due to illegal gold panning, people still can secure permits from higher authorities.

Although SSM is a leading environmental and social problem in many countries, it still gained global importance. Reasons are due to being a source of subsistence and people's means of livelihood. Poverty and limited employment opportunities in urban areas make SSM activities thrive.

Researchers conducted a comprehensive field survey in areas where minerals are currently mined. It is observed during field reconnaissance; there are different mining methods applied in the exploration of gold. Drilling, deep tunneling (sinking and drifting) in the farmland and mountainous areas, and *banlas or* open cut minings in the mountains have been documented. Miners use hydraulic mining, panning and amalgamation operations. Hydraulic mining is locally known as "*banlas*" is surface mining with the use of hydraulic hose and sluice boxes. It is operated by three or more persons in an open area positive for gold. This practice is done next to creeks or rivers to drain quickly the water that washes the ores with gold. This is prohibited because it destroys vegetation, causes soil erosion, exposes tree roots, silts and pollutes rivers through the use of mercury. Furthermore, this operation comes with the illegal cutting of trees for slabs in assembling sluice boxes. Inherent digging of soil, sluicing, water siltation, erosion, and land degradation with rivers used for mining are issues need to be addressed.

In the past few years, the application of Geographic Information System (GIS) has increased [4, 5, 6, 7]. It is an excellent alternative mapping technique for monitoring, investigating, and assessing geo-hazards areas. More researchers are utilizing GIS and applied probabilistic models in forecasting landslide hazard zones. The models such as logistic regression [8], fuzzy logic [5, 9, 6, 4, 10, 11, 12] and artificial neural network [8] are used. Fuzzy Logic (FL) has a significant impact on intelligent autonomous systems design and is an essential problem-solving method. The fuzzy system has become popular for control, expert systems, prediction and decision making. FL method offers more flexible combinations of weighted maps, and readily implements GIS modeling language [6].

The current situation of the community and problems encountered by some government officials imposing the responsible mining in CARAGA Region have motivated the researchers to conduct a study and develop a fuzzy logic controller to assess landslide hazard in mining areas particularly in Surigao del Norte province. They integrated the use of GIS application in mapping for efficient data warehousing. This will aid the local government in their planning the disaster prevention programs.

5.2 Theoretical Framework

Beck 'risk society' theory deals with the intellectual mapping of endemic catastrophic risks [13]. It is society's calls for concerted efforts from various disciplines to mitigate possible effects of risks in vulnerable sectors of the society which closely linked to poverty. This study is anchored on the Beck's theory since SSM activities involve people in the marginalize areas who are forced to work in this precarious livelihood due to poverty. In many cases, small-scale miners have no safety equipment and use hand tools only [14]. There is a necessity to employ security measures to minimize the risk of their work.

Figure 5.1 presents the framework of the study. Data Sources, Geodatabase, Fuzzy Logic System, Landslide Hazard Mapping, and Reports are terminators that need to be communicated to the system to ensure efficiency in the entire process.

The first step includes data collection. Landslide assessment is then established. In parallel, the landslide causative factors are determined. Factors are then statistically tested, and the active factors are selected for hazard analysis. Data sources cover the range of operations by which spatial data from a global positioning system and other sources are transformed into an ordinal format. The attributes associate a numerical code to each cell or set of coordinates and for each variable, either to represent actual values or to connote specific data types.

Geodatabase refers to the way in which spatial data are structured and organized according to their location, interrelationship, and attribute design.

The procedure using a fuzzy logic system for analysis is performed to obtain useful information from data previously entered into the system. Data manipulation embraces causative factors utilized as crisp inputs to the controllers; the Mamdani Inference System is used for fuzzification and Center of Gravity for defuzzification method.

The data output is the GIS map. Two types of data entered into GIS are geographic references and attributes. The geographic reference data are the coordinates,

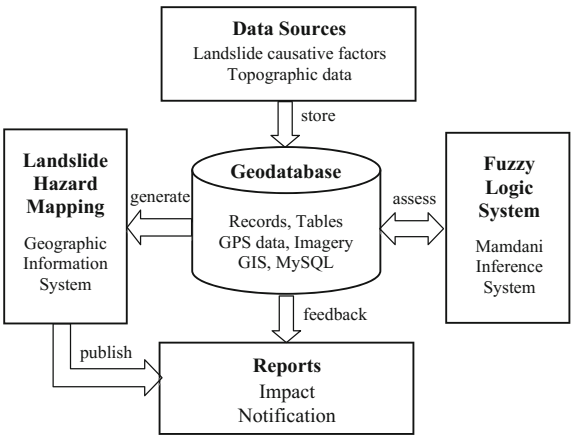


Fig. 5.1 Framework of the study

which are latitude and longitude or the columns and rows, which give a result of the location based on the entered information.

Reports include the impact of mining-related incidents and notification. Linkages between MGB and SSM people to establish capability and security dimensions on the situation of the miners. It also reduce the risks through planning and implementation of mitigating factors, and establish policy for responsible mining.

5.3 Methods

This study employed the Rapid Application Development approach that creates an increasingly functional version of the system in short cycles. Each version is reviewed to produce requirement that will feed the next phase. This process is repeated until all functionality is developed. Each cycle provides the user an opportunity to have feedback, refine requirements, and view progress. The fuzzy logic algorithm Mamdani method employed determines landslide hazard rate. This study also uses the integration of GIS for mapping landslide hazard zones. The outcome produced a landslide hazard mapping in SSM areas in Surigao del Norte.

5.3.1 Study Area

Surigao del Norte is one of the provinces of the CARAGA region and is located in Mindanao in the Northeastern part between 125º15' to 126º15' east longitude and 9º18' to 10º30' north latitude. Its boundaries are the Pacific Ocean on the northern and eastern sides, the provinces of Agusan del Norte and Surigao del Sur respectively in the southern part, while in the western portion is the Surigao Strait. The area is composed of 20 municipalities, one city, and 395 barangays.

The abundant mineral reserves of the province include gold, manganese, silica, cobalt, copper, chromite, limestone, and silver. Also, nickel is the largest deposit found in the island of Nonoc, and the world's biggest iron deposits are found in Claver, Surigao del Norte. Diverse species of fish and aquatic life are abundat within its territorial sea boundaries. The province is under the tropical climate type and prone to brief afternoon downpours and thunderstorms. The soil in the area is clay and sandy loam type. In the mainland area, it is classified as loam soil characterized as permeable, moderately drained and highly suitable for agriculture [15]. Most of the active mining areas are located in moderate to steep rugged terrain underlain by volcanic rocks and terrace gravel deposits and are structurally controlled along Philippine fault line. Figure 5.2 shows the map of Surigao del Norte.

Fig. 5.2 Map of Surigao del Norte province

5.3.2 *Risk Indicators of Landslides*

Published literature about the indicators of landslide risk areas reveals eight causative factors of landslides [6, 7] such as slope gradient, vertical displacement, drainage density, the rate of weathering, lithology, ground stability, soil type and vegetation. Field surveys conducted from September 2013 to February 2014 and December 2014 to January 2015 respectively, for gathering baseline data for landslide inventory reveal the presence of the enumerated eight causative factors of landslides in areas under study. They are naturally occurring. In the research site, one or two factors may be present. However, unregulated SSM activities can make mining sites susceptible to landslides or cause the presence of the eight causative factors of landslides. Table 5.1 depicts the rate of landslide hazard parameters with their corresponding linguistic definitions [5]. Further, the researchers conducted a geological investigation to prove areas of SSM. The field surveys included assessing and photographing the sites, verifying their location by taking GPS points. Interviews were conducted with some of the present owners of tunnels and mill sites. Figure 5.3 shows some landslide photos in the SSM areas of the province.

5.3.3 *Fuzzy Decision Model for Landslide Hazard Assessment*

The fuzzy logic system theory was introduced in 1965 by Professor Lotfi A. Zadeh of University of California in Berkeley is utilized in this study to assess landslide

Table 5.1 Rate of landslide hazard parameters

Variable		Classes	Rate
Slope gradient	FU	• Flat and undulating (0°–<8°)	1
	MS	• Moderately sloping (8°–<18°)	3
	HS	• Hilly and Moderately steep (18°–<26°)	6
	SP	• Steep (26°–<=90°)	9
Vertical displacement	SH	• Shallow (<2 m)	1
	MV	• Moderate (2–9 m)	3
	DP	• Deep (10–25 m)	6
	VD	• Very deep (>25 m)	9
Drainage density	LD	• Low (0–1 drainage)	1
	MD	• Moderate (2–4 drainage)	3
	HD	• High (5–10 drainage)	6
	DD	• Very High (>10 drainages)	9
Weathering	SW	• Slightly weathered (stain along joints)	1
	MW	• Moderately weathered (less than half of decomposed rock material)	3
	HW	• Highly weathered (discolored rock throughout; more than half of decomposed rock material)	6
	CW	• Completely weathered	9
Lithology	WF	• Weakly fractured/weathered (andesite porphyry; diorite; agglomerate; hydrothermal breccia)	1
	MF		3
	HF	• Moderately fractured/weathered (andesite porphyry; diorite; agglomerate; siltstone; hydrothermal breccia)	6
	VF	• Alluvium; highly weathered/fractured and delighting structure (siltstone; andesite porphyry; ultramafic; agglomerate; hydrothermal breccia)	9
		• Highly weathered/Fractured Sandstone; terrace gravel; highly fractured and highly weathered (siltstone; andesite porphyry; ultramafic; agglomerate; diorite; hydrothermal breccia) with tension cracks/or landslide scarp	
Ground stability	ST	• Stable with no identified landslide scars, it is either old, recent or active	1
	SC		3
	IL	• Soil creep and other indications for possible landslide occurrence are present	6
	AL	• Inactive landslides evident; tension cracks present	9
		• Active landslides are evident with tension cracks, bulges, terracettes, seepage present	
Soil type	SG	• Silts and clays, liquid limit 50 % or greater (50 % or more of material is smaller than no. 200 sieve size.) [18]	1
	SL		3
	SD	• Silts and clays, liquid limit less than 50 % (50 % or more of material are smaller than no. 200 sieve size.) [18]	6
	GV	• Sands, more than 50 % of coarse fraction smaller than no. 4 sieve size (more than 50 % of the material is larger than no. 200 sieve size) [18].	9
		• Gravels, 50 % more than of coarse fraction larger than no. 4 sieve size. (more than 50 % of the material is larger than no. 200 sieve size) [18]	

(continued)

Table 5.1 (continued)

Vegetation	PG	• Primary growth (untouched, pristine forest that exists in its original condition/characterized by a full ceiling canopy and usually several layers of understory) • Secondary growth (rainforest that has been disturbed in some way, naturally or unnaturally/degraded forest recovering from selective logging to areas cleared by slash-and-burn agriculture that has been reclaimed by forest.) • Tertiary growth (vegetation permanently cultivated land areas such as coconut trees, bananas, rice, corn, coffee, palm oil, rubber tree, tobacco, shrubs/cogon grasses and fruit trees.)	1
	SG		5
	TG		9

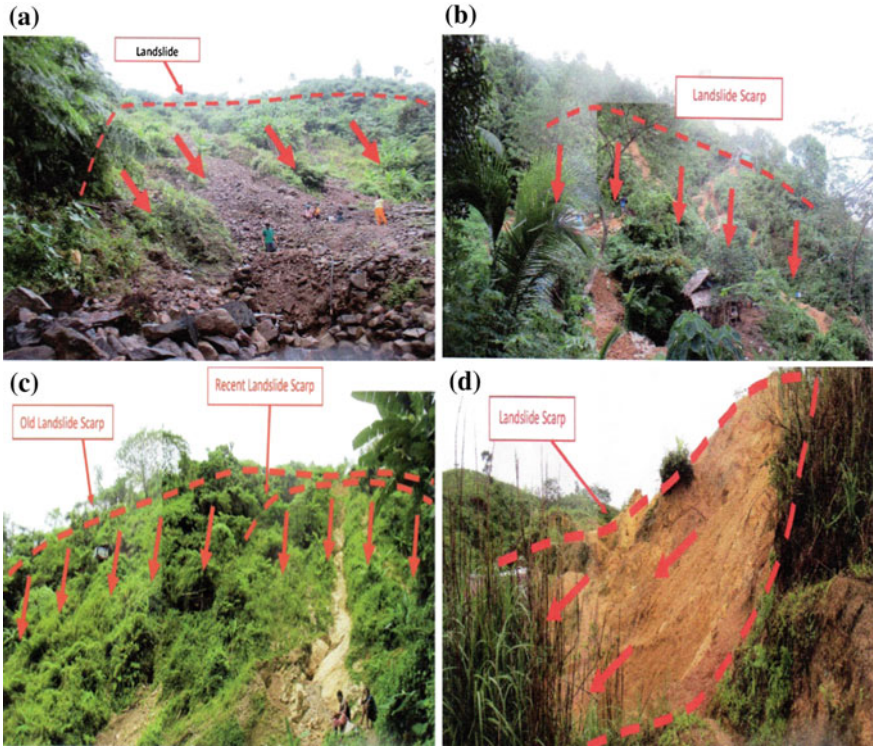


Fig. 5.3 Panoramic view of some landslide in small-scale mining areas **a** San Pedro, Alegria; **b** Cansayong, Malimono; **c** Masgad, Malimono; **d** Nabago, Surigao City

hazard rate of SSM areas. This theory is about the non-linear mapping of input data set to a scalar output data.

The eight factors that affect are the slope gradient, vertical displacement, drainage density, the rate of weathering, lithology, ground stability, soil type and vegetation used as crisp inputs to feed into the controllers. The crisp input values

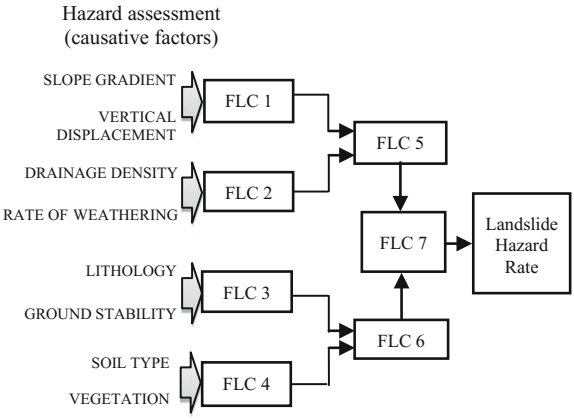


Fig. 5.4 Framework of multiple controllers using Mamdani fuzzy model

were converted to fuzzy values through input membership function. The two-input single output Mamdani fuzzy model was used in the study. Fuzzy Logic Controllers (FLC) 1, 2, 3, and 4 were the primary controllers to fuzzify. Defuzzification with Center of Gravity Method follows it. Input for controllers 5 and 6 were from the crisp output of the four controllers. The output of the last to FLCs was once again used as crisp input for controller 7. The crisp output of FLC7 is the rate of landslide hazard within the area. Figure 5.4 shows the framework for landslide hazard assessment using the Mamdani fuzzy logic controller.

The phase of investigation on the site was the first to be looked into in collecting necessary information needed for the analysis. This information was numbered "height of the slope" for example, or could be in words "degree of rock weathering". The fuzzy logic for the quantification of site's parameters was extensively applied. The second is the reasoning phase, where, with the help of a geologic expert, rules were set out to be used for reasoning over the different classes of various parameters. A reasoning rule could be: "If the slope is steep and the vertical displacement is deep, then the landslide hazard is very high".

Fuzzy inference technique or the so-called Mamdani method was used in the process and performed in four phases: (1) Fuzzification of the input variables, (2) Rule evaluation (inference), (3) Aggregation of the output rules (composition), and (4) Defuzzification.

5.3.3.1 Fuzzification of the Input Variables

The fuzzy set has graphically represented its membership functions. The universe of discourse is represented by the x-axis, while the y-axis shows degrees of membership at 0, 1 interval. Fuzzy sets are represented through triangular fuzzy numbers as well as the consequent computed firing level as the product of antecedents firing levels. The three parameters (a, b, c) show triangular function. They are defined as follows: *a* is a lower bound, *c* is an upper bound, and a value *b*, where a < b < c.

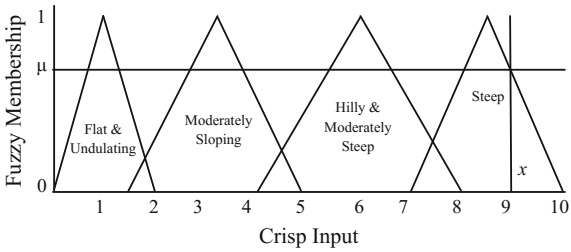


Fig. 5.5 Fuzzy membership functions for slope gradient

Refer to (5.1) for the computation of the degree of membership μ, with its crisp input x.

$$\mu \mathrm{A}(x) = \begin{cases} 0, & x \le a. \\ \frac{x-a}{b-a}, & a \le x \le b. \\ \frac{c-x}{c-b}, & b \le x \le c. \\ 0, & c \le x. \end{cases} \tag{5.1}$$

Slope gradient, vertical displacement, drainage density, the rate of weathering, lithology, ground stability and soil type use same crisp input ranges which start from 0–2, 1.5–5, 4–8, and 7–10. While, there are only three hedges from 0–3, 2–7 and 6–10 for vegetation. Figure 5.5 illustrates sample fuzzification method.

5.3.3.2 Rule Evaluation

The operation of the FLC in performing linguistic computations and its repository of knowledge of the system is described in the fuzzy rule base. IF-Then rules are obtained from an expert, who infers the rules from the system's behavior. Heuristics guidelines determine the matrix with the following statements and their converses:

(1) When the Slope Gradient is Steep, and the Vertical Displacement is Deep, then Landslide Hazard Rating is Very High.

or IF **SG** IS **SP** AND **VD** IS **DP** THEN SET **FLC1** TO **VH**
(**SG** ∧ **VD** → **VH**)

(2) When Slope Gradient is Flat and Undulating, and Vertical Displacement is Shallow, and then Landslide Hazard Rating is Low.

or IF **SG** IS **FU** AND **VD** IS **SH** THEN SET **FLC1** TO **L**
(**SG** ∧ **VD** → **L**)

Table 5.2 Fuzzy rule base for fuzzy logic controller 1

	Vertical displacement				
Slope gradient		SH	MD	DP	VD
	FU	L	M	H	H
	MS	M	H	VH	VH
	HS	H	VH	VH	VH
	SP	H	VH	VH	VH

(3) And so on…

The antecedents and consequent are represented by the matrix or rule base. In determining landslide hazard, hedges used for FLCs were (**L**) for Low Hazard; (**M**) which means Moderately Hazardous; (**H**) for Highly Hazardous; and (**VH**) for Very Highly Hazardous. Table 5.2 represents the sample of the fuzzy rule base for fuzzy logic controller 1.

5.3.3.3 Aggregation of the Output Rules

Computation of the antecedent degrees of membership is done with the use of fuzzy operator AND T-norm. The equation used for triangular membership function of intersection is defined in minimum criteria as shown below.

$$\mu A \cap B = \min(\mu \mathrm{A}, \mu B) \tag{5.2}$$

where $\mu A \cap B$ is the calculated fuzzy membership function, μA is the membership value for map A at a particular location and μB is the value for map B, and so on.

5.3.3.4 Defuzzification

In this method, the crisp outputs were calculated through the use of the center of gravity method (COG). This involves weighing and combining some fuzzy sets which resulted from the calculation of fuzzy inference process, which yield a crisp value for each output.

$$COG = \frac{\int_a^b A(x)x dx}{\int_a^b A(x) dx} \tag{5.3}$$

5.3.4 *Geographic Information System for Landslide Hazard Mapping*

Geo-technologies elevated the importance of geography to the unprecedented level in the discipline's history [16]. GIS lead a transformation in the meaning of technology. It is a powerful software technology that allows a virtually unlimited amount of information to link to a geographic location. It is also comprehensive that is specifically designed to compile, process, analyze, display, and archive large volumes of data. Crucial to the success of mining operation is a well-designed, implemented, and supported GIS management structure [17].

On the base map, layers of information were overlaid namely the locations of potential landslide sites, drifting and sinking sites, ball mill, open-pit mines, drainage canal, tension cracks, and creeks [5]. This phase was used for mapping procedure such as data acquisition, georeferencing and digitizing.

5.4 Results and Discussion

This study presented the running output of the landslide hazard fuzzy logic controller. The controller accepts data, and an image of a particular area is displayed (Fig. 5.6). The unique strength of the FLC is the overlapping of membership functions. Certain robustness is happening to the controller which for every observation there is at least one rule which is ultimately fired. Landslide hazard rate is used as feed data to GIS to produce a geo-hazard map.

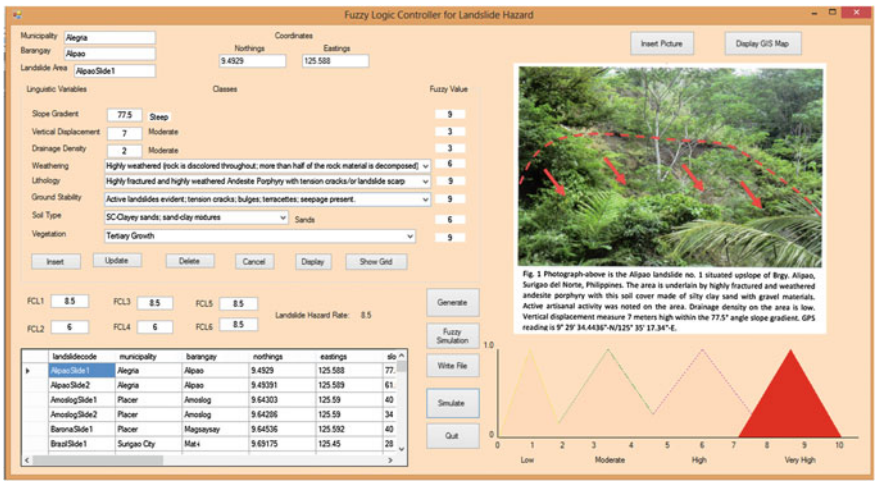


Fig. 5.6 IDE of the landslide hazard controller

Table 5.3 Landslide results from FLCs

Hazard zones	FLC1	FLC2	FLC3	FLC4	FLC5	FLC6	FLC7
Low	–	8.08	–	18.18	–	–	–
Moderate	–	70.71	–	54.55	–	–	–
High	4.04	19.19	–	18.18	8.08	–	–
Very high	95.96	2.02	100	9.09	91.92	100	100

Landslide hazard analysis result is verified using known landslide locations. There are 31 classes for eight causative factors. The factors are the slope gradient, vertical displacement, drainage density, the rate of weathering, lithology, ground stability, soil type and vegetation. Based on results show 16 barangays are within the six towns of Surigao del Norte province which are involved in the small-scale mining activities. Ninety-nine landslide areas are observed during the comprehensive survey.

Based on the result of FLC7 (Table 5.3) the 16 barangays within the six towns of Surigao del Norte are all very highly prone to landslide. The figure shows that 100 % of the study area is very highly prone to landslide with the eight causative factors taken as variables for calculating landslide hazard indicators. The landslide rate explains how the model and factors predict the landslides. The rate illustrates the fact that more percentage of landslides occur in very highly susceptible zone compared to other areas.

As depicted in the same table, very high susceptible zones obtained 95.96 % in FLC1. They consist of a steep slope and moderate vertical displacement. FLC2, which is moderate hazard areas, has obtained 70.71 % which means the areas are low in drainage density and high rate of weathering. FLC3 obtained 100 % prediction for very high hazard areas are caused by lithology and ground stability variables. Highly fractures and highly weathered classes with tension cracks or landslide scarp pose lithology hazard. When ground stability is affected by tension cracks, bulges, terracettes, and seepage, it would likely result to an active landslide. FLC4 has only obtained 54.55 % which means it is moderately a hazard zone since most of its areas are found in tertiary growth vegetation.

Findings asserted that mining areas are very highly hazardous to a landslide. Results from the survey revealed that SSM impact will cause soil erosion and landslides and which will also lead to forest destruction, agricultural lowlands, sedimentation or siltation of the water bodies. Moreover, poor mining practices also result to flash floods and flooding.

Figure 5.7 presents the identified small-scale mining areas and landslide hazard map in scale 1:10,000.

Fig. 5.7 Landslide hazard map in GIS

5.5 Conclusion

FLC application for landslide hazard forecast yields an accurate result even with overlapping. The model can map the areas and can be used as a tool to visualize the identified causative factors of the landslide. System scalability accurately calculates the rate of landslide hazard. The result reveals the system's ability to predict the emergence of danger to a landslide corollary to the heuristic prediction of MGB. The study further shows how geographic information system successfully maps landslide hazard zones. Hence, it is beneficial for local government and non-government entities in their strategic planning as ultimate guide in implementing some measures to mitigate inevitable landslide risk.

5.6 Recommendation

It is recommended that small-scale miners must be taught how to mitigate effects of landslides. Further, an intensive information drive on policy/guideline for responsible mining must be conducted.

Acknowledgements The authors would like to thank Dr. Maria Tavita Q. Lumintac and Professor Jonas Robert L. Miranda for their brilliant pieces of advice on reviewing this study and also to Professor Iryn E. Cavite for the editing she has done in this book.

References

1. Smith K (2013) Environmental hazards, assessing risk and reducing disaster, 6th edn.
2. Petly D (2011) Mining-related landslides in the Philippines. http://blogs.agu.org/landslideblog/2011/04/28/mining-related-landslides-in-the-philippines/
3. IRINnews.org (2012) PHILIPPINES: artisanal mining compounds landslide risk. http://www.irinnews.org/report/94629/philippines-artisanal-mining-compounds-landslide-risk
4. Ghosh JK, Bhattacharya D, Sharma SK (2012) Fuzzy knowledge based GIS for zonation of landslide susceptibility. In: Applications of chaos and nonlinear dynamics in science and engineering. Understanding complex systems, vol 2, © Springer-Verlag, Berlin, Heidelberg. doi:10.1007/978-3-642-29329-0_2
5. dela Cerna MA, Maravillas EA (2015) Landslide hazard GIS-based mapping using Mamdani fuzzy logic in small scale mining areas of Surigao del Norte, Philippines. In: Proceedings of the world congress on engineering and computer science 2015, WCECS 2015. Lecture notes in engineering and computer science. San Francisco, USA, pp 901–906. Accessed 21–23 Oct 2015
6. Kayastha P, Bijukchhen SM, Dhital MR, De Smedt F (2013) GIS based landslide susceptibility mapping using a fuzzy logic approach: a case study from Ghurmi-Dhad Khola Area, Eastern Nepal. J Geol Soc India 82:249–261
7. Yu G, Zeng Q, Yang S, Hu L, Lin X, Che Y, Zheng Y (2010) On the intensity and type transition of land use at the basin scale using RS/GIS: a case study of the Hanjiang river basin. Environ Monit and Assess 160(1–4):169–179
8. Yilmaz I (2010) Comparison of landslide susceptibility mapping methodologies for Koyulhisar, Turkey: conditional probability, logistic regression, artificial neural networks, and support vector machine. Environ Earth Sci 61:821–836. doi:10.1007/s12665-009-0394-9
9. Sdao F, Lioi DS, Pascale S, Caniani D, Mancini IM (2013) Landslide susceptibility assessment by using a neuro-fuzzy model: a case study in the Rupestrian heritage rich area of Matera. Nat Hazards Earth Syst Sci 13:395–407
10. Pourghasemi HR, Pradhan B, Gokceoglu C (2012) Application of fuzzy logic and analytical hierarchy process (AHP) to landslide susceptibility mapping at Haraz watershed, Iran. Nat Hazards 63(2):965–996. http://link.springer.com/article/. doi:10.1007/s11069-012-0217-2
11. Kayastha P (2012) Application of fuzzy logic approach for landslide susceptibility mapping in Garuwa sub-basin, East Nepal. Front Earth Sci 6(4):420–432
12. Dadios EP (2012) Fuzzy logic-algorithms, techniques, and implementations. http://www.intechopen.com/books/fuzzy-logic-algorithms-techniques-and-implementations
13. Beck U (2009) Critical theory of world risk society: a cosmopolitan vision. Constellations vol 16, No 1, The Author, Journal compilation. Blackwell Publishing Ltd.
14. Portal Magazine (2013) Your insight into the Philippines. First Quarter 2013, vol 2, No. 1
15. Department of tourism (2009) http://en.wikipedia.org/wiki/Surigao

16. Guo D, Mennis J (2009) Spatial data mining and geographic knowledge discovery-an introduction. Comput Environ Urban Syst 33:403–408
17. Phifer M (2012) GIS in mining. InfoMine Inc. http://technology.infomine.com/reviews/GIS/welcome.asp?view=full
18. Unified soil classification system. http://gozips.uakron.edu/~mcbelch/documents/UnifiedSoilClassification.pdf

Chapter 6
Laser Scanning as a Tool for the Analysis of Historical Buildings

Jerzy Szolomicki

6.1 Introduction

One of the many possible applications of reverse engineering, dealing with the process of transferring real structures to the virtual world, is scanning real objects to obtain geometric data to create their graphical computer models. Possession of such information for historical buildings, which often have a complex structure, is extremely important in the context of the implementation of numerical static-strength analysis and the design of any reinforcements [1].

The name reverse engineering is related to the methodology which creates the engineering design. In this methodology a real object is first is examined and then a virtual computer model is obtained by using 3D laser scanners. So it is a reverse process compared with the traditional design. Reverse engineering which uses laser scanning technology is an important element in the structure of computer aided design (CAD). 3D laser scanning is used in the digitization of historical buildings with an aim to receive detailed geometric data. The resulting point cloud consisting of millions of points can then be used to create a three-dimensional image of the scanned structure. The point cloud can be converted into a geometric grid of triangles or a set of parametric surfaces by using appropriate software [2]. Then with the use of CAD programs, the correct numerical model for the Finite Element Method (FEM) is obtained [3, 4]. The article aims to investigate the process of analysis of historical buildings with the use of 3D laser scanning and also the processing of the obtained virtual image in the FEM numerical model.

J. Szolomicki (✉)
St. Wyspianskiego 27 Street, 50-370 Wroclaw, Poland
e-mail: jerzy.szolomicki@pwr.edu.pl

S.-I. Ao et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-981-10-2717-8_6

6.2 2D Documentation

The aim of this article is to demonstrate the importance of realistic geometric modeling of historical buildings for the assessment of safety state and also to discuss the methods of data collection used in the preservation of historical structures, with particular attention devoted to the technology of laser scanners. During the period of analog studies the basic forms of documentation were plans, sections and elevations as line drawings, and also photo-plans performed on light-sensitive photographic materials. After the introduction of computer processing of digital images, forms of 2D documentation remain the same, but have changed in character. Line drawings are created in the computer memory as vector drawings, and can be easily printed in the form of CAD files. The digital form of photo-plans, due to their enormous content of information (natural colors and detail) and relatively high ease of obtaining of color print work, have become more desirable documentary products. Among 2D documentation, the following types can be distinguished:

- **Vector Documentation—plans, sections and elevations**

Digital Autograph VSD is used for studies concerning on the inventory of historical buildings. It is mainly used for vectors studies, draw plans, sections, and views of elevation. VSD is also used to obtain metrics data for creation process of raster studies.

- **Raster Documentation—digital photo-plans**

Digital photo-plans are usually performed for the presentation of views of a façade and walls, especially when it is necessary to present their brick and stone bonds. The basic geometric parameter of digital photo-plans is the size of object pixel, crucial for the development of detail On the basis of research, it is assumed that the pixel size of an object in the case of architectural buildings, in relation to their size, should be in the range of 1-3 mm. The surfaces of historical buildings that are presented in photo-plans can be divided into:

- flat or almost flat (e.g. wall surfaces),
- non-planar, regular and developable (e.g. cylindrical vault surfaces),
- non-planar, regular but non-developable (e.g. spherical dome surfaces),
- non-planar, irregular (e.g. damaged wall surfaces).

Depending on the type of surface, various methods of creating photo-plans are applied.

A geometrical survey of the structure is important for both an understanding of the current condition of the structure, including any damage, and also the destruction of the material and the structural modeling for structural analysis. There are various inventory techniques that allow the collection of accurate data a geometric description of the building. Tacheometers are traditionally used for this purpose. They are useful for surveys of edge of important points, however, it is

difficult to study complex surfaces using this method. Other techniques include close-range photogrammetry and laser scanners are used, which are faster and more accurate, and also allow for extensive surveys and tests.

6.3 Types of 3D Scanners

Laser scanners provide realistic data about objects with very accurate measurements of thousands of points made in a relatively short period of time. In the process of transformation, collected data can be used to develop two- and three dimensional models that are useful in many applications. Laser scanners can be based on the principle of triangulation, on a return time of the reflected signal or the phase difference. Triangulation lasers are devices that transfer a laser line or pattern on the object and measure the deformation of the pattern and thus determine the geometry of the object. Scanners operating on a return time of the reflected signal are based on measuring the time between sending a short laser signal, and receiving its reflection from the object. These two types of scanners are relatively slow, but can measure points at a distance of 1 km without losing on accuracy. Phase scanners use modulated continuous laser beam instead of short signals, which enables rapid measurement, but the range for such scanners does not exceed 50–80 m. Laser scanners have been available for more than a decade, but the last three years have seen a big development in their applications and quality. Development of software for effective, efficient handling and visualization of point clouds has ushered in a new era in the field of measurement, called "high definition survey"—HDS. With these technologies the actual condition can be documented in an efficient, accurate and reliable way. Inaccurate or even erroneous data about the state of the actual object were, and still are, a major problem for engineers. The latest laser scanners available today can create massive data sets in a very short period of time. Completely new approaches and concepts have been developed to manage these large data sets. HDS supports the generation of results and data using differentiated solutions that combine and integrate various scanning technologies and software modules.

6.4 Possibilities of Using Data Obtained from Measurements Using Laser Scanners

As a result of processing terrestrial laser scanning data, the following final products can be obtained:

- Filtered and oriented point clouds (e.g. in ASCII format) usually used as the starting material for further studies; they contain basic information about object;

- Orthoplans of objects generated from the point cloud—are developing on the grayscale showing the strength of the returning signal from the object through replacement of individual points from the acquired clouds on the pixels of a given terrain size. In this way a product analogous to classical orthophotoplan is obtained;
- Vector models, used primarily for the analysis of the differences between the designed and actual surface; they are mainly used for the control of performed elements, but also in the measurement of displacements and strains;
- Line developments: plans, sections, elevations, which are so far the most frequently ordered by customer;
- Contour models of scanned objects;
- Hypsometric models of scanned objects;
- 3D models of objects.

The last three groups of products are usually used for visualization purposes of certain object characteristics.

6.5 Application of 3D Laser Scanners to Determine the Geometry of Historical Buildings

The principle of laser scanner operation is based on illumination of the investigated object with a laser beam, which when encountering this object is reflected and recorded by the device. An effect of line distortion in the form of a light beam illuminating the surface of the object, the so-called Moire pattern [5] is utilized. The measured object is illuminated by a raster with known parameters (set of lines with known density). Control software allows the distance of the point from the scanner on the basis of the phase shift between the input and output light beam to be determined. The distance and information on the current position of the mirror and the angle of rotation of the scanner allow the coordinates x, y, z for each point of the object to be calculated. In order to produce color scans the execution of a series of digital images are also required. Then, using the software, the resulting images are automatically superimposed on the resulting scanning point cloud in such a way that every point is assigned to an appropriate color. It is necessary to carry out several measurements in order to fully scan objects that are only partially located in the field of view of the scanner. For the purpose of scanning, measuring control points are assumed around the object by using the technique of global navigation satellite systems GNSS. Measuring control points are used for geo-referencing and connection of the particular scans. Ground-based laser scanning is performed with positions oriented on the measuring fields set on measuring control points. The scan of the object is made up of dozens of stations, with an average resolution of approximately 1 cm. The combination and transformation of scans can be performed in a variety of programs (e.g. Cyclone program [6]), with a view to reaching the exact fit for the horizontal and height components. The type of scan described

above belongs to terrestrial scanning. Terrestrial scan data are usually insufficient to model roofs. Such information can be obtained from aerial scan data, but this technique is not suitable for the complete modeling of building façade. Therefore, complete data needed for geometrical modeling of analyzed structures can be obtained by combining the sets of data from aerial and terrestrial laser scanning. However, in this case problems can arise at the stage of combining data sets, resulting from different accuracy of both techniques. In the case of needing to improve the match between aerial and terrestrial sets of data, appropriate transformation algorithms can be used. These algorithms may much homologous surfaces by using a model of the method of least squares [7].

The elements constituting the process of creating the historical buildings model are the following:

(a) Approximation of scan data using planes (all elements of the buildings and architectural details are approximated by planes).
(b) Modeling of building edge as a result of the intersection of various planes.
(c) Checking and possible adjustment of topology which occur during the modeling of roofs.
(d) The creation of a 3D vector model for construction.

6.6 3D Modeling in Analysis of Historical Buildings

Modern technology related to 3D scanning significantly influences the process of research and documentation of historical buildings [8, 9]. The first step in this process is to acquire digital data which defines the geometry of the structure. The measured values are shown in formatted files (ASCII, PTS, LAS, E57, etc.), presenting the position, intensity and color of each point in the resulting point cloud. Appropriate processing of digital data makes it possible to apply collected information in the next step (Fig. 6.1). The data received are converted to the following file formats: DXF, DWG and edited in appropriate programs CAD, FEA and BIM (Building Information Modeling). The process of modeling and numerical

Fig. 6.1 A digital image of the interior of the Franciscan church in Gniezno (3D Scanning, scanning3d.pl)

Fig. 6.2 A digital image of the Franciscan church in Gniezno (3D Scanning, scanning3d.pl)

static-strength analysis of historical buildings is implemented by means of these programs. Generally, BIM programs are used to easily define a space with a small number of polygons. In contrast, 3D graphics programs and then FEM programs are used for the parametric design and analysis of complex geometric shapes created with polygons and splines. An important advantage of 3D programs is the generation of form by means of polygon mesh, which affects its very precise definition (Fig. 6.2). It is very important for the analysis of historical buildings because of the possibility of drawing up detailed documentation, accurate representation of geometry and making a calculation in 3D space by using FEM. Objects are created with a small and large number of polygons due to the assumed degree of accuracy level and priority in the model. Many graphical 3D programs allow the execution of Boolean operations by using the CSG technique (Constructive Solid Geometry) [7]. These operations include defining objects by a particular action executed on solids, i.e. union, subtraction and intersection. The presented technique is particularly useful in generating a model which allows for gap filling or the restoration of missing elements within a building. When carrying out analysis of historical buildings, it is very important to examine the precise geometry of the structure. The basic objective is to understand the actual condition of the structure concerning the damage and destruction of material and to perform numerical statical-strength analysis. It is necessary to precisely identify the geometry of the structure and its individual components (e.g. for historical churches: pillars, walls, buttresses and many different types of vaults) before adopting the appropriate mathematical model based on FEM. The geometry of these structures is often very complicated and preparation of a complete geometric model is only possible by using the technology of 3D scanning [10]. Despite this support, this task is very difficult and time-consuming. However, the biggest advantage of this approach is the ability to carry out a comprehensive analysis of the structure. This analysis illustrates the behavior of the structure in 3D.

However, there are many relevant parameters for the examination of these structures. These parameters include: dimensions of elements, shape of masonry blocks, bonding patterns, type of mortar and general homogeneity of the material. All historical buildings must be analyzed individually due to the wide variety of historical buildings in terms of geometry and static scheme.

6.7 The Procedures Needed to Obtain a Digital Model of Analyzed Structures

The numerical model of the building can be represented in many different ways after obtaining the point cloud with the use of 3D scanning. However, the most important in engineering practice are the following models [7] (Fig. 6.3), in the form of a regular square grid (GRID), which may be completed by characteristic points and digital lines (hybrid method), in the form of an irregular triangle mesh (TIN). Selection of mesh type is very important for the presentation of surface shape. The most commonly used mesh is TIN triangle and GRID regular square. The irregular triangle mesh is spread between the points designated in the measurements. An interpolation algorithm which enables the calculation of the height of the mesh node is used for the construction of the GFRID mesh. In the method of creating a digital model of buildings by using 3D scanning we can distinguish the following stages [7]:

(a) *3D scanning of a building*

In the first stage the execution of 3D scanning of a building which gives the disordered point cloud is carried out. The resulting point cloud is then recorded by using a mesh of control points (i.e. markers). It is impossible to capture the entire building using a single scan, so a scan with different positions is required. Therefore, the automatic correlation of images plays a very important role in the process of model generation. It is based on the automatic search and measurement of homologous points in two or more scans. For this purpose, numerous correlation methods (e.g. Area-Based Matching, Based Feature Matching, Relational Matching) have been developed [7].

Regardless of the method, the correlation process consists of the following activities:

- selection of elements which will be matched,
- finding their counterparts in subsequent images,
- calculation of the spatial coordinates of the matched elements,
- calculation of fitting accuracy.

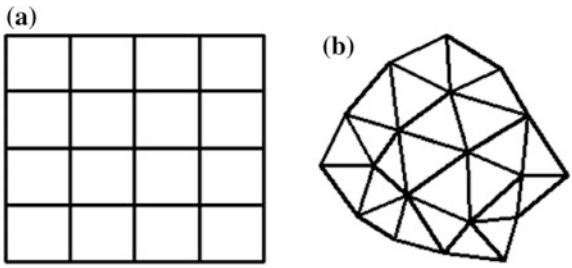


Fig. 6.3 **a** GRID mesh; **b** TIN mesh

The main correlation algorithm used in the process of creating a digital model of the 3D scan is an algorithm from the FBM group (Feature Based Matching). It is based on comparison of objects on the image. During the visualization process, local objects (points, edge elements, sections, small regions) as well as global objects (polygon, complex structure) are compared. Each of these elements has its own attributes, such as: coordinates in the imaging system, average optical density, orientation, gradient, length and curvature of the line. For each image a list of objects is compiled, which are then subjected to a process of matching. The method of least squares (minimum sum of the squares of deviations between attributes) is used to determine the similarity measure.

(b) *Filtering and cleaning up the point cloud*

This stage involves improving the quality of the point cloud in order to make the best representation of the analyzed object. For this purpose an appropriate filtering tool is used (e.g. software CATIA), which allows unnecessary points to be removed (called noise) as well as those that are repetitive. This tool can work in the unification and adaptive mode of the mesh. This stage is very important, because by removing unnecessary and incorrect points, the amount of data used by the system in subsequent steps is substantially reduced.

(c) *Generation of the polygonal surface*

In this step, the point cloud is combined into an adjacent triangle mesh. Delaunay triangulation is mostly used for creating the triangle mesh [7, 11]. The elements of Delaunay triangulation should be based on the circle criterion. It is satisfied if the circle which is circumscribed on the triangle does not contain points other than the vertices of the triangle (Fig. 6.4).

Moreover, interior angles should be fulfilled, which is related to the condition of maximizing the minimum angles. After the division of the triangle containing the data point on three smaller triangles (if the point does not lie on the edges), it is checked whether the operation satisfies the criterion of angles. This criterion determines the relationship that for any two adjacent triangles the sum of the interior angles adjacent to the common side is greater than the sum of the other two angles.

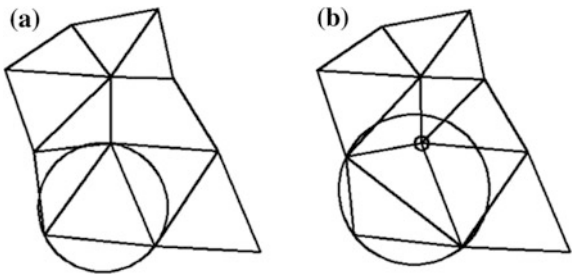


Fig. 6.4 Examples of: **a** correct triangles; **b** incorrect triangles

Optimal triangulation is defined by measuring angles, edge length, height or surface elements, whereas finite element approximation error in the FEM is usually associated with a minimum angle of elements. In order to eliminate gaps in the polygonal model new vertices can be added and the coordinates of existing vertices can be adjusted. Additionally, imperfections of the triangle mesh can be removed using a smoothing function. However, despite the filling of gaps in the mesh and smoothing of the surface, the model may be characterized by insufficient accuracy in certain zones. This model cannot be qualified as incorrect. This is due to the specific nature of 3D scanning. Intersection curves are smoothed in the next stage and occurring inaccuracies practically do not affect the final model of the analyzed structure.

(d) *Creating the wireframe model*

At this stage, there is a digital model, which is burdened with some errors. Because it is not parameterized, it cannot be modified. The purpose of further analysis is to obtain a parametric surface model. A key element of activity is the use of family intersection curves which are parallel with the others and located at equal intervals both horizontally and vertically. Their orientation may coincide with the main coordinate system, or be correlated with other systems and geometries. When selecting the position of the planes, attention should be given to characteristic places, i.e. areas where the shape of the surface is changed accordingly. This is particularly important in view of the representation of the final surface of the analyzed building model. In this way a mesh of profiles is obtained which are required to define the surface elements.

(e) *Creating the base area and surface model*

The next step is to generate surface elements based on the mesh that was previously created. Adjustment of the formed surface to an existing wireframe is possible after generating the control points. When modeling all surface elements, they should be cut and consolidated. It is important to check the correct direction of material orientation, as well as the proper sequence of combining planes.

6.8 Evaluation of the Accuracy of Created 3D Models

Algorithms for modeling and automation of this process are the subject of intensive research, and assessment of the accuracy of the 3D model is not a common issue in scientific literature. This is due to the fact that there is not a precisely defined level of accuracy of such models. Accuracy is usually reduced to provide a level of detail in modeling LoD and visual quality assessment. In the assessment of the accuracy of a 3D model, the following errors, which affect the final accuracy of the model, can be distinguish:

1. The accuracy of the terrestrial scanning.
2. The accuracy of the aerial scanning.
3. Integration errors of both data sets.
4. Generalization of the model (detail modeling) and coherence identification of individual model elements (roughness modeled surface).

Characteristics of both sensors show that the accuracy of terrestrial laser scanning, both planimetric and altitude is at the level of few centimeters, and it is about an order of magnitude better than the accuracy of aerial scanning (used for scanning roofs). The integration method of both cloud points must have an impact on the final accuracy of the model. Numerical terrain models interpolated from aerial scanning data are generally slightly, from a few to a dozen centimeters above the actual ground surface [12]. This is mainly due to the roughness of the scanned surface. A major source of error in the final model may be a generalization of architectural elements and roughness, or the lack of coherent identification of the modeled area. For example, the roof of modeled building is covered with tiles, which in the model is approximated by a plane. The ambiguity of identification of such a roof is at a level of 10 cm.

6.9 Software

Computer software is required at each stage of the laser scanning process. This includes the operation of the scanner, the processing of the collected data and the visualization and utilization of any delivered digital product. The choice of software will be based on a number of factors, including data quantity and the type of deliverable output required. Such software will offer a three-dimensional viewer that can be used to preview the dataset. It will enable the view to be rotated, zoomed and panned, colors to be changed and data to be clipped from view. The software will have been designed specifically to handle large volumes of three-dimensional measurements. Mainstream software for CAD, GIS or 3D modelling was not originally designed to handle the large datasets now routinely generated by laser scanning. In some cases specialist point cloud processing engines can be obtained to improve the performance of these mainstream tools, making it possible to use a familiar software environment.

Among the computer programs supporting laser scanning, the following can be distinguished: AutoCAD, Revit, Microstation, Geomagic, PolyWorks, Rapidform, Maya, Blender, Skyline, ArcGis, ArcInfo, Erdas, etc. The most popular programming languages used for the visualization of 3D models are VRML (VRML1, VRML2) and X3D. Very useful tools in this process are the library OPEN GL and DirectX/Direct 3D. They enable the creation of computer graphics in 2D and 3D, and draw complex three-dimensional structures from simple primitives.

6.10 Fem Analysis of Historical Building

The last step in the process of creating a digital model of historical buildings is to use the final form of the model for Finite Element Analysis. The obtained point cloud from 3D scanning can be converted to DWG or DXF files. Next, after importing the point model of the structure into FEM programs, each point is assigned a number and position in the coordinate system. These points are treated as nodes. In the next step, after performing a series of geometrical transformations, it is possible to generate a finite element mesh. For Finite Element Analysis of historical buildings various types of elements are used. Solid elements are used for modeling walls, pillars and buttresses, while for vaults shell elements are applied. The main problem of this analysis is the adoption of an appropriate model for a masonry structure, after obtaining the exact geometry of the scanned building. Basically, the calculation should take into account the two-component characteristic of masonry. Currently, the most accurate type of analysis is micro-modeling which assumes the separate properties of bricks, mortar and contact surface between brick and mortar. This type of analysis is suitable for structures with particularly strong heterogeneous areas of stress and deformation. When using micro-modeling all possible failure mechanisms can be identified and this enables mechanical phenomena at the level of the microstructure to be examined. However, this type of analysis has large limitations for applications in engineering practice. In this analysis, finite element discretization of masonry must be at least equivalent to the real brick bonding with mortar. Additional density of finite element mesh is often necessary. Application of this method is related to a considerable amount of computing time at relatively high hardware requirements. Also needed are a very large number of material parameters of masonry components supported by experimental studies.

In such a situation, implementation of homogenous material with equivalent uniform mechanical properties seems to be the ideal solution. This approach is referred to as homogenization. The mechanical parameter values of the homogenized masonry model, including Young's modules, Kirchhoff's modulus and Poisson's ratio is determined from the model of the representative masonry cell.

The historical masonry church (Fig. 6.5) with cross-ribbed vaults was analyzed with the use of FEM (Fig. 6.6). To span over rectangular bays, cross-ribbed vaults were applied. A cross vault is obtained by the intersection of two or more cylindrical vaults forming diagonal arches over the space to be covered. The arch action of each barrel vault brings all the loads as compressive forces down to the springing with an outward thrust at the supports. Buttressing forces are required to stabilize these diagonal ribs.

Computer calculations were performed using the LUSAS system ver. 14.7.

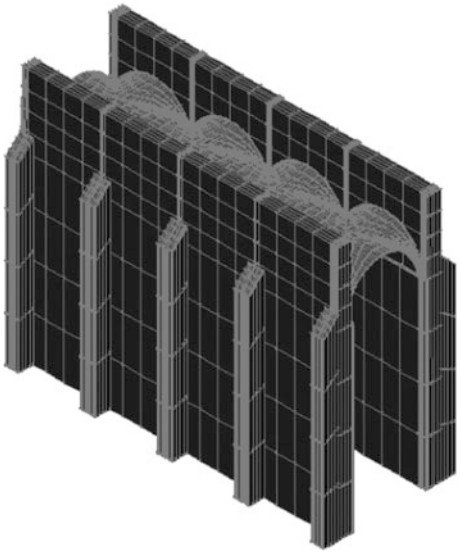


Fig. 6.5 Finite element analysis of the historical masonry church

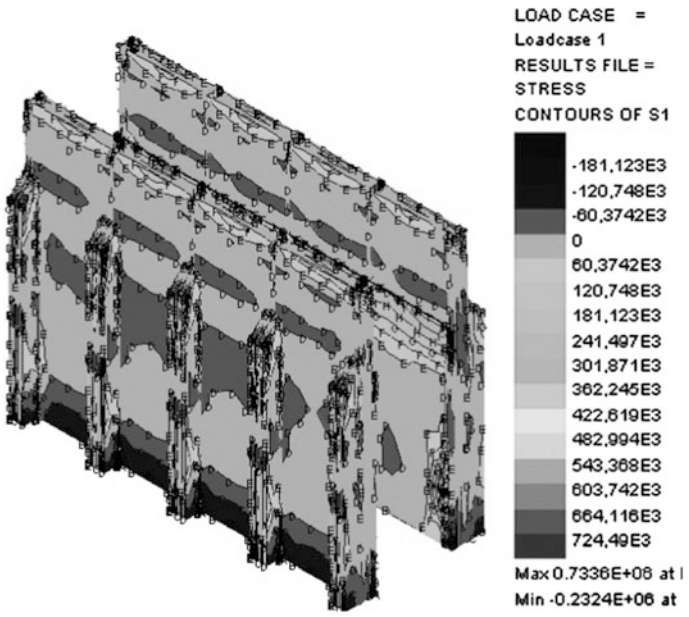


Fig. 6.6 Finite Element Analysis of the historical church, distribution of stress S1 (kPa)

6.11 Conclusions

The article contains an analysis of the application of 3D scanning in the process of creating a very precise numerical model of historical buildings. The obtained digital model allows for FEM analysis to be performed. 3D scanning is used in inventory documentation with the visualization of a building and the creation of 3D models in most cases. Despite the fact that laser scanning is now a leading and constantly developing measurement technique, processing and conversion of data in the form of a point cloud is still complicated and time-consuming. In the case of using a digital model of historical buildings, complexity of this model is twofold. Firstly, it is complexity of the architectural elements. Apart from walls and roof structures, there are balconies, cornices, pediments, columns, vaults, domes, carved decorative elements etc. These elements may come from different periods of the history of architecture and constitute themselves as complex structures. The second type of complexity of this subject is based on the number of architectural elements required to create the analyzed structures. Such an understanding of the building structure (with plans, images and a point cloud obtained from laser scanning) is important to correctly fit all parts of a structure in relation to adjacent elements. However, only simple geometric shapes are automatically adjusted to the points in the cloud, so they do not include such architectural elements as column grooves, volutes and vaults. Furthermore, only visible parts of the structure are visualized. There are always parts of the components that cannot be captured by the camera or laser scanner. However, for architectural elements their location is predictable. For example, the hidden or damaged part of the capital of Corinthian columns can be determined by knowledge of its construction. Finally, all used measurement techniques require knowledge of architecture and geometric design to restore constructional elements of an analyzed structure according to structural rules.

The overall conclusion is that measurement technology based on 3D scanning is slowly replacing traditional measurement inventory methods meeting the needs of the monitoring and protection of historical buildings. This technology not only allows a reduction in working time, by improving the efficiency, but also enables the creation of a 3D model, making precise drawings in CAD software and performing accurate statical-strength FEM analysis. The evolution of the 3D laser scanner is developed in order to increase the speed and accuracy of measurements and extend the research area from one measuring position. It provides the necessary impetus for further development of the use of laser scanning to protect historical buildings on a larger scale.

References

1. Szolomicki J (2015) Application of 3D laser scanning to computer model of historic buildings. In: Lecture notes in engineering and computer science: proceedings of the world congress on engineering and computer science 2015, WCECS 2015, 21–23 October 2015, San Francisco, USA, pp 858–863
2. Hichri N, Stefani C, De Luca L, Verona P, Hamon G (2013) From point cloud to BIM a survey of existing approaches. In: XXIV International CIPA Symposium, 2–6 Sept 2013, Strasbourg, France, pp 343–348
3. Cardaci A, Mirabella Roberti G, Versaci A (2011) From the continuous to the discrete model a laser scanning application to conservation projects. In: International archives of the photogrammetry, remote sensing and spatial information sciences, workshop, 2–4 March 2011, Trento, Italy, vol. XXXVIII-5/W16, pp 437–444
4. Jonak J (2013) CAD and FEA systems in practical applications. Monograph, Part 1, Lublin University of Technology, 2013, Lublin, Poland, pp 1–80 (in Polish)
5. Rychlik M (2015) Scanners—gateway to the virtual world (in Polish). http://www.designnews.pl
6. Leica Cyclone Model 8.1. Converting laser scans into implemented data. http://www.hds.leica-geosystems.com
7. Foryś A (2005) Comparison of effectiveness of the automatic generation of digital terrain model from aerial photographs for the various systems. MSc Thesis, Faculty of Mining Surveying and Environmental Engineering, AGH University of Science and Technology, 2005, Cracow, Poland, pp. 1–142 (in Polish)
8. Fregonese L, Barbieri G, Biolzi L, Bocciarelli M, Frigeri A, Taffurelli L (2013) Surveying and monitoring for vulnerability assessment of an ancient building. Sensors 13:9747–9773
9. Guarnieri A, Milan N, Vettore A (2008) Combined techniques for the stability control of historical structures. In: 13th FIG symposium on deformation measurement and analysis, 12–15 May 2008, Lisbon, Portugal, 15 pp
10. Kościuk J (2012) Modern laser scanning 3D for modeling documentation and conservation of architectural monuments. J Heritage Conserv 32:82–88
11. Jurczyk T (2000) Generation of unstructured triangular meshes by using Delaunay's triangulation. MSc Thesis, Department of Computer Science, AGH University of Science and Technology, 2000, Cracow, Poland, pp 1–74 (in Polish)
12. Borkowski A, Jóźków G (2012) Accuracy assessment of 3D modeling using laser scanning data, case study of Silesian Piast dynasty castle in Brzeg. Archiwum Fotogrametrii, Kartografii i Teledetekcji, vol 23, 2012, pp 37–47 (in Polish)

Chapter 7
A Simulation Study on Performance Validation of Indoor Location Estimation Based on the Radial Positive Distribution

Kosuke Okusa and Toshinari Kamakura

7.1 Introduction

We study the problem of analyzing indoor spatial location estimation by statistical radial distribution model. Recently we daily use the global positioning systems (GPS) for obtaining the location for car navigation. These systems are very convenient, but we sometimes need to the location estimation in the indoor environments for obtaining the nursing care information like in hospitals. Indoor location estimation based on the GPS is very difficult, because the GPS signals are difficult to receive in the indoor situation.

A study on the indoor spatial location estimation is very important in the fields of the marketing science, design for the public space. For instance, indoor spatial location estimation is one of the important methods to make the space planning based on the evacuation model, and shop layout planning [5, 6, 10–12].

Recently, indoor spatial location estimation is mainly based on Received Signal Strength (RSS) method [7, 18, 23], Angle of Arrival (AoA) method [14, 20] and Time of Arrival (ToA) method [4, 19, 22].

RSS is cost-effective approach, this method is able to using general radio signals (e.g. Wi-Fi networks). However, the signal strength is affected by signal reflections and attenuation and hence it is not robust. Therefore, RSS method's location estimation accuracy is very low. AoA is most high accuracy approach, this method is able to use signal arrival directions and estimated distances. However, this system is very high cost, because this system requires array signal receivers. On the other hand, ToA method is only using the distance between transmitter and receiver. This

K. Okusa (✉)
Kyushu University, 4-9-1 Shiobaru, Minami-ku, Fukuoka 815-8540, Japan
e-mail: okusa@design.kyushu-u.ac.jp

T. Kamakura
Chuo University, 1-13-27 Kasuga, Bunkyo-ku, Tokyo 112-8551, Japan
e-mail: kamakura@indsys.chuo-u.ac.jp

S.-I. Ao et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-981-10-2717-8_7

method's accuracy is higher than RSS method, and cost is lower than AoA method. For this reason, previous research has suggested that ToA method is most suitable for the practical indoor location estimation system [17].

In this research, we use the ToA data based measurement system. ToA data based measurement system's location estimation algorithm is mainly based on the least squares method. Least squares method based approach is difficult to process the outlier value, and such data is frequently observed in the ToA data. To cope with this, Kamakura and Okusa [8], Okusa and Kamakura [15] proposed that indoor location estimation algorithm base on the marginal likelihoods of radial distribution generated by positive distribution among the several transmitter radio sites placed in the room. However, this research only focusing on the radial Weibull distribution. Generally, there are more positive distributions exists and their work is not validate the other distributions performance.

We shall compare the radial Weibull distribution and other radial positive distribution based location estimation. In this paper we compare the performance of radial Weibull distribution and radial log-normal distribution.

To compare the performance of each distribution based approach, we conducted simulation study of location estimation. As a result, radial Weibull distribution based method shows high accuracy for the indoor spatial location estimation.

The organization of the rest of the paper is as follows. In Sect. 7.2, we discuss the feature and problem of ToA data. In Sect. 7.3, we present our models for indoor location estimation based on the radial Weibull distribution and radial log-normal distribution. In Sect. 7.4, we present some performance results showing the effectiveness of our model based on the simulation study. We conclude with a summary in Sect. 7.5.

7.2 Time of Arrival (ToA) Data

ToA is one of the methods of estimate the distance between transmitter and receiver. This method is computed from radio signal's travel time between transmitter and receiver. When the transmitter and receiver's time has been completely synchronized, distance d between transmitter and receiver is calculated as follows:

$$d = C(r_t - r_r) \tag{7.1}$$

where r_t, r_r are transmitted and received time, respectively. C is the speed of light. In the ideal circumstance, d provides accurate distance between transmitter and receiver, which called Line-of-Sight. In this case, subject location is easy to estimate by trilateration (Fig. 7.1).

However, in many cases, distance d includes error components, which called Non Line-of-Sight (NLoS) [2, 21] (Fig. 7.2).

NLoS conditions is mainly caused by obstacles between transmitter and receiver, signal reflections. In this case, observed distance d will be larger than true distance.

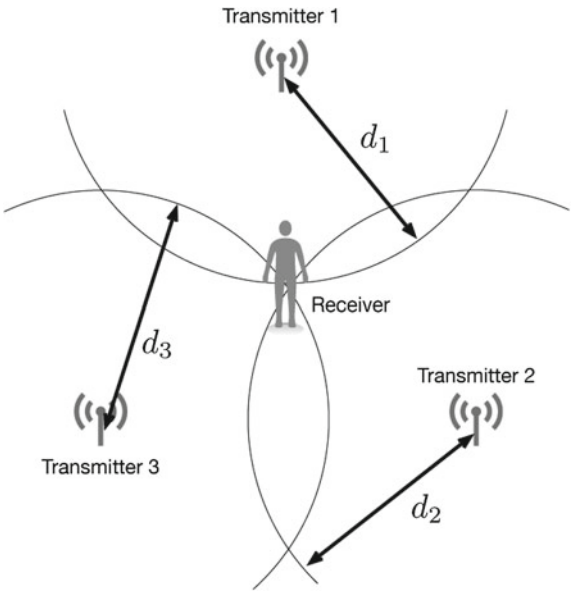


Fig. 7.1 Location estimation by trilateration (ideal case)

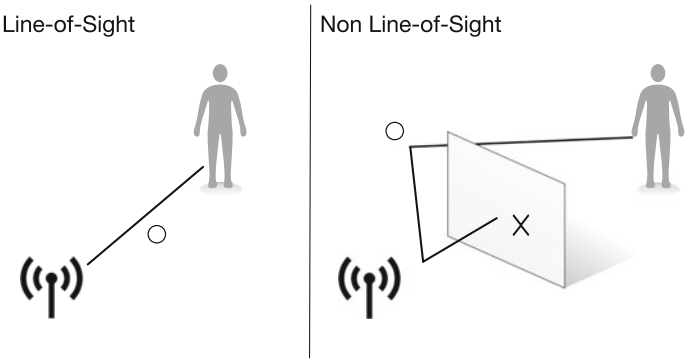


Fig. 7.2 LoS, NLoS circumstance

Fujita et al. [3] reported that observed distance of LoS and NLoS are defined as follows:

$$d_{k,LoS} = \sqrt{(x - c_{k1})^2 + (y - c_{k2})^2} + e_k \tag{7.2}$$

$$d_{k,NLoS} = \sqrt{(x - c_{k1})^2 + (y - c_{k2})^2} + e_k + b_k \tag{7.3}$$

LoS case, observed value is distributed from true distance with error term $e_k \sim N(0, \sigma_k^2)$, where $N(\cdot)$ is normal distribution. On the other hand, NLoS case, observed

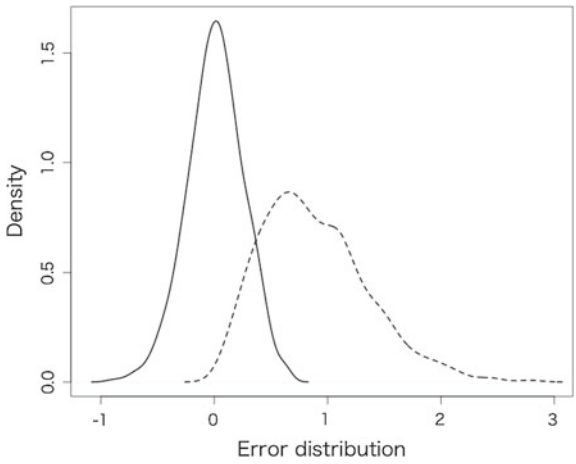


Fig. 7.3 Error density between true distance and observed distance

value is contains the additional bias term $b_k \sim U(0, B_{\max})$, where $U(\cdot)$ is uniform distribution. $B_{\max}$ means possible maximum bias value of observed value.

Figure 7.3 shows example of error density between true distance and observed distance. Here solid line represents LoS case density, and dashed line represents NLoS case. Figure 7.3 indicates that NLoS case's estimated distance will not distributed in true distance.

To cope with this, some researchers focus on the model based MLE approach [9, 13, 16], but these methods were modelized by 1-D distribution. ToA data is only provided distance data, and not provided angle data. We can assume that the observed data is a 2-D distribution and propose the statistical radial distribution.

7.3 Indoor Location Estimation Algorithm

In this section, we describe the proposed indoor location estimation algorithm. The proposed method is based on the marginal likelihoods of radial distribution generated by positive distribution. Firstly, we explain the probability density function and likelihood function of radial Weibull distribution. Secondly, we explain the probability density function and likelihood function of radial log-normal distribution. Finally, we describe the location estimation method based on the radial distribution.

7.3.1 Radial Weibull Distribution

Considering that the obtained distance data are all positive, we propose the following circular distribution based on the 3-parameter Weibull distribution:

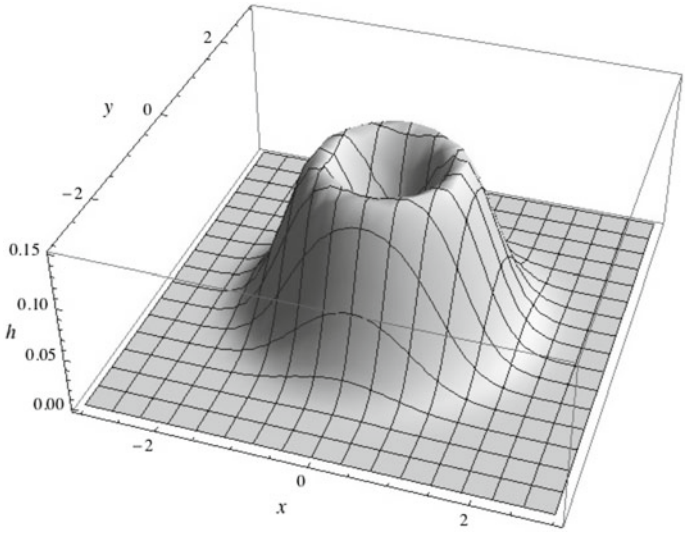


Fig. 7.4 Radial Weibull distribution ($m = 3, \lambda = 0.5, g = 0$)

$$f(r,\theta) = \frac{1}{2\pi}\left(\frac{m}{\eta}\right)\left(\frac{r-g}{\eta}\right)^{m-1} \exp\left\{-\left(\frac{r-g}{\eta}\right)^{m}\right\}$$
$$(r, g, \eta, m > 0,\ 0 \le \theta < 2\pi). \tag{7.4}$$

Here g is location parameter, η, m are Weibull distribution's shape, scale parameter.

From Eq. 7.4, converts from Polar to Cartesian coordinates:

$$g(x,y) =$$
$$\frac{\lambda m}{2\pi}(\sqrt{x^2+y^2} - g)^{m-2} \exp\left\{-\lambda(\sqrt{x^2+y^2} - g)^{m}\right\}$$
$$(x, y, g, \lambda, m > 0). \tag{7.5}$$

Here $\lambda = 1/\eta^m$ (Fig. 7.4).

Assuming that each transmitter station observers independent measurements, the likelihood based on the data set is calculated as follows:

$$L(\lambda_1, m_1, g_1, ..., \lambda_K, m_K, g_K) =$$
$$\prod_{i=1}^{K}\prod_{j=1}^{n_i} \frac{\lambda_i m_i}{2\pi}(\sqrt{(x_{ij}-c_{i1})^2 + (y_{ij}-c_{i2})^2} - g_K)^{m-2}$$
$$\exp\left[-\lambda_i\{\sqrt{(x_{ij}-c_{i1})^2 + (y_{ij}-c_{i2})^2} - g_K\}^{m}\right]. \tag{7.6}$$

Here K is the number of stations and for each station i the sample size is n_i. The observed data set for station i is (x_{ij}, y_{ij}). The coordinates (c_{i1}, c_{i2}) are given transmitter station positions (Fig. 7.1).

We can assume that most high likelihood value point $(\hat{x}, \hat{y})$ as a subject estimated location.

7.3.2 Radial Log-Normal Distribution

As same as the radial Weibull distribution, considering that the estimated distance data are all positive, we propose the following circular distribution based on the log-normal distribution:

$$f(r,\theta) = \frac{1}{\sqrt{2\pi}\sigma r}\exp\left\{-\frac{(\log r-\mu)^2}{2\sigma^2}\right\}$$
$$(\mu,\sigma,r>0,\ 0\le\theta<2\pi). \tag{7.7}$$

Here μ is location parameter, σ is scale parameter.

From Eq. 7.7, converts from Polar to Cartesian coordinates (Fig. 7.5):

$$g(x,y) = \frac{1}{\sqrt{2\pi}\sigma(x^2+y^2)}\exp\left\{-\frac{(\log\sqrt{x^2+y^2}-\mu)^2}{2\sigma^2}\right\}$$
$$(x,y,\mu,\sigma>0). \tag{7.8}$$

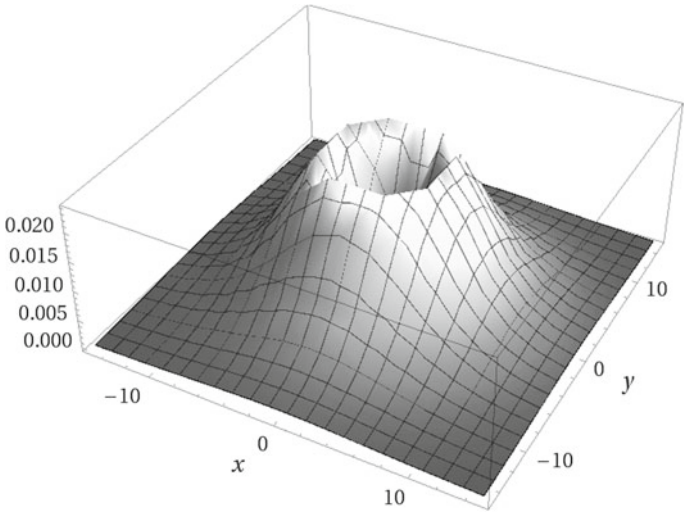


Fig. 7.5 Radial log-normal distribution ($\mu = 2, \sigma = 0.5$)

Assuming that each transmitter station observers independent measurements, the likelihood based on the data set is calculated as follows:

$$L(\mu_1, \sigma_1, ..., \mu_K, \sigma_K) = \prod_{i=1}^{K} \prod_{j=1}^{n_i} \frac{1}{\sqrt{2\pi}\sigma_i\{(x_{ij} - c_{i1})^2 + (y_{ij} - c_{i2})^2\}} \times \exp\left\{ - \frac{(\log \sqrt{(x_{ij} - c_{i1})^2 + (y_{ij} - c_{i2})^2} - \mu_i)^2}{2\sigma_i^2} \right\}. \tag{7.9}$$

Here K is the number of stations and for each station i the sample size is n_i. The observed data set for station i is (x_{ij}, y_{ij}). The coordinates (c_{i1}, c_{i2}) are given transmitter station positions (Fig. 7.1). We can assume that most high likelihood value point $(\hat{x}, \hat{y})$ as a subject estimated location.

7.3.3 Location Estimation Based on the Radial Distribution

The idea of location estimation using the radial positive distribution is based on the marginal likelihoods. Firstly, we estimate the each transmitter station's distribution parameter using observed value. Secondly, we calculate the joint estimated likelihood of each transmitter station's radial distribution. We can assume that most high likelihood value point as a subject estimated location (Fig. 7.6).

From this idea, most high likelihood value point $(\hat{x}, \hat{y})$ of radial Weibull distribution is calculated as follows:

$$(\hat{x}, \hat{y}) = \arg\max(\hat{\lambda}_1, \hat{m}_1, \hat{g}_1, ..., \hat{\lambda}_K, \hat{m}_K, \hat{g}_K) = \prod_{i=1}^{K} \frac{\lambda_i m_i}{2\pi} (\sqrt{(x - c_{i1})^2 + (y - c_{i2})^2} - g_K)^{m-2} \exp\left[- \lambda_i \{ \sqrt{(x - c_{i1})^2 + (y - c_{i2})^2} - g_K \}^m \right]. \tag{7.10}$$

As same as the radial Weibull distribution, most high likelihood value point $(\hat{x}, \hat{y})$ of radial log-normal distribution is calculated as follows:

$$(\hat{x}, \hat{y}) = \arg\max(\hat{\mu}_1, \hat{\sigma}_1, ..., \hat{\mu}_K, \hat{\sigma}_K) = \prod_{i=1}^{K} \frac{1}{\sqrt{2\pi}\sigma_i\{(x - c_{i1})^2 + (y - c_{i2})^2\}} \times \exp\left\{ - \frac{(\log \sqrt{(x - c_{i1})^2 + (y - c_{i2})^2} - \mu_i)^2}{2\sigma_i^2} \right\}. \tag{7.11}$$

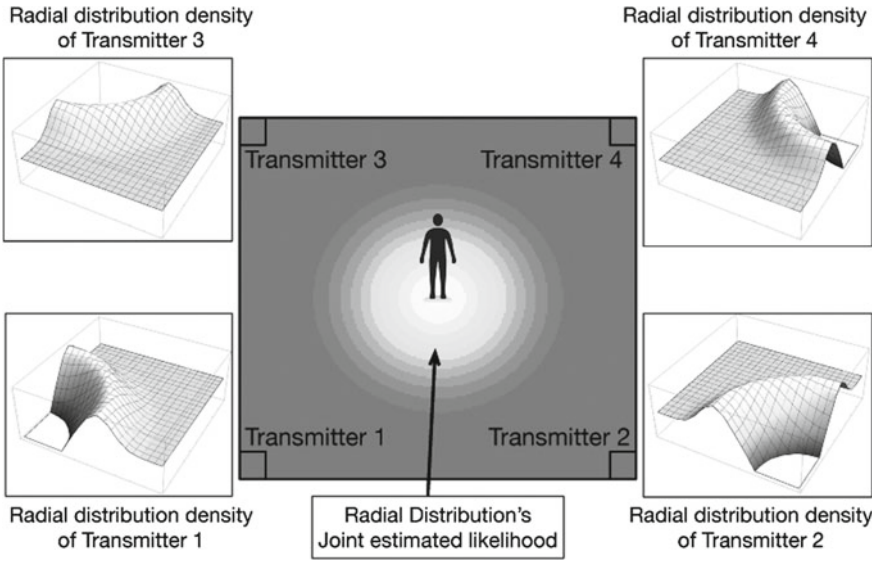


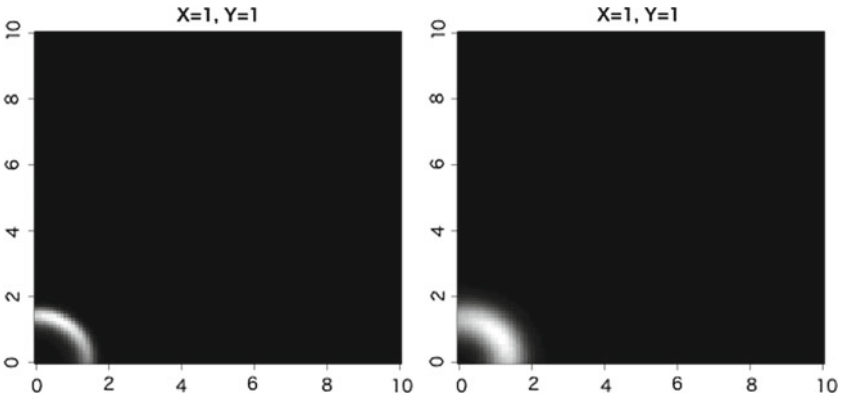
Fig. 7.6 Location estimation based on the radial distribution

In next section, we compare the performance of radial Weibull distribution and radial log-normal distribution for the indoor location estimation based on the simulation study.

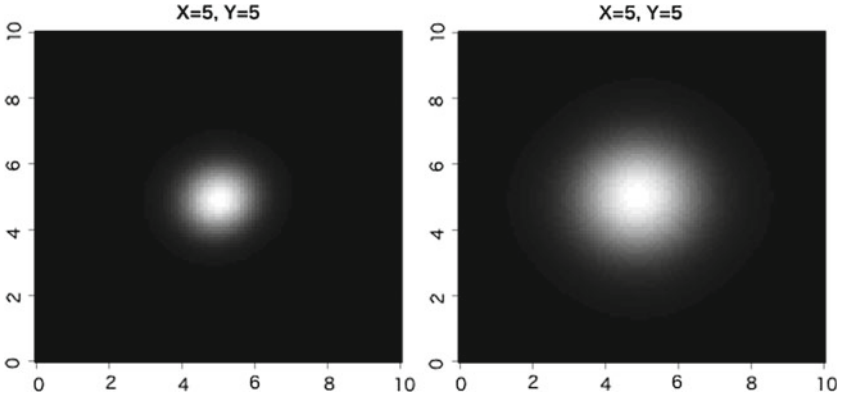
7.4 Simulation Study

To compare the performance of each distribution based approach, we conducted simulation study of location estimation. We consider the NLoS situation simulation of 10[m] × 10[m] chamber, and estimate the subject location at (1, 1), (5, 5), (5, 2.5).

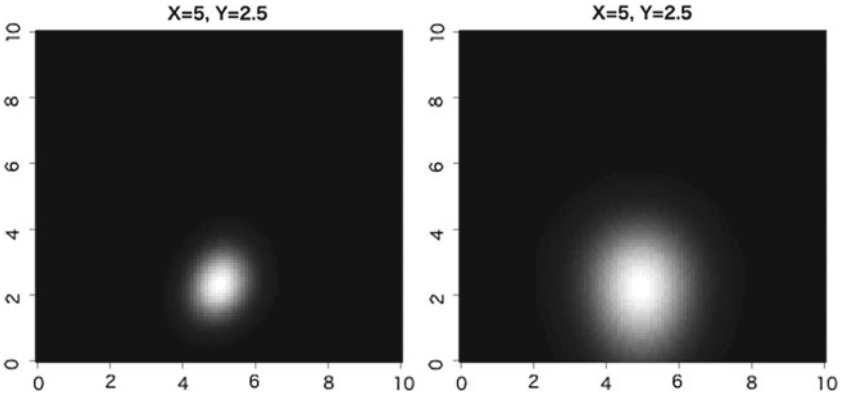
NLoS random value is generated by $d_{k,NLoS} = \sqrt{(x - c_{k1})^2 + (y - c_{k2})^2} + e_k + b_k$ from Eq. 7.2, where e_k, b_k are $N(0, \sigma^2)$, $B(0, B_{\max})$ random value, respectively. We set the σ^2 of normal distribution's random value to variance value of observed data from Okusa and Kamakura [15]. $B_{\max}$ of uniform distribution's random value to $U(0, 2)$, sample size $n = 100$ at each location. The number of times of simulation is 10,000 and compare the performance of each distribution based on the AIC [1] and average error distance from true location.



(a) subject location $(1,1)$



(b) subject location $(5,5)$



(c) subject location $(5,2.5)$

Fig. 7.7 Radial Weibull distribution (*left side*), Radial log-normal distribution (*right side*)'s marginal likelihood

Table 7.1 AIC and average error distance from true location of each distribution based approach

	AIC	Average error dist. (m)
Radial Weibull distribution	115.3	0.47
Radial log-normal distribution	242.1	0.91

7.4.1 Simulation Results

Figure 7.7 is radial Weibull distribution (left side) and radial log-normal distribution (right side)'s marginal likelihood of subject location at $(1, 1)$, $(5, 5)$, $(5, 2.5)$. Figure color means marginal likelihood value, and light tone means high probability. From this figure, radial Weibull distribution based method is able to estimates subject location precisely than radial log-normal distribution based method.

Table 7.1 is AIC and average error distance from true location of each distribution based approach. As same as the marginal likelihood results, radial Weibull distribution based approach is better performance than radial log-normal based approach.

7.5 Conclusions

In this article, we compared the location estimation performance of radial Weibull distribution and radial log-normal distribution based on the simulation study. From the simulation results, radial Weibull distribution based approach is better performance than radial log-normal based approach based on the AIC and average error distance.

Note that the this article's approach is focusing on the ToA based dataset. The reason is because the radial log-normal distribution based approach is good performance for RSS dataset. Further studies are needed in order to verify the radial distribution's performance.

In next phase, we need to compare our method with more another indoor location estimation method, and its calculation performance. Additionally, we need to implement the indoor location estimation system based on the proposed method and demonstrate it.

References

1. Akaike H (1973) Information theory and an extention of the maximum likelihood principle. In: International symposium on information theory, pp 267–281
2. Chan YT, Tsui WY, So HC, Ching PC (2006) Time-of-arrival based localization under NLOS conditions. IEEE Trans Veh Technol 55(1):17–24

3. Fujita T, Ohtsuki T, Kaneko T (2007) Low complexity TOA localization algorithm for NLOS environments. Res Paper IPS 15:69–74
4. He J, Geng Y, Pahlavan K (2012) Modeling indoor TOA ranging error for body mounted sensors. In: IEEE International symposium on personal indoor and mobile radio communications (PIMRC), pp 682–686
5. Herrera JC, Hinkenjann A, Ploger PG, Maiero J (2013) Robust indoor localization using optimal fusion filter for sensors and map layout information. In: Proceedings of the IEEE conference on indoor positioning and indoor navigation, pp 1–8
6. Ikeda T, Kawamoto M, Sashima A, Tsuji J, Kawamura H, Suzuki K, Kurumatani K (2014) An indoor autonomous positioning and navigation service to support activities in large-scale commercial facilities. Serviceol Serv:191–201
7. Kaemarungsi K, Krishnamurthy P (2004) Properties of indoor received signal strength for WLAN location fingerprinting. In: IEEE confernce on in mobile and ubiquitous systems: networking and services, pp 14–23
8. Kamakura T, Okusa K (2013) Estimates for the spatial locations of the indoor objects by radial distributions. In: Proceedings of the ISI world statistics congress 2013 (ISI2013), CPS109, pp 3417–3422
9. Knapp CH, Carter GC (1976) The generalized correlation method for estimation of time delay. IEEE Trans Acoust Speech Signal Process 24(4):320–327
10. Li N, Becerik-Gerber B (2011) Performance-based evaluation of RFID-based indoor location sensing solutions for the built environment. Adv Eng Inf 25(3):535–546
11. Liu H, Darabi H, Banerjee P, Liu J (2007) Survey of wireless indoor positioning techniques and systems. IEEE Trans Syst Man Cybern Part C: Appl Rev 37(6):1067–1080
12. Martin W (2014) Indoor location-based services—prerequisites and foundations. Springer International Publishing. ISBN: 978-3-319-10698-4
13. Moses RL, Krishnamurthy D, Patterson RM (2003) A self-localization method for wireless sensor networks. EURASIP J Appl Signal Process 348–358
14. Niculescu D, Nath B (2003) Ad hoc positioning system (APS) using AOA. INFOCOM 3:1734–1743
15. Okusa K, Kamakura T (2015) Indoor Location Estimation based on the statistical spatial modeling and radial distributions. In: Lecture notes in engineering and computer science: proceedings of the world congress on engineering and computer science 2015, WCECS 2015, 21–23 Oct 2015, San Francisco, USA, pp 835–840
16. Patwari N, Hero AO III, Perkins M, Correal NS, O'dea, R. J. (2003) Relative location estimation in wireless sensor networks. IEEE Trans Signal Process 51(8):2137–2148
17. Patwari N, Ash JN, Kyperountas S, Hero AO III, Moses RL, Correal NS (2005) Locating the nodes: cooperative localization in wireless sensor networks. Signal Process Mag 22(4):54–69
18. Roos T, Myllymaki P, Tirri H, Misikangas P, Sievnen J (2002) A probabilistic approach to WLAN user location estimation. Int J Wirel Inf Netw 9(3):155–164
19. Shen J, Molisch AF, Salmi J (2012) Accurate passive location estimation using TOA measurements. IEEE Trans Wirel Commun 11(6):2182–2192
20. Taponecco L, D'Amico AA, Mengali U (2011) Joint TOA and AOA estimation for UWB localization applications. IEEE Trans Wirel Commun 10(7):2207–2217
21. Venkatraman S, Caffery J Jr, You HR (2004) A novel TOA location algorithm using LOS range estimation for NLOS environments. IEEE Trans Veh Technol 53(5):1515–1524
22. Watabe T, Kamakura T (2010) Localization algorithm in the indoor environment based on the ToA data. In: Australian statistical conference 2010 program and abstracts, p 243
23. Yamada I, Ohtsuki T, THisanaga T, Zheng L (2008) An indoor position estimation method by maximum likelihood algorithm using received signal strength. SICE J Control Meas Syst Integr 1(3):251–256

Chapter 8
Semi-supervised Classification with Modified Kernel Partial Least Squares

Paweł Błaszczyk

8.1 Introduction

Feature extraction, classification, and clustering are the basic methods used to analyze and interpret multivariate data. For classification task, the datasets contain vectors of features belonging to certain classes. These vectors are called samples. On the other hand, for the purpose of clustering, we do not have information about proper classification of objects. In datasets for classification tasks, the number of samples is usually much smaller compared to the number of features. In this situation, the small number of samples makes it impossible to estimate the classifier parameters properly; therefore, the classification results may be inadequate. In the literature, this phenomenon is known as the Curse of Dimensionality. In this case, it is important to decrease the dimension of the feature space. This can be done either by feature selection or feature extraction. Some of the linear feature extraction methods are for example Principal Component Analysis (PCA) and Partial Least Squares (PLS). These methods are often applied in chemometrics, engineering, computer vision, and many other applied sciences. However, the classical approach to feature extraction is based on the mean and the sample covariance matrix. It means that these methods are sensitive to outliers. Moreover, when the features and the target variables are non-linearly related, linear methods cannot properly describe the data distribution. Different non-linear versions of PCA and PLS have been developed (see [1–3]). In real classification task, we often have the dataset with relatively small amount of labeled data and a huge amount of data without labels. In real applications, we frequently encounter the problem with obtaining labeled data, as it is both time-consuming and capital-intensive. Sometimes, it requires specialized equipment or expert knowledge. Labeled data is very often associated with intense human labor, as in most applications, each of the examples need to be marked manually. In such situations,

P. Błaszczyk (✉)
University of Silesia, Katowice, Poland
e-mail: pawel.blaszczyk@us.edu.pl

S.-I. Ao et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-981-10-2717-8_8

semi-supervised learning can have a great practical value. The semi-supervised techniques allow us to use both labeled and unlabeled data. Including the information coming from unlabeled data and semi-supervised learning, we can improve the feature extraction task. Unlabeled data, when used in conjunction with a small amount of labeled data, can improve learning accuracy.

This paper is a revised and extended version of [4], presented on the World Congress on Engineering and Computer Science 2015. In this paper, we present a semi-supervised method for nonlinear feature extraction. We propose to combine a kernel for modified Partial Least Squares method with a Gaussian Mixture Model (GMM) (see [5, 6]) clustering algorithm. The supervised kernel exploits the information conveyed by the labeled samples and the cluster kernel from the structure of the data manifold. The proposed semi-supervised method was successfully tested in economical datasets and compared with other well known classification algorithms.

The paper is organized as follows. Section 8.2 reviews the standard PLS algorithm, modified version of PLS with different separation cirteria and a classification task. The Sect. 8.2.3 describe the GMM model. In next Sect. 8.2.5 we present the semisupervised algorihm for semi-supervise learning. In Sect. 8.3 we describe our experiment and present the the experimental results. Finally, Sect. 8.4 concludes this paper.

8.2 Methodology

Let us assume that we have the L-classes classification problem and let $(x_i, y_i) \in X \times \{C_1, \ldots, C_L\}, x \in \mathbb{R}^p$ where matrix of sample vectors X and response matrix Y are given by the following formulas:

$$X = \begin{bmatrix} x_{11} & \ldots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \ldots & x_{np} \end{bmatrix}, \quad Y = \begin{bmatrix} 1 & 0 & \ldots & \ldots & 0 \\ \vdots & & & & \vdots \\ 0 & 0 & \ldots & 0 & 1 \end{bmatrix}. \tag{8.1}$$

Each row of the matrix Y contain 1 in a position denoting the class label.

8.2.1 Partial Least Squares

One of the commonly used feature extraction methods is the Partial Least Squares (PLS) Method (see [7–9]). PLS uses of the least squares regression method [10] in the calculation of loadings, scores and regression coefficients. The idea behind the classic PLS is to optimize the following objective function:

$$(w_k, q_k) = \arg\max_{w^T w=1; q^T q=1} \operatorname{cov}\left(X_{k-1} w, Y_{k-1} q\right) \tag{8.2}$$

under conditions:

$$w_k^T w_k = q_k q_k^T = 1 \quad \text{for } 1 \leq k \leq d, \tag{8.3}$$

$$t_k^T t_j = w_k^T X_{k-1}^T X_{j-1} w_j = 0 \text{ for } k \neq j, \tag{8.4}$$

where $\text{cov}\left(X_{k-1}w, Y_{k-1}q\right)$ is a covariance matrix between $X_{k-1}w$ and $Y_{k-1}q$, vector t_k is the k-th extracted component, w_k is the vector of weights for k-th component and d denotes the number of extracted components. The matrices X_k, Y_k arise from X_{k-1}, Y_{k-1} by using so called deflation technique which removes the k-th component using the following formulas:

$$X_{(k+1)} = X_k - t_k t_k^T X_k \tag{8.5}$$

$$Y_{(k+1)} = Y_k - t_k t_k^T Y_k \tag{8.6}$$

Extracted vector w_k corresponds to the eigenvector connected with the largest eigenvalue of the following eigenproblem:

$$X_{k-1}^T Y_{k-1} Y_{k-1}^T X_{k-1} w_k = \lambda w_k \tag{8.7}$$

Let S_B denote the between scatter matrix and S_W within scatter matrix, given by:

$$S_B = \sum_{i=1}^{L} p_i \left(M_i - M_0\right)\left(M_i - M_0\right)^T, \tag{8.8}$$

$$S_W = \sum_{i=1}^{L} p_i E\left[\left(X - M_i\right)\left(X - M_i\right)^T | C_i\right] = \sum_{i=1}^{L} p_i S_i, \tag{8.9}$$

where S_i denotes the covariance matrix, p_i is a priori probability of the appearance of the i-th class, M_i is the mean vector for the i-th class, and M_0 is given by:

$$M_0 = \sum_{i=1}^{L} p_i M_i. \tag{8.10}$$

These matrices are often used to define separation criteria for evaluating and optimizing the separation between classes. For the PLS, a separation criterion is used to find vectors of weights that provide an optimal separation between classes in the projected space. In PLS method the matrix in each k-th step is:

$$X_k^T Y_k Y_k^T X_k = \sum_{i=1}^{L} n_i^2 \left(M_i - M_0\right)\left(M_i - M_0\right)^T \tag{8.11}$$

This matrix is almost identical to the between class scatter matrix S_B. Hence, we can say that the separation criterion in the PLS method is based only on the between scatter matrix. It means that the classic PLS method is that it does not properly separates the classes. To provide a better separation between classes we can use weighted separation criterion (see [11]) denoted by:

$$J = \text{tr}\left(\gamma S_B - (1-\gamma)S_W\right). \tag{8.12}$$

where γ is a parameter from interval $[0, 1]$, S_B and S_W are between scatter matrix and within scatter matrix, respectively. Applying a linear transformation criterion, condition (8.12) can be rewritten in the following form:

$$J(w) = \text{tr}\left(w^T\left(\gamma S_B - (1-\gamma)S_W\right)w\right). \tag{8.13}$$

which is more suitable for optimization. The next step is to optimize the following criterion:

$$\max_{w_k} \sum_{k=1}^{d} w_k^T\left(\gamma S_B - (1-\gamma)S_W\right)w_k, \tag{8.14}$$

under the conditions:

$$w_k^T w_k = 1 \quad \text{for } 1 \le k \le p. \tag{8.15}$$

The solution to this problem can be found using the Lagrange multipliers method.

Theorem 8.1 *Under the conditions (8.15) for the optimization problem (8.14) the solution is defined by eigenvector connected with the largest eigenvalue for the following eigenproblem:*

$$\left(\gamma S_B - (1-\gamma)S_W\right)w_k = \alpha w_k. \tag{8.16}$$

Proof The solution to this problem can be found with the use of the Lagrange multipliers method. Let

$$L\left(w_k, \alpha\right) = \sum_{k=1}^{p} w_k^T\left(\gamma S_B - (1-\gamma)S_W\right)w_k - \alpha_k\left(w_k^T w_k - 1\right),$$

be the Lagrangian function where $\alpha = \left[\alpha_1, \cdots, \alpha_p\right]$ is the vector of Lagrange multipliers connected with the conditions (8.15). For such defined optimization task we maximize function L with respect to w. This problem can be solved by maximization of the function L by providing the following conditions:

$$\begin{aligned} \frac{\partial L\left(w_k, \alpha\right)}{\partial w_k} &= 0 \quad \text{for } 1 \le k \le p \\ \alpha_k &\ge 0 \quad \text{for } 1 \le k \le p. \end{aligned} \tag{8.17}$$

Calculating the partial derivative of L with respect to w_k we obtain:

$$\frac{\partial L}{\partial w_k} = 2\left(\gamma S_B - (1-\gamma)S_W\right) w_k - 2\alpha w_k.$$

Then, from (8.17), we get:

$$\left(\gamma S_B - (1-\gamma)S_W\right) w_k = \alpha w_k.$$

This completes the proof.

To find the correct value of the parameter γ, we used the following metric:

$$\rho(C_1, C_2) = \min_{c_1 \in C_1, c_2 \in C_2} \rho(c_1, c_2), \tag{8.18}$$

where C_i is the i-th class for $i \in \{1, 2\}$. The value of the parameter γ was chosen by the using the following formula:

$$\gamma = \frac{\min_{i,j=1,\ldots,L, i \neq j} \left\{\rho(C_i, C_j)\right\}}{1 + \min_{i,j=1,\ldots,L, i \neq j} \left\{\rho(C_i, C_j)\right\}}. \tag{8.19}$$

Parameter γ equals 0 if and only if certain i and j classes exist for which $\rho(C_i, C_j) = 0$. This means that at least one sample which belongs to classes C_i and C_j, exist. If distance between classes increase, the value of γ also increases. Therefore the importance of the component S_W becomes greater.

To improve separation between classes in classic PLS method, we replace the matrix (8.11) with the matrix from our separation criterion (8.13) (see [11]) to optimize the objective criterion

$$w_k = \arg\max_w \left(w^T(\gamma S_B - (1-\gamma)S_W)w\right), \tag{8.20}$$

under the following conditions:

$$w_k^T w_k = 1 \quad \text{for } 1 \leq k \leq d \tag{8.21}$$

$$t_k^T t_j = w_k^T X_{k-1}^T X_{j-1} w_j = 0 \quad \text{for } k \neq j, \tag{8.22}$$

We call this extraction algorithm, i.e., Extraction by applying Weighted Criterion of Difference Scatter Matrices (EWCDSM).

Theorem 8.2 *For the input matrices X and Y the extracted vector w_k corresponds to the eigenvector connected with the largest eigenvalue for the following eigenproblem:*

$$(\gamma S_B - (1-\gamma)S_W)w = \lambda w. \tag{8.23}$$

Also, the k-th component t_k corresponds to the eigenvector related to the largest eigenvalue for the following eigenproblem:

$$X_{k-1}X_{k-1}^T\left(D-(1-\gamma)I\right)t_k=\lambda t_k. \tag{8.24}$$

Matrix $D=[D_j]$ is an $n\times n$ block-diagonal matrix, where D_j is a matrix in which all elements equals $1/nn_j$, where n_j is the number of samples in the j-th class.

Proof From the definition of the between scatter matrix we have

$$S_B=X^TDX, \tag{8.25}$$

where $D=[D_j]$ is an $n\times n$ block-diagonal matrix, where D_j is a matrix in which all elements equals $1/nn_j$, where n_j is the number of samples in the j-th class. Also we know that covariance matrix S_T has the following form

$$S_T=X^TX. \tag{8.26}$$

Using (8.25) and (8.26) we obtain

$$S_B-(1-\gamma)S_T=X^TDX-(1-\gamma)X^TX=X^T\left(D-(1-\gamma)I\right)X, \tag{8.27}$$

where I is identity matrix. In the EWCDSM the criterion WCDSM is used. So because of Theorem 8.1 we have that in the EWCDSM algorithm in k-th step we extract k-th component by calculate the eigenvector connected with the biggest eigenvalue for the following eigenproblem

$$(\gamma S_B-(1-\gamma)S_W)w=\lambda w.$$

Using (8.27) we get

$$X_{k-1}^T\left(D-(1-\gamma)I\right)X_{k-1}w=\lambda w.$$

Then we calculate the k-th component t_k as a linear combination of features of the matrix X

$$t_k=X_{k-1}w_k.$$

After extraction we delete the extracted component from the matrix X and Y using the so called "deflation technique". The new matrix X and Y we can write as follow:

$$X_{k+1}=X_k-t_kp_k^T,\qquad Y_{k+1}=Y_k-t_kc_k^T,$$

where

$$p_k = (X^T t_k)\backslash(t_k^T t_k), \qquad c_k = (Y_{k-1}^T t_k)\backslash(t_k^T t_k).$$

From the above we infer that eigenproblem for extraction component t_k has formula:

$$X_{k-1}X_{k-1}^T\,(D-(1-\gamma)I)\,X_{k-1}w_k = \lambda X_{k-1}w_k,$$

what is equivalent to

$$X_{k-1}X_{k-1}^T\,(D-(1-\gamma)I)\,t_k = \lambda t_k.$$

This completes the proof.

A proper features extraction for nonlinear separable is difficult and could be inaccurate. Hence, for this problem we designed a nonlinear version of our extraction algorithm. We used the following nonlinear function $\boldsymbol{\Phi} : x_i \in \mathbb{R}^N \to \boldsymbol{\Phi}(x_i) \in F$ which transforms the input vector into a new, higher dimensional feature space F. Our aim is to find an EWCDSM component in F. In F, vectors w_k and t_k are given by the following formulas:

$$w_k = (D-(1-\gamma)I)\mathbf{K}_k w_k \tag{8.28}$$

$$t_k = \mathbf{K}_k w_k \tag{8.29}$$

where $\mathbf{K}$ is the kernel matrix.

Theorem 8.3 *In the dual feature space F, vectors w_k and t_k are given by the following formulas:*

$$w_k = w_k^\star = (D-(1-\gamma)I)\boldsymbol{K}_k w_k^\star \tag{8.30}$$

$$t_k = \boldsymbol{K}_k w_k^\star \tag{8.31}$$

where $\boldsymbol{K}_k = \boldsymbol{\Phi}_k\boldsymbol{\Phi}_k^T$.

Proof Because the vector w_i lies in the primary space we must transform them into the dual space. Like for kernel PLS method [3] we use formula:

$$\beta_k w_k = \boldsymbol{\Phi}_k^T w_k^\star, \tag{8.32}$$

where w_k is a weight vector in primary space, whereas vector $w_k^\star$ is weight vector in the dual space. For the nonlinear version of extraction algorithm we implement the deflation of the matrix D. Let D_k denote the matrix D in $k+1$ step. Using (8.32) into $w_k = X^T(D-(1-\gamma)I)Xw_{k-1}$ we get the recursive form of the vector $w_k^\star$:

$$w_{k,l}^\star = \left(D_k-(1-\gamma)I\right)\boldsymbol{\Phi}_k\boldsymbol{\Phi}_k^T w_{k,l-1}^\star. \tag{8.33}$$

Next we normalize the vector $w^{\star}_{k,l}$ i.e.

$$w^{\star}_{k,l} = (w^{\star}_{k,l})\backslash(\|w^{\star}_{k,l}\|). \tag{8.34}$$

We repeat this procedure for the vector $w^{\star}_{k,l}$ until the convergence. Finally, we get that vector $w^{\star}_{k,l}$ correspond to the vector $w^{\star}$ in the dual space, which is the rescaled version of the vector w_k. Therefore in the dual space the components t_k are given by the formula:

$$t_k = \beta_k \boldsymbol{\Phi}_k \boldsymbol{\Phi}_k^T w^{\star}_k. \tag{8.35}$$

Let $\boldsymbol{K}_k = \boldsymbol{\Phi}_k \boldsymbol{\Phi}_k^T$. We have:

$$w^{\star}_k = (D - (1-\gamma)I)\boldsymbol{K}_k w^{\star}_k \tag{8.36}$$

$$t_k = \boldsymbol{K}_k w^{\star}_k. \tag{8.37}$$

This completes the proof.

We called the matrix $\boldsymbol{K}$ the kernel matrix. From the definition of the kernel matrix $\boldsymbol{K}$ we have that:

$$\boldsymbol{\Phi}\boldsymbol{\Phi}^T_{(k+1)} = \left(\boldsymbol{\Phi}_k - tp^T\right)\left(\boldsymbol{\Phi}_k - tp^T\right)^T = \boldsymbol{K}_{k+1} = \left(I - t_k t_k^T\right)\boldsymbol{K}_k\left(I - t_k t_k^T\right)$$

Theorem 8.4 *In the dual space F the extracted vector w_k corresponds to the eigenvector connected with the largest eigenvalue for eigenproblem:*

$$\left(D_k - (1-\gamma)I\right)\boldsymbol{\Phi}_k \boldsymbol{\Phi}_k^T w^{\star}_k = \lambda w^{\star}_k. \tag{8.38}$$

Also, the k-th component t_k corresponds to the eigenvector connected with largest eigenvalue for the following eigenproblem:

$$\boldsymbol{K}_{k-1}(D_{k-1} - (1-\gamma)I)t = \lambda t, \tag{8.39}$$

8.2.2 Classification Using PLS Method

Let us assume that X_{train} and X_{test} are the realizations of the matrix X for train and test datasets respectively. The idea of a training step is to extract vectors of weights w_k and components t_k by using the train matrix X_{train} and to store them as a columns in matrices W and T respectively. To classify samples into classes, we use train matrix X_{train} to compute the regression coefficients by using the least squares method [10] given by:

$$Q = W\left(P^T W\right)^{-1} U^T, \tag{8.40}$$

where,

$$U = YY^T T\left(T^T T\right)^{-1}, \tag{8.41}$$

$$W = X^T U, \tag{8.42}$$

$$P = X^T T\left(T^T T\right)^{-1}. \tag{8.43}$$

We then multiply test matrix X_{test} by the coefficients of the matrix Q. To classify samples corresponding to the Y_{test} matrix, we use the decision rule:

$$y_i = \arg\max_{j=1,\ldots,L} Y_{test}(i,j). \tag{8.44}$$

The final form of the response matrix is the following:

$$Y_{test} = \left[y_1 y_2 \cdots y_L\right]^T. \tag{8.45}$$

Like for the linear version of the algorithm, if want make a prediction, first we must compute the regression coefficient using the formula (8.46)

$$Q = \boldsymbol{\Phi}^T U(T^T \boldsymbol{K} U)^{-1} T^T Y, \tag{8.46}$$

where T is matrix of the components and matrix U has the following form

$$U = YY^T C. \tag{8.47}$$

We make a prediction by multiplying the test matrix data $\boldsymbol{\Phi}_{test}$ by matrix Q, i.e.

$$\hat{Y} = \boldsymbol{\Phi}_{test} Q, \tag{8.48}$$

and then by using the decision rule

$$y_i = \arg\max_{j=1,\ldots,L} \hat{Y}(i,j). \tag{8.49}$$

Finally, the response matrix has the following form

$$Y_{test} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_L \end{bmatrix} \tag{8.50}$$

Like in classic kernel PLS algorithm, if we want make a prediction for the data from test dataset, we use the following formula.

$$\hat{Y} = \mathbf{K}U\left(T^T\mathbf{K}U\right)^{-1}T^TY = TT^TY, \tag{8.51}$$

and the decision rule has the following formula

$$y_i = \underset{j=1,\ldots,L}{\arg\max}[TT^TY](i,j). \tag{8.52}$$

Finally, the response matrix is given by

$$Y_{test} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_L \end{bmatrix}. \tag{8.53}$$

8.2.3 Gaussian Mixture Model

Gaussian mixture model (GMM) (see [5, 6]) is a kind of mixture density model, which assumes that each component of the probabilistic model is a Gaussian density component, i.e., given by formula

$$p(x|\theta_k) = \frac{1}{\sqrt{2\pi}^p\,|\Sigma_k|}exp{-\frac{1}{2}}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k) \tag{8.54}$$

where $\theta_k = (\mu_k, \Sigma_k)$ are the parameters of the Gaussian distribution, including the mean μ_k and positive defined covariance matrix Σ_k. Hence, the Gaussian Mixture Model is the probability density on $\mathbb{R}^p$ given by formula

$$p(x|\theta) = \sum_{k=1}^{M} p(x|\theta_k)p(k) \tag{8.55}$$

where $\theta = (\mu_1, \Sigma_1, \mu_2, \Sigma_2, \ldots, \mu_M, \Sigma_M,)$ is the vector of the model parameters, $p(k)$ represents a priori probabilities, which sum to one. In GMM method, we assume that the covariance matrices are diagonal; hence, the GMM is specified by $(2p + 1)M$ parameters. The parameters are learned form the training dataset by classical Expectation-Maximization (EM) algorithm (see [12]). With Gaussian components, we have two steps in one iteration of the EM algorithm. E-step is the first step in which we re-estimate the expectation based on the previous iteration

$$p(k|x) = \frac{p(k)p(x|k)}{\sum_{i=1}^{M} p(k)p(x|k)} \tag{8.56}$$

$$p(k)^{new} = \frac{1}{n}\sum_{i=1}^{n} P(k|x). \tag{8.57}$$

The second step is so-called M-step in which we update the model parameters to maximize the log-likelihood

$$\mu_i = \frac{\sum_{j=1}^{n} p(k|x_j)x_j}{\sum_{j=1}^{n} p(k|x_j)} \tag{8.58}$$

$$\Sigma_i = \frac{\sum_{j=1}^{n} p(k|x_j)(x_j - \mu_i)(x_j - \mu_i)^T}{\sum_{j=1}^{n} p(k|x_j)}. \tag{8.59}$$

The initial values of μ_i are randomly chosen from a normal distribution with the mean $\mu_0 = \frac{1}{n}\sum_{i=1}^{n} x_i$ and the covariance $\Sigma_0 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_0)(x_i - \mu_0)^T$. Using the Bayes rule, it is possible to obtain a posteriori probability $\pi_{i,k}$ of x belonging to cluster k by the following formula

$$\pi_{i,k} = \frac{p(x|k)p(k)}{p(x)} \tag{8.60}$$

where $p(x|k)$ is the conditional probability of x given the cluster k. It means that GMM is a linear combination of Gaussian density functions. The GMM clustering is fast and provides posterior membership probabilities.

8.2.4 Support Vector Machine (SVM)

Support Vector Machines (SVM) are relatively young and highly effective classification methods. This method was proposed by V. Vapnik [13, 14]. In the basic form it is linear classifier and is designed for two class problems. Let the training dataset will be given by the following formula

$$X_{train} = \left\{ \left(x_1, y_1\right), ..., \left(x_N, y_N\right) \in R^p \times \{-1, +1\} \right\}$$

Moreover if $g(x) \geq 0$, then the classified sample belong to class denoted as a 1 and if $g(x) < 0$ the classified sample belongs to class -1. In the training step of the SVM classifier we find a hyperplane given by $w^T x + w_0 = 0$, where x is the classified sample, w is weight vector and b is the bias. In the linear separable case we can find such a hyperplane that all samples will be correctly divided between classes. An optimal (in the sense of SVM) hyperplane is such a hyperplane for which the distance

between the hyperplane and the nearest sample form the training dataset is maximal. For the discriminant function $g(x) = w^T x + w_0$ the distance r between sample x and the hyperplane is given by $r = \frac{g(x)}{\|w\|}$. By a support vectors SV we mean the vectors for which $\alpha_i > 0$ where $\alpha = [\alpha_1, \ldots, \alpha_{n_{train}}]$ are Lagrange's multipliers. From the sets of optimal hyperplanes we choose the canonical hyperplane it means hyperplane for which $|w^T x + w_0| = 1$. For the training SVM classifier we optimize the following function

$$J(w, w_0) = \frac{1}{2}\|w\|^2 \tag{8.61}$$

under conditions

$$y_i(w^t x_i + w_0) \geq 1 \quad \text{for } i = 1, \ldots, N, \tag{8.62}$$

where N is the number of samples in the training dataset. This problem can be solved using the Langrange's multiplier method to optimize the following function

$$Q(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j x_i^T x_j \tag{8.63}$$

under the conditions $\alpha = \left[\alpha_1, \ldots, \alpha_{n_{train}}\right]$ and

$$\sum_{i=1}^{N} \alpha_i y_i = 0, \quad 0 \leq \alpha_i \leq C. \tag{8.64}$$

When we find optimal values of the α_i, then the weights are given by formula

$$w = \sum_{i=1}^{N} \alpha_i y_i x_i \tag{8.65}$$

and the bias $w_0 = 1 - w^T x^{(s)}$ where $x^{(s)}$ is the support vector. If the dataset is not linear separable we can transform input dataset into the higher dimensional space, where that dataset will be linear separable. Let $\boldsymbol{\Phi} : \mathbb{R}^p \to \mathbb{R}^d$ be some nonlinear function. Then the function $\boldsymbol{\Phi}$ has $\boldsymbol{\Phi}$-separable property if for each sample from the training dataset the following condition

$$w^T \boldsymbol{\Phi}(x_i) > 0 \quad for \quad y_i = 1, \tag{8.66}$$

$$w^T \boldsymbol{\Phi}(x) < 0 \quad for \quad y_i = -1. \tag{8.67}$$

holds.

Using Cover theorem (see [15]), when $d \to \infty$ than the probability $P(n, d) \to 1$. Main problem here is to find such a nonlinear function $\boldsymbol{\Phi}$ for which samples from

classes C_1 and C_2 be linearly separable in new d-dimensional space. In practice, using so-call kernel function there is no need to know how exactly this function is defined. The kernel function is a basic parameter of SVM classifier. The most commonly used kernel functions are: polynomial and Gaussian kernel functions. Dual optimization problem for the problem of finding optimal hyperplane can be defined similar as problem (8.63). This can be solved by Lagrange's multiplier method. Than we get the following optimization problem

$$Q(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \tag{8.68}$$

under conditions $\alpha = [\alpha_1 \ldots \alpha_{n_{train}}]$ and

$$\sum_{i=1}^{N} \alpha_i y_i = 0 \quad \text{dla } 0 \leq \alpha_i \leq C. \tag{8.69}$$

The weight vector describing the optimal hyperplane is given by formula

$$w = \sum_{i=1}^{N} \alpha_i y_i \boldsymbol{\Phi}(x_i), \tag{8.70}$$

Discriminant function of nonlinear SVM classifier is given by formula:

$$g(x) = w^T \boldsymbol{\Phi}(x) + w_0 = \sum_{i=1}^{N} \alpha_i y_i K(x_i, x_j) + w_0. \tag{8.71}$$

The bias w_0 can be calculated using the Kuhn-Tucker condition i.e. using the following formula

$$w_0 = \frac{1}{n_{sv}} \sum_{j=1}^{N} \left[y_j - \sum_{i=1}^{N} \alpha_i y_i K(x_i, x_j) \right], \tag{8.72}$$

where n_{sv} is the number of support vectors.

8.2.5 Proposed Semi-supervised Modified PLS Method

Like in [16], in this paper, we propose using a Gaussian Mixture Model (GMM) to perform the clustering, which is fast and provides posterior probabilities that typically lead to smoother kernels (see [16, 17]). The proposed cluster kernel will be the

combination of a kernel computed from labeled data and a kernel computed from clustering unlabeled data (using GMM), resulting in following algorithm:

1. Compute the kernel for labeled data using the following formula

$$\mathbf{K}_s(x_i, x_j) = \boldsymbol{\Phi}(x_i)^T \boldsymbol{\Phi}(x_j) \tag{8.73}$$

2. Run the GMM algorithm n times with different initial values and number of clusters This results in $q \cdot t$ cluster assignments where each sample has its corresponding posterior probability vector $\boldsymbol{\pi}_i \in \mathbb{R}^m$, where m is the number of clusters.
3. Compute the kernel for all (labeled and unlabeled) data. The kernel is the mean of inner products maximum posterior probabilities $\boldsymbol{\pi}_i$ and $\boldsymbol{\pi}_j$. The kernel is given by following formula:

$$\mathbf{K}_u(x_i, x_j) = \frac{1}{N} \sum_{k=1}^{t} \sum_{l=1}^{q} \boldsymbol{\pi}_i^T \boldsymbol{\pi}_j \tag{8.74}$$

   where m is the number of clusters, N is normalization factor.
4. Compute the final kernel using the following formula

$$\mathbf{K}(x_i, x_j) = \delta \mathbf{K}_s(x_i, x_j) + (1 - \delta)\mathbf{K}_u(x_i, x_j) \tag{8.75}$$

   where $\delta \in [0, 1]$ is a scalar parameter tuned during validation.
5. Use the computed kernel into kernel PLS method.

Because the kernel in (8.74) corresponds to a summation of inner products in $t \cdot q$-dimensional spaces, the above kernel in (8.74) is a valid kernel. Additionally the summation of (8.75) leads also to valid Mercer's kernels.

8.3 Experiments

We applied the new extraction method to commonly accessible economical datasets: *Australian Credit Approval* and *German Credit Data*.We compared our method with PLS based on the Australian Credit Approval available at [18]. The Australian Credit Approval was introduced in papers [19, 20]. This dataset contains information from credit card application form divided into two classes denoted as 0 and 1. Class 1 contains information about people who receive positive decision regarding credit card application. Class 0 contains information about people who receive negative decision regarding credit card application. This dataset contains 690 samples, where 307 samples are those taken from class 0. The remaining 383 samples belong to class 1. Each sample is represented by 14 features. The second dataset, German Credit Data available at [18] contained 1000 samples divided into two classes: class 0 and

Table 8.1 Classification performance (percent) of economic datasets

	SS Kernel EWCDSM	PLS	LibSVM	BPN
Australian	95.65	63.91	94.00	83.36
German	94.21	83.78	44.83	77.83

class 1. Each sample is represented by 30 features. Both datasets contained some non-numerical features. In order to apply extraction algorithm to those datasets, the data had to be relabeled. We assigned natural numbers as new values of non-numerical features. To examine the classification performance of proposed method, we used the following experimental scheme. First, we normalized each dataset. For each dataset, we randomly chose 10 % of samples as a labeled data (5 % from each class). To define the $(q \cdot t)$ cluster centers and the posterior probabilities for each of them, we used 200 samples as unlabeled samples per each class in both datasets. In all cases, we tuned the parameter δ from 0 to 1 in 0.05 intervals. When the mixture models were computed, we chose the most probable Gaussian mode and computed the K_c kernel. We used the nonlinear version of EWCDSM with the Gaussian kernel and parameter σ. The result for all datasets are presented in the Table 8.1. We used the jackknife method [21] to find the proper value of parameters δ and σ. Classification performance is computed by dividing the number of samples classified properly by the total number of samples. This rate is know as a standard error rate [21]. The performance of PLS and SS Kernel EWCDSM was compared with LibSVM [22] and Back Propagation Neural Network (BPN63,91).

8.4 Conclusions

We introduced a new kernel version of an algorithm for semi-supervised feature extraction. Our algorithm uses weighted separation criterion to find the weights vector, which allows for the scatter between the classes to be maximal and for the scatter within the class to be minimal. When comparing the new criterion with the other well known ones, it can be seen that the new one can be used in a situation where the number of samples is small and the costs of computation are lowered. The new extraction algorithm can distinguish between between high-risk and low-risk samples for two different economical datasets. Moreover, we have shown that our method had significantly higher classification performance compared to classical the PLS method. The presented method performs well in solving classification problems. However, to draw some more general conclusions, further experiments should be conducted using other datasets.

References

1. Kramer MA (1991) Nonlinear principal component analysis using autoassociative neural networks. AIChE J 37(2):233–243
2. Roweis ST, Saul LK (2000) Nonlinear dimensionality reduction by locally linear embedding. Science 290(5500):2323–2326
3. Shawe-Taylor J, Cristianini N (2004) Kernel methods for pattern analysis. Cambridge University Press (2004)
4. Blaszczyk P (2015) Semi-supervised feature extraction method using partial least squares and gaussian mixture model. Lecture notes in engineering and computer science: proceedings of the world congress on engineering and computer science 2015, WCECS 2015, 21–23 October, 2015, San Francisco, USA, pp 802–806
5. Chatfield K, Lempitsky VS, Vedaldi A, Zisserman A (2011) The devil is in the details: an evaluation of recent feature encoding methods. BMVC Springer 2(4):8
6. Wang J, Lee J, Zhang C (2003) Kernel GMM and its and its application to image binarization. In: Proceedings. 2003 International conference on multimedia and expo, 2003. ICME'03, vol 1. IEEE
7. Garthwaite PH (1994) An interpretation of partial least squares. J Am Stat Assoc 89:122
8. Höskuldsson A (1988) A PLS regression methods. J Chemometr 2:211–228
9. Wold H (1975) Soft modeling by latent variables: the non-linear iterative partial least squares (NIPALS) approach. In: Perspectives in probability and statistics. Papers in Honour of M. S. Bartlett, pp 117–142
10. Gren J (1987) Mathematical statistic. PWN, Warsaw (1987) (in polish)
11. Blaszczyk P, Stapor K (2009) A new feature extraction method based on the partial least squares algorithm and its applications. Adv Intell Soft Comput 179–186
12. Roberts SJ (1997) Parametric and non-parametric unsupervised cluster analysis. Pattern Recogn 30(2):261–272
13. Vapnik VN (1998) Statistical learning theory. Wiley, New York
14. Vapnik VN (2000) The nature of statistical learning theory 2nd edn. Springer
15. Ripley BD (1996) Pattern recognition and neural networks. Cambridge University Press, Cambridge
16. Izquierdo-Verdiguier E, Gomez-Chova L, Bruzzone L, Camps-Valls G (2012) Semisupervised nonlinear feature extraction for image classification. In: IEEE workshop on machine learning for signal processing, MLSP'12
17. Gomez-Chova L, Camps-Valls G, Bruzzone L, Calpe-Maravilla J (2010) Mean map kernel methods for semisupervised cloud classification. IEEE Trans Geosci Remote Sens 48(1):207–220
18. UC Irvine Machine Learning Repository. http://archive.ics.uci.edu/ml/
19. Quinlan JR (1987) Simplifying decision trees. Int J Man-Machine Stud 27:221–234
20. Quinlan JR (1993) C4.5: programs for machine learning. Morgan Kaufmann
21. Duda R, Hart P (2000) Pattern classification. Wiley, New York
22. http://www.csie.ntu.edu.tw/~cjlin/libsvm

Chapter 9
2D-DOAE Improvement Using ESPRIT with Doppler Correction

Youssef Fayad, Caiyun Wang and Qunsheng Cao

9.1 Introduction

The 2D-DOAE is a very important process in the TWS radar system [1, 2] because it is the creator of the angular two tracking gate dimensions. The efforts which have been done by researchers through the years in all aspects of the DOAE field resulted in several methods [3–12]. However, the eigen decomposition of all these methods are in complex-value space, which leads to high-computational complexity, especially with high snapshots and sensors. In this work a modified ESPRIT algorithm, which bases firstly on T-ESPRIT corrected with Doppler frequency, and secondly on the combination of this T-ESPRIT with S-ESPRIT using TDOA is introduced to estimate 2D-DOA of a radiated source in order to increase the estimation accuracy with low computational load.

Y. Fayad (✉) · Q. Cao
College of Electronic and Information Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China
e-mail: yousseffyad595@yahoo.com; yousseffyad@nuaa.edu.cn

Q. Cao
e-mail: qunsheng@nuaa.edu.cn

C. Wang
College of Astronautics, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China
e-mail: wangcaiyun@nuaa.edu.cn

S.-I. Ao et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-981-10-2717-8_9

9.2 T-ESPRIT Method for 2D Case

In order to carry out the proposed algorithm, there are K narrow-band signals are considered, and have same $\omega_0 = 2\pi f$, the kth signal is:

$$s_k(t) = E_k e^{j(\omega_0 t + \Psi_k)}, k = 1, 2, \ldots, K \tag{9.1}$$

where, $s_k(t)$ of the kth emitting source at time instant t, Ψ_k are assumed to be random variables. E_k is the Incident electric field can be written as components form. As a general expression, we omit the subscript, then,

$$\vec{E} = E_\theta \hat{e}_\theta + E_\varphi \hat{e}_\varphi \tag{9.2}$$

Defining the regions of the auxiliary polarization and polarization phase difference angles $\gamma \in [0, \pi/2]$, $\eta \in [-\pi, \pi]$ respectively, and E_θ, E_φ are the vertical and the horizontal components of E respectively. Then,

$$E_\varphi = |\vec{E}| \cos\gamma, E_\theta = \left|\vec{E}\right| \sin\gamma e^{j\eta} \tag{9.3}$$

The incident field can be also expressed in Cartesian coordinate system,

$$\begin{aligned} \vec{E} = E_\theta \hat{e}_\theta + E_\varphi \hat{e}_\varphi = &\left(E_\theta \cos\theta \cos\varphi - E_\varphi \sin\varphi\right)\hat{e}_x \\ &+ \left(E_\theta \cos\theta \sin\varphi + E_\varphi \cos\varphi\right)\hat{e}_y + (E_\theta \sin\theta)\hat{e}_z \end{aligned} \tag{9.4}$$

For a receiving planar antenna array oriented in *xoy* plane; has elements indexed *L*, *I* along *y* and *x* directions, respectively, and for any pairs (*i*, *l*), its coordinate is $(x, y) = ((i-1)\Delta_x, (l-1)\Delta_y)$ [1], where *i = 1,..., I, l = 1,..., L,* Δ_x and Δ_y are reference displacements between neighbor elements along *x* and *y* axis. The space phase factors are expressed as,

$$p_i(\theta_k, \varphi_k) \equiv p_i^k = e^{j\frac{2\pi(i-1)\Delta_x}{\lambda}\sin\theta_k \cos\varphi_k} \tag{9.5}$$

$$q_l(\theta_k, \varphi_k) \equiv q_l^k = e^{j\frac{2\pi(l-1)\Delta_y}{\lambda}\sin\theta_k \sin\varphi_k} \tag{9.6}$$

where, (θ_k, φ_k) are denoted for the kth source. The measurement vector is,

$$z_{i,l}(t) = \sum_{k=1}^{K} u_k s_k(t) p_i(\theta_k, \varphi_k) q_l(\theta_k, \varphi_k) + w_{i,l}(t) \tag{9.7}$$

$$[Z(t)] = [z_{1,1}(t) \cdots z_{1,L}(t) \cdots z_{I,1}(t) \cdots z_{I,L}(t)]^T \tag{9.8}$$

$$u_k = \begin{pmatrix} \sin\gamma_k \cos\theta_k \cos\varphi_k e^{j\eta_k} - \cos\gamma_k \sin\varphi_k \\ \sin\gamma_k \cos\theta_k \sin\varphi_k e^{j\eta_k} + \cos\gamma_k \cos\varphi_k \end{pmatrix} \tag{9.9}$$

The whole receiving factors in subspaces matrix are included in $[a(\theta_k, \varphi_k)]$,

$$a(\theta_k, \varphi_k) \underline{\underline{\text{def}}} p(\theta_k, \varphi_k) \otimes q(\theta_k, \varphi_k) \tag{9.10}$$

where $\otimes$ denotes the Kronker product, so

$$\begin{aligned} A(\theta_k, \varphi_k) = [&u_k^T p_1^k q_1^k u_k^T p_1^k q_2^k \ldots u_k^T p_1^k q_L^k \quad u_k^T p_2^k q_1^k \quad u_k^T p_2^k q_2^k \cdots u_k^T p_2^k q_L^k \\ &\cdots u_k^T p_I^k q_1^k u_k^T p_I^k q_2^k \cdots u_k^T p_I^k q_L^k]^T \end{aligned} \tag{9.11}$$

where a is the sensor element response corresponding to measured signal direction, and A is the steering vectors of the measured signals. The receiving model can be rewritten as,

$$[Z(t)] = [A]S(t) + [W(t)] \tag{9.12}$$

where;

$S(t) \underline{\underline{\text{def}}} [s_1(t) \cdots \quad \cdots s_K(t)]^T$, and, $[A] \underline{\underline{\text{def}}} [A(\theta_1 \varphi_1) \cdots \quad \cdots A(\theta_K, \varphi_K)]$.

Using T-ESPRIT method for TWS radar system with dwell interval T_D. the whole data is divided into M snapshots at each time $T = T_0/M$ second [2].

Where, T_0 is the total signal time after the radar pulse integration. Then it picks up enough data r enclosed by each snapshot m with time period $\tau = \frac{T}{r}$. So, from Eq. (9.12) each receiving signal measurement value through mth subspace is given as,

$$[z^m(\tau)] = [A]s^m(\tau) + [w^m(\tau)] \tag{9.13}$$

The index m runs as $m = 1, 2, \ldots, M$ snapshots. Then; for the mth subspace data matrix can be expressed as,

$$\begin{bmatrix} z_{1,1}^m(0) & \ldots & z_{1,1}^m(\tau_1) \\ \vdots & \ddots & \vdots \\ z_{I,L}^m(0) & \ldots & z_{I,L}^m(\tau_1) \end{bmatrix} \tag{9.14}$$

where, $\tau_1 = (r-1) \times \tau$.

For T-ESPRIT scheme the ESPRIT algorithm is used in an appropriate picked data represented in Eq. (9.14) for each (m) subspace in parallel for the same sampling accuracy thus reducing the calculations load and consequently saving time is achieved.

The ESPRIT algorithm is based on a covariance formulation that is,

$$\hat{R}_{zz} \underline{\underline{\text{def}}} E[Z(\tau)Z^*(\tau)] = A\hat{R}_{ss}A^* + \sigma^2 \sum_w \tag{9.15}$$

$$\hat{R}_{ss} = E[S(\tau)S^*(\tau)] \tag{9.16}$$

The correlation matrix of $\hat{R}_{zz}$ can be done for eigenvalue decomposition,

$$\hat{R}_{zz}\underline{\underline{\text{def}}}\hat{E}_{\text{S}}\Lambda\hat{E}_{\text{S}}^{H}+\sigma^{2}\hat{\text{E}}_{N}\Lambda\hat{\text{E}}_{N}^{H} \tag{9.17}$$

The eigenvectors $\hat{E}_S=[\hat{e}_1,\hat{e}_2,\cdots,\hat{e}_K]$ for larger K eigenvalues spans the signal subspace, the rest $2(I\times L)-K$ smaller eigenvalues $\hat{E}_N=[\hat{e}_{K+1},\cdots,\hat{e}_{2(I\times L)}]$ spans the noise subspace which is orthogonal to the signal subspace. Therefore, there exists a unique nonsingular matrix Q, such that,

$$\hat{E}_S=[A]Q=\left[u_k^T\otimes p_i(\theta_k,\varphi_k)\otimes q_l(\theta_k,\varphi_k)\right]Q \tag{9.18}$$

In Eq. (9.12), A_{P1} and A_{P2} are assumed the first and the last $2L\times(I-1)$ rows of A respectively, they differ by the factor $\Delta p_k=e^{j\frac{2\pi\Delta_x}{\lambda}\sin\theta_k\cos\varphi_k}$ along the x direction. So $A_{P2}=A_{P1}\varnothing_P$, where $\varnothing_P$ is the diagonal matrix with diagonal elements Δp_k. Consequently, $\hat{E}_{\text{P1}}$ and $\hat{E}_{\text{P2}}$ will be the first and the last $2L\times(I-1)$ sub-matrices formed from $\hat{E}_{\text{S}}$. Then the diagonal elements p_k of $\varnothing_P$ are the eigenvalues of the unique matrix $\Psi_P=Q^{-1}\varnothing_P Q$, that satisfies

$$\hat{E}_{\text{P2}}=\hat{E}_{\text{P1}}\Psi_P \tag{9.19}$$

Similarly, the two $2I\times(L-1)$ sub-matrices A_{q1} and A_{q2} consist of the rows of A numbered $2L\times(i-1)+l$ and $2L\times(i-1)+l+2$ respectively, differ by the space factors $\Delta q_k=e^{j\frac{2\pi\Delta_y}{\lambda}\sin\theta_k\sin\varphi_k}$ along the y direction, $l=1,\ \ldots,\ 2(L-1)$. Then $A_{q2}=A_{q1}\varnothing_q$, where $\varnothing_q$ is the diagonal matrix with diagonal elements Δq_k. Consequently, $\hat{E}_{\text{S}}$ forms the $2I\times(L-1)$ two sub-matrices $\hat{E}_{q1}$ and $\hat{E}_{q2}$. Then the diagonal elements Δq_k of $\varnothing_q$ are the eigenvalues of the unique matrix $\Psi_q=Q^{-1}\varnothing_q Q$, that satisfies,

$$\hat{E}_{q2}=\hat{E}_{q1}\Psi_q \tag{9.20}$$

For multiple signals processing the ordering of the eigenvectors of Ψ_q is generally permuted relative to the ordering of the eigenvectors of Ψ_P. It is essential to pair the eigenvalues of them. Thus, for all possible permutations;

$$\left\{k_1^0,k_2^0,\ldots.,k_K^0\right\}=\frac{\Delta p_{k_u}}{\Delta q_{k_v}}-\Delta q_{k_t} \tag{9.21}$$

For, $k_u,k_v,k_t=1,2,\ldots,\text{K}$

Therefore, it is easy to derivate the arrival angles (θ_k,φ_k) as follows:

$$\hat{\theta}_k=sin^{-1}\left\{\frac{\lambda}{2\pi}\left[\left(\frac{arg(\Delta p_k)}{\Delta_x}\right)^2+\left(\frac{arg(\Delta q_k)}{\Delta_y}\right)^2\right]^{1/2}\right\} \tag{9.22}$$

$$\hat{\varphi}_k = tan^{-1}\left[\frac{\Delta_x}{\Delta_y} \cdot \frac{arg(q_k)}{arg(p_k)}\right] \tag{9.23}$$

9.3 Spatial Subspace Method

The S-ESPRIT algorithm is realized by arranging the main planar array as a multiple uniform sub-planar arrays related to the common reference point to get a unified phase shifts measurement point for all sub-arrays. The T-ESPRIT algorithm is applied on each sub-array separately, and in the same time with the others to realize time and space parallel processing. Then, the TDOA [13–15] technique is applied and combined the multiple sub-arrays to calculate the optimum DOAE value at the reference antenna. Figure 9.1 introduces a planar array arranged as sub-planar arrays and indexed with *N, G* along *y* and *x* directions respectively. Figure 9.2 shows a sub-planar array(*n, g*) where $n = 1, \ldots, N$, $g = 1, \ldots, G$. For any pairs (*i, l*), its coordinates with respect to the reference point (R.P) along *y* and *x* directions respectively are $\left(y^{il,n}, x^{il,g}\right)$, where $i = 1,\ldots, I$, $l = 1,\ldots, L$.

The space phase factors along *x* and *y* directions are expressed as:

$$p_i\left(\theta_k^{n,g}, \varphi_k^{n,g}\right) \equiv p_i^k(n,g) = e^{j\frac{2\pi(i-1)\Delta_x}{\lambda}\sin\theta_k^{n,g}\cos\varphi_k^{n,g}} \tag{9.24}$$

$$x^{il,g} = ((i-1)\cdot\Delta_x) + ((g-1)\cdot I\cdot\Delta_x) \tag{9.25}$$

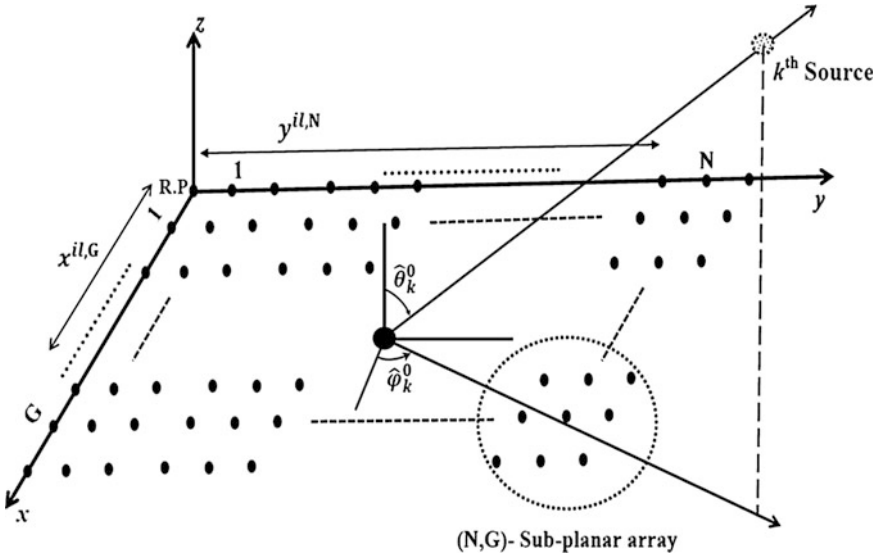


Fig. 9.1 Introduction of planar antenna sub-arrays

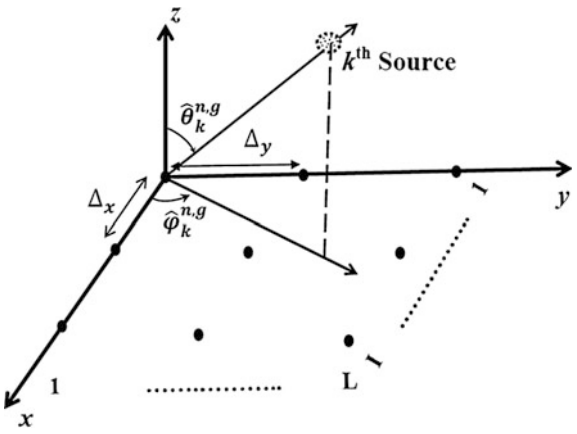


Fig. 9.2 Distribution of sub-Planar antenna array

$$q_l\left(\theta_k^{n,g},\varphi_k^{n,g}\right)\equiv q_l^k(n,g)=e^{j\frac{2\pi(l-1)\Delta_y}{\lambda}\sin\theta_k^{n,g}\sin\varphi_k^{n,g}} \tag{9.26}$$

$$y^{il,n}=\left((l-1)\cdot\Delta_y\right)+\left((n-1)\cdot L\cdot\Delta_y\right) \tag{9.27}$$

where $\left(\theta_k^{n,g},\varphi_k^{n,g}\right)$ denote the kth source estimated angles with respect to $(n,\ g)$ sub-array, Δ_x and Δ_y are reference displacements between neighbor elements along x and y directions within any $(n,\ g)$ sub-array also they are the reference displacements between neighbor sub-arrays. So, for each $(n,\ g)$ sub-array, the receiving model can be expressed as:

$$[Z^{n,g}(t)]=[A^{n,g}][S(t)]+[W(t)] \tag{9.28}$$

$$[A^{n,g}]\underline{\underline{\text{def}}}\left[u_k^T\left(\gamma_k,\eta_k,\theta_k^{n,g},\varphi_k^{n,g}\right)\otimes p_i\left(\theta_k^{n,g},\varphi_k^{n,g}\right)\otimes q_l\left(\theta_k^{n,g},\varphi_k^{n,g}\right)\right] \tag{9.29}$$

The T-ESPRIT method is applied for each $(n,\ g)$ sub-array. Therefore, the arrival angles (θ_k,φ_k) for each sub-array can be calculated as:

$$\hat{\theta}_k=sin^{-1}\left\{\frac{\lambda}{2\pi}\left[\left(\frac{arg(\Delta p_k)}{\Delta_x}\right)^2+\left(\frac{arg(\Delta q_k)}{\Delta_y}\right)^2\right]^{1/2}\right\} \tag{9.30}$$

$$\hat{\varphi}_k=tan^{-1}\left[\frac{\Delta_x}{\Delta_y}\cdot\frac{arg(q_k)}{arg(p_k)}\right] \tag{9.31}$$

For sub-Planar pair shown in Fig. 9.3, the TDOA (δ^{ng}) for the kth received signal is calculated as follows [14]:

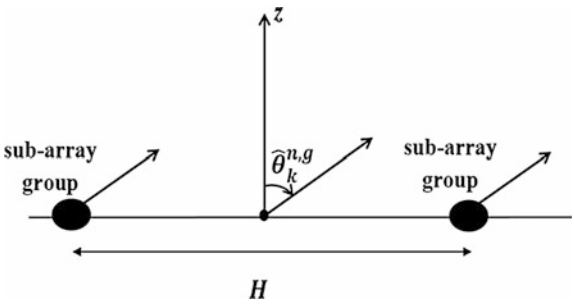


Fig. 9.3 Distribution of a sub-planar pair

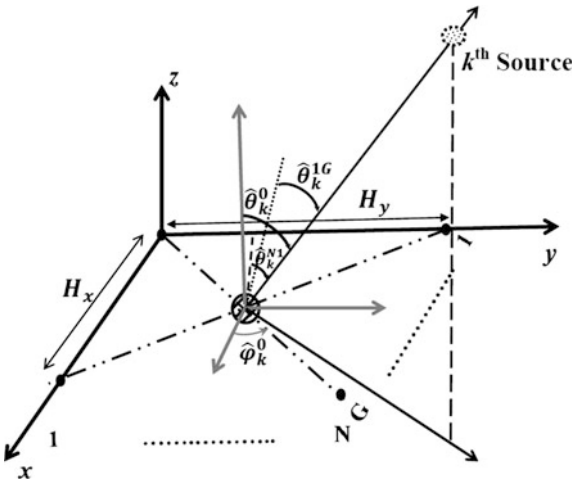


Fig. 9.4 Distribution of sub-planar groups

$$\delta_k^{ng} = \frac{H}{c}\sin\hat{\theta}_k^{ng} \tag{9.32}$$

where H is the distance between two pair (n, g) shown in Fig. 9.3. For the sub-planar groups shown in Fig. 9.4, the TDOA between group (1, 1) and group (N, G) for kth source is denoted by δ_k^{1G}, and the TDOA between group (N, 1) and group (1, G) for kth source is denoted by δ_k^{N1}. So, from Eq. (9),

$$\delta_k^{1G} = \frac{H}{c}\sin\hat{\theta}_k^{1G} \tag{9.33}$$

$$\delta_k^{N1} = \frac{H}{c}\sin\hat{\theta}_k^{N1} \tag{9.34}$$

and,

$$H = \sqrt{(H_x)^2 + (H_y)^2} \tag{9.35}$$

where H_x, H_y are for the first and last sub-array arranged along x and y directions respectively. And,

$$H_x = x_c^{g=G} - x_c^{g=1} \tag{9.36}$$

$$H_y = y_c^{n=N} - y_c^{n=1} \tag{9.37}$$

From Eqs. (9.25) and (9.27), we get,

$$x_c^g = \left(\frac{(I-1)}{2} \cdot \Delta_x\right) + ((g-1) \cdot I \cdot \Delta_x) \tag{9.38}$$

$$y_c^n = \left(\frac{(L-1)}{2} \cdot \Delta_y\right) + \left((n-1) \cdot L \cdot \Delta_y\right) \tag{9.39}$$

and,

$$\hat{\theta}_k^{1G} = \frac{\hat{\theta}_k^{1,1} + \hat{\theta}_k^{N,G}}{2} \tag{9.40}$$

$$\hat{\theta}_k^{N1} = \frac{\hat{\theta}_k^{N,1} + \hat{\theta}_k^{1,G}}{2} \tag{9.41}$$

where the value of θ_k for any (n, g) sub-array group was estimated as in (30). Substitute Eqs. (9.35)–(9.41) into Eqs. (9.33) and (9.34) then the TDOA $\left(\delta_k^{1G}, \delta_k^{N1}\right)$ of the kth received signal are obtained. So, as shown in Fig. 9.4 the optimal DOAE values $(\theta_k^0, \varphi_k^0)$ are measured at the center of the sub-array groups which is located in the cross point of the straight lines pass through group (1, 1) and group (N, G) and through group (N, 1) and group (1, G). Thus, $(\theta_k^0, \varphi_k^0)$ values can be represented as follows [14, 15]:

$$\sin\hat{\theta}_k^{1G} = \sin\hat{\theta}_k^0 \sin\hat{\varphi}_k^0 = \frac{c\delta_k^{1G}}{H} \tag{9.42}$$

$$\sin\hat{\theta}_k^{N1} = \sin\hat{\theta}_k^0 \cos\hat{\varphi}_k^0 = \frac{c\delta_k^{N1}}{H} \tag{9.43}$$

From Eqs. (9.42) to (9.43) the DOAE values are obtained:

$$\hat{\varphi}_k^0 = \tan^{-1}\left(\frac{\delta_k^{1G}}{\delta_k^{N1}}\right) \tag{9.44}$$

$$\hat{\theta}_k^0 = \sin^{-1}\left(\frac{c}{H}\sqrt{\left(\delta_k^{1G}\right)^2 + \left(\delta_k^{N1}\right)^2}\right) \tag{9.45}$$

9.4 Doppler Correction for TS-ESPRIT Method

Target movement causes fluctuation of the measured phase due to wavelength change which degrades the accuracy of DOAE. Method of Zhang [16] had been introduced to reduce the computational complexity, but these works are suitable for stationary targets, the compensation of the target movement within the DOAE values using ESPRIT method had not been introduced. Jinli [17] method had paid attention to the effect of Doppler shift on the DOAE accuracy, but it employed the Fourier transformer to extract each target components from each range cell and then it spliced the separated target components that distributed along all the across range cells which results in large computational load. Additionally, this method has paid attention to target velocity (v_s) but the effect of target maneuver (v_s, α) has not been introduced. In our method the correction is applied according to the fluctuation of the echo signal wavelength change from the transmitted one which induced by the target maneuver (v_s, α), that is,

$$f_d \propto |\vec{v}_s| \cos\alpha \tag{9.46}$$

In order to apply the Doppler correction we should firstly detect and describe the target components. In [18] Doppler frequency had been estimated via using the FFT, but it has a high order computation time for a large number of samples. Other methods apply ESPRIT technique to compute Doppler shift [19, 20]. However, these methods require intensive matrix computations. A new method to estimate the Doppler shift with low computational loads is introduced in this work. In order to estimate f_d using T-DIT method, the T-ESPRIT is applied on a signal (Z_T) generated from a transmitted signal sample and has its phase in order to reduce the computational loads, then its space phase factors along x and y directions are compared with the space phase factors of the received signal along x and y directions respectively. This comparison enables the Doppler frequency shift estimation via measuring the fluctuation wavelength change between the transmitted and received signal. Due to target movement the arrival space phase factors for the co-located planar array at the receiver should be calculated as follows,

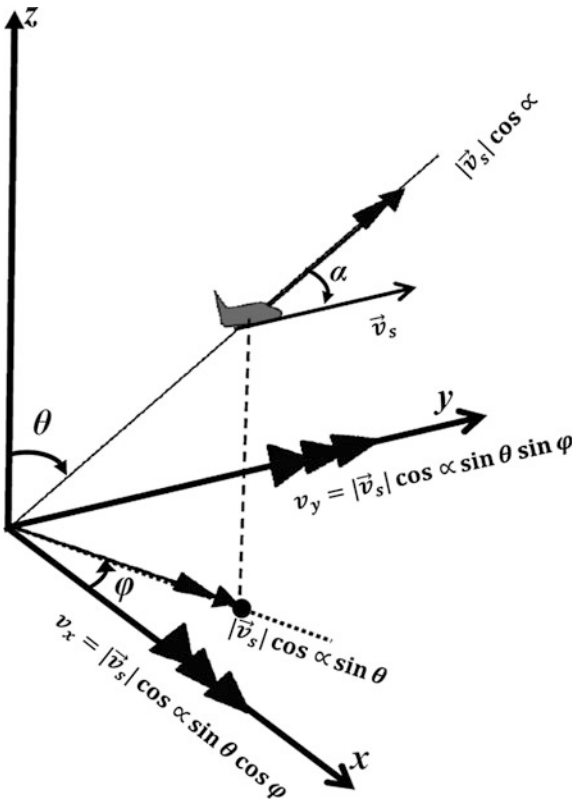


Fig. 9.5 Definition of target linear velocity components in antenna plane

$$arg_A(\Delta p_k) = \frac{2\pi}{\lambda_x}\Delta_x \sin\theta_k \cos\varphi_k \tag{9.47}$$

$$arg_A(\Delta q_k) = \frac{2\pi}{\lambda_y}\Delta_y \sin\theta_k \sin\varphi_k \tag{9.48}$$

where λ_x, λ_y are the fluctuated wavelength components of the received wave into antenna plane and differ from the transmitted wavelength because of f_d, From Fig. 9.5, the wavelength λ_x and λ_y are caused by the velocity components v_x and v_y along x and y directions, as follows,

$$\lambda_x = \frac{\lambda(c + v_x)}{c} \tag{9.49}$$

$$\lambda_y = \frac{\lambda(c + v_y)}{c} \tag{9.50}$$

Substitute into Eqs. (9.47) and (9.48), then,,

$$arg_A(\Delta p_k) = \frac{2\pi c}{\lambda(c+v_x)}\Delta_x \sin\theta_k \cos\varphi_k \tag{9.51}$$

$$arg_A(\Delta q_k) = \frac{2\pi c}{\lambda(c+v_y)}\Delta_y \sin\theta_k \sin\varphi_k \tag{9.52}$$

where,

$$v_x = |\vec{v}_s|\cos\alpha\ \sin\theta_k \cos\varphi_k = \frac{c\cdot f_d}{2f}\sin\theta_k \cos\varphi_k \tag{9.53}$$

$$v_y = |\vec{v}_s|\cos\alpha\ \sin\theta_k \sin\varphi_k = \frac{c\cdot f_d}{2f}\sin\theta_k \sin\varphi_k \tag{9.54}$$

$$|\vec{v}_s| = \frac{c\cdot f_d}{2f\cdot\ \cos\alpha} \tag{9.55}$$

For the transmitted signal, the departure arguments is introduced as,

$$arg_d(\Delta p_k) = \frac{2\pi\Delta_x}{\lambda}\sin\theta_k \cos\varphi_k \tag{9.56}$$

$$arg_d(\Delta q_k) = \frac{2\pi\Delta_y}{\lambda}\sin\theta_k \sin\varphi_k \tag{9.57}$$

It is noted that departure and arrival angles are almost equal for the co-located case.

Substitute Eqs. (9.53), (9.54) into Eqs. (9.51), (9.52) respectively, then

$$arg_A(\Delta p_k) = \frac{2\pi\Delta_x}{\lambda}\cdot\ \sin\theta_k \cos\varphi_k \cdot \frac{2f}{2f+f_d^k\cdot\ \sin\theta_k \cos\varphi_k} \tag{9.58}$$

$$arg_A(\Delta q_k) = \frac{2\pi\Delta_y}{\lambda}\cdot\ \sin\theta_k \sin\varphi_k \cdot \frac{2f}{2f+f_d^k\cdot\ \sin\theta_k \sin\varphi_k} \tag{9.59}$$

And substitute Eqs. (9.56), (9.57) into Eqs. (9.58), (9.59) respectively, we have

$$\frac{arg_d(\Delta p_k)}{arg_A(\Delta p_k)} = \frac{2f+f_d^k\cdot\ \sin\theta_k \cos\varphi_k}{2f} \tag{9.60}$$

$$\frac{arg_d(\Delta q_k)}{arg_A(\Delta q_k)} = \frac{2f+f_d^k\cdot\ \sin\theta_k \sin\varphi_k}{2f} \tag{9.61}$$

Table 9.1 Judgment of target direction

$\sin(\theta_k)$	v_y	$v_y \cdot \sin(\theta_k)$	Target direction
+ve	+ve	+ve	Away from antenna
+ve	−ve	−ve	Close to antenna
−ve	+ve	−ve	Close to antenna
−ve	−ve	+ve	Away from antenna

Thus,

$$f_d^k = \left(\frac{2f}{\sin\theta_k}\right)\left[\left(\frac{arg_d(\Delta p_k)}{arg_A(\Delta p_k)} - 1\right)^2 + \left(\frac{arg_d(\Delta q_k)}{arg_A(\Delta q_k)} - 1\right)^2\right]^{1/2} \tag{9.62}$$

The variable θ_k is the coarse estimated value of the elevation angle stated from (30).

It is clear from Table 9.1 that our main purpose is to determine v_y sign. Substitute Eq. (9.57) into (9.52), then it has follow form,

$$v_y = c \cdot \left(\frac{arg_d(\Delta q_k)}{arg_A(\Delta q_k)} - 1\right) \tag{9.63}$$

So,

$$\begin{cases} arg_A(\Delta q_k) < arg_d(\Delta q_k), & if\ v_y = +ve \\ arg_A(\Delta q_k) > arg_d(\Delta q_k), & if\ v_y = -ve \\ arg_A(\Delta q_k) = arg_d(\Delta q_k), & if\ v_y = 0 \end{cases} \tag{9.64}$$

9.4.1 DOAE Using the Modified TS-ESPRIT Method

The more accurate T-ESPRIT algorithm should consider the effect of the Doppler frequency shift due to the target movement.

From Eqs. (9.58) and (9.59),

$$\begin{aligned} &arg_A(\Delta p_k) + arg_A(\Delta p_k) \cdot \left(f_d^k / 2f\right) \sin\theta_k \cos\varphi_k \\ &\quad = (2\pi\Delta_x/\lambda) \cdot \sin\theta_k \cos\varphi_k \end{aligned} \tag{9.65}$$

$$\begin{aligned} &arg_A(\Delta q_k) + arg_A(\Delta q_k) \cdot \left(f_d^k / 2f\right) \sin\theta_k \sin\varphi_k \\ &\quad = (2\pi\Delta_y/\lambda) \cdot \sin\theta_k \sin\varphi_k \end{aligned} \tag{9.66}$$

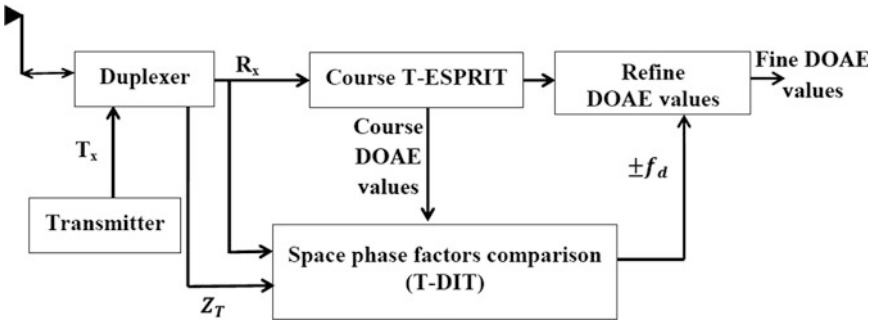


Fig. 9.6 Doppler correction with T-DIT technique

So,

$$arg_A(\Delta p_k) = \sin\theta_k \cos\varphi_k \cdot \left[(2\pi\Delta_x/\lambda) - \left(arg_A(\Delta p_k) \cdot \left(f_d^k/2f\right)\right)\right] \tag{9.67}$$

$$arg_A(\Delta q_k) = \sin\theta_k \sin\varphi_k \cdot \left[(2\pi\Delta_y/\lambda) - \left(arg_A(\Delta q_k) \cdot \left(f_d^k/2f\right)\right)\right] \tag{9.68}$$

Thus,

$$\hat{\theta}_k = sin^{-1}\left\{\left[\left(\frac{arg_A(\Delta p_k)}{(2\pi\Delta_x/\lambda) - \left(arg_A(\Delta p_k) \cdot \left(f_d^k/2f\right)\right)}\right)^2 + \left(\frac{arg_A(\Delta q_k)}{(2\pi\Delta_y/\lambda) - \left((arg_A(\Delta q_k) \cdot \left(f_d^k/2f\right)\right)}\right)^2\right]^{1/2}\right\} \tag{9.69}$$

$$\hat{\varphi}_k = tan^{-1}\left[\frac{arg_A(\mathrm{H}q_k)}{arg_A(\Delta p_k)} \cdot \frac{(2\pi\Delta_x/\lambda) - \left(arg_A(\Delta p_k) \cdot \left(f_d^k/2f\right)\right)}{(2\pi\Delta_y/\lambda) - \left(arg_A(\Delta q_k) \cdot \left(f_d^k/2f\right)\right)}\right] \tag{9.70}$$

Figure 9.6 shows using of T-DIT to refine T-ESPRIT method.

9.5 Simulation Results and Analyses

9.5.1 Spatial Subspace Method Application

Considering 2D-DOAE process with the AWGN, the parameters are given as $T_D = 1$ ms, and $f_s = 25$ MHz. Assuming total 25 temporal snapshots, pickup enclosed data $r = 20$ times, and 200 independent Monte Carlo simulations. In order to validate the proposed TS-ESPRIT method, it has been used in the planar case with number of elements, such as $(I, L) = (3, 3)$, $(N, G) = (2, 2)$ with displacement values $\Delta_x = \Delta_y = \lambda/2$, at initial values of $\theta = 45^o$ and $\varphi = 60^o$. Table 9.2 represents the computational time and complexity of the proposed method in term of number

Tab. 9.2 Comparison of the required computational time and complexity

	Conventional	T-ESPRIT	TS-ESPRIT
Time (ms)	17.7	1.62	1.14
Complexity	$O([4d(IGLN)^2 + 8(IGLN)^3])$	$O(M + 4r(IGLN)^2 + 8(IGLN)^3)$	$O([GNM + 4r(IL)^2 + 8(IL)^3])$

of flip-flops. It is obvious that the computational load has been reduced as a result of employing TS-ESPRIT method to enable a simultaneous processing for M sub-spaces with each has r snapshots for sub-array with (I, L) elements instead of processing for one space has ($d = Mr$) snapshots for array with (IG, LN) elements. From Table 9.2, it has been found that our algorithm requires $O([GNM + 4r(IL)^2 + 8(IL)^3])$ flops [21]. Also, the proposed TS-ESPRIT algorithm requires about 70.4 % of the computational time compared with that of the T- ESPRIT algorithm.

Figure 9.7 is plotted for the RMSEs of the T-ESPRIT algorithm applied on the array with elements (6,6), and the RMSEs of the proposed TS-ESPRIT algorithm for (I, L) = (3, 3), and (N, G) = (2, 2). The proposed algorithm errors are getting closer to the CRB as a result of enabling multiple phase deference measurements with different resolutions.

Figure 9.8 is given the RMSEs of the proposed algorithm (TS-ESPRIT with Doppler correction). The results shown in Fig. 9.8 are indicated that the proposed algorithm errors are getting closer to the CRB as a result of applying subspace concept with Doppler correction. The accuracy improvement of the 2D-DOAE using proposed algorithm has been verified by comparing the resulted RMSEs with the RMSEs of ESPRIT-Like, and the 2D-Beamspace ESPRIT algorithms used in Refs. [11, 22] respectively.

The comparison results displayed in Fig. 9.9 showed that the proposed algorithm has a better performance, especially at a low SNR. This upgrade has been realized due to the increase of DOAE accuracy when combining the T-ESPRIT with S-ESPRIT algorithm which increases the resolution of phase deference

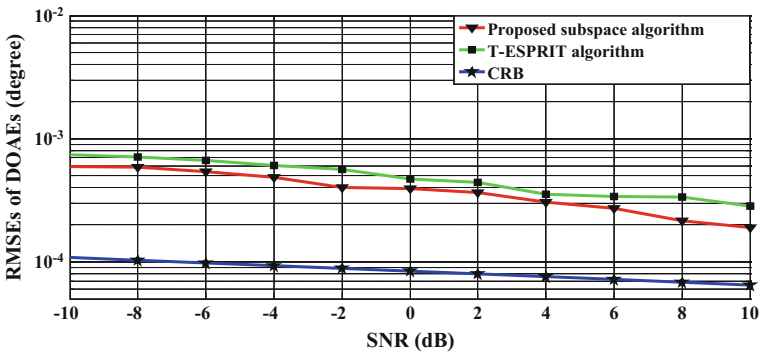


Fig. 9.7 RMSEs versus SNR for T-ESPRIT algorithm and the proposed TS-ESPRIT algorithm

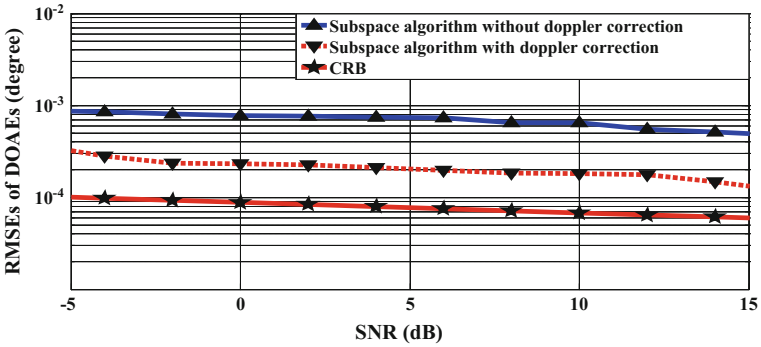


Fig. 9.8 RMSEs versus SNR for the proposed TS-ESPRIT combined algorithm

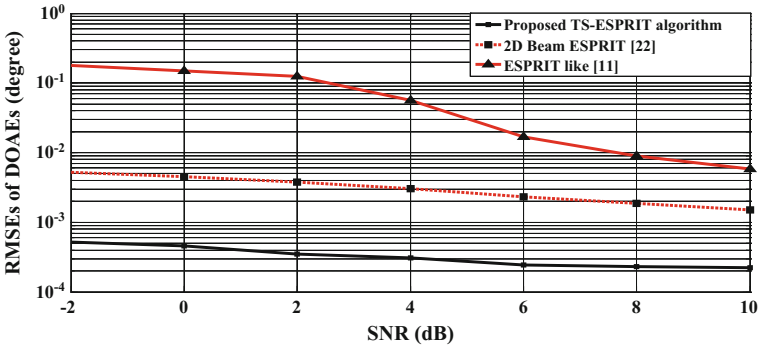


Fig. 9.9 RMSEs versus SNR for the proposed algorithm against different ESPRIT algorithms

measurement. Additionally, this improvement is due to Doppler correction which reduces the DOAE uncertainty associated with effect of the target movement. In order to introduce the improvement in the computation time compared to estimate Doppler frequency via using the FFT method, Table 9.3 is listed a comparison between the proposed method computation time and the different FFT [23]. The results listed in Table 9.3 are indicated the computation time is reduced to 11.09 % compared to the best computation time which has been realized via using PWZFFT method [23]. This improvement has been realized because of using the temporal subspaces concept which enables parallel processing.

Table 9.3 Comparisons of the required computational time

Algorithms	Computational time (s)
Direct calculation	80.3702
FFT method	1.5112
ZFFT method	0.1604
PWZFFT method	0.0835
Proposed T-DIT method	0.00926

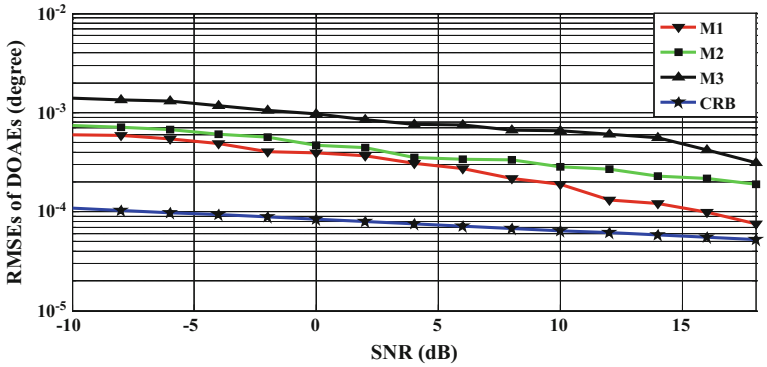


Fig. 9.10 RMSEs versus SNR of T-ESPRIT for different maneuver cases

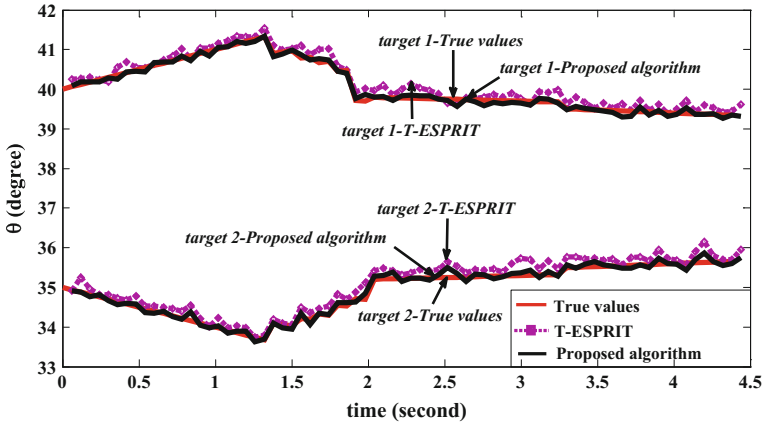


Fig. 9.11 The moving target elevation estimated angles

Figure 9.10 showed the errors of the proposed method for different maneuver cases according to α. For simplicity we pointed to the three cases by M_1, M_2, and M_3 due to different values of α, where M_1 at $\alpha = 0°$, $\alpha = 95°$ for maneuver M_2, and M_3 at $\alpha = 89°$. The displayed results showed that the lowest maneuver (M_1) RMSEs curve is the closest to CRB on the contrary of the highest maneuver (M_3) RMSEs curve. For more validation, a comparison between true and estimated angles of two moving targets is showed at Figs. 9.11 and 9.12.

It is clearly that we find that the proposed algorithm makes the estimated angles more close to the real angles which means reducing of angles errors (up to 0.2°–0.25°) as shown in Figs. 9.11 and 9.12. This reduction in errors improves the radar resolution, which increases the targets separation ability within the TWS radar system (e.g. for radar has an angular resolution 0.75°; this error reduction improves the separation ability about 30 %). Figure 9.13 is illustrated the importance of

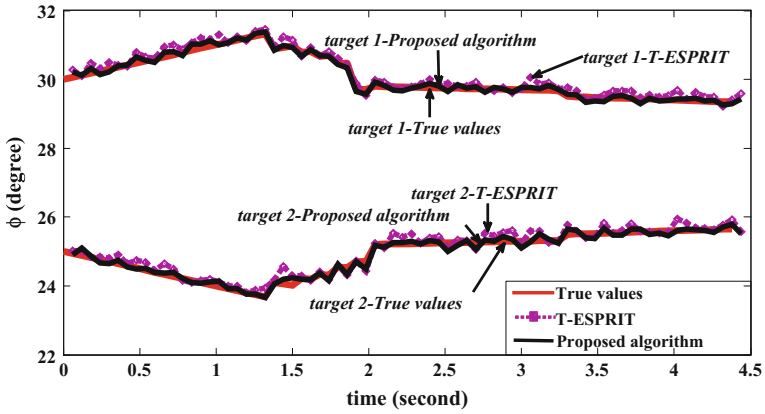


Fig. 9.12 The moving target azimuth estimated angles

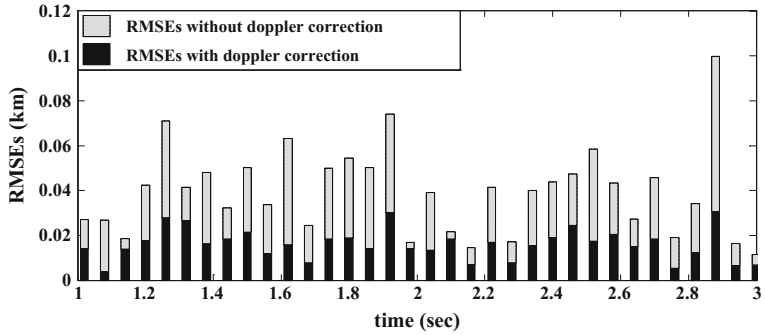


Fig. 9.13 The moving target RMSEs during the flight course

Doppler correction where it is noticed that the RMSEs of the target estimated locations during the flight course without implementing the proposed Doppler correction reach about 100 m, but the implementation of the proposed Doppler correction method reduce it that its maximum value reaches about 30 m which means that the proposed method reduces the errors with about 70 %.

9.6 Conclusions

The problem statement at the source of this work was to upgrade the accuracy but with low computational load which lead to enhance the TWS radar system performance. To address these issues, the algorithm firstly enabled reducing the effects of the non-linear relation between the measured values and the target required

angles; then secondly it enabled multiple phase deference measurements with different resolutions which increased the accuracy of the ESPRIT estimation method; third it curbed the effects of target maneuver on the DOAE accuracy via correcting the DOA estimated values with Doppler frequency which produced from the target maneuver. Finally, this algorithm realized all previous contributions for accuracy improvement with low computational load via combining between temporal and spatial subspace concepts, which enabled parallel processing into both of time and space domains simultaneously. It is clear from the previously mentioned results that the radar angular resolution was increased about 30 %; also it reduced the computational loads and computational time with about 94 % for DOAE and 88.91 % of the computational time of Doppler estimation. Briefly, this approach achieves high DOAE accuracy with low computational load, leading to enhance the TWS radar system performance.

References

1. Fayad Y, Wang C, Cao Q (2015) Doppler correction for spatial subspace technique. In: Lecture notes in engineering and computer science: proceedings of the world congress on engineering and computer science 2015, WCECS 2015, 21–23 Oct 2015, San Francisco, USA, pp 508–513
2. Fayad Y, Wang C, Cao Q (2016) A developed ESPRIT for moving target 2D-DOAE. Eng. Lett. **24**(1), 30–37 (2016)
3. Burg JP (1975) Maximum entropy spectral analysis. In: Proceedings of the 37th meeting society exploration geophysicists, 1967, Stanford Thesis
4. Capon J (1969) High resolution frequency-wavenumber spectrum analysis. Proc IEEE 57 (8):1408–1418
5. Schmidt RO (1979) Multiple emitter location and signal parameter estimation. In: Proceedings of the RADC spectrum estimation workshop, pp 243–258, Griffiths AFB, NY
6. Marcos S, Marsal A, Benidir M (1995) The Propagator method for source bearing estimation. Signal Process 42(2):121–138
7. Roy R, Paulraj A, Kailath T (1986) ESPRIT-estimation of signal parameters via rotational invariance techniques, acoustic, speech, and signal processing. In: IEEE international conference on ICASSP, vol 11, pp. 2495–2498
8. Roy R, Kailath T (1989) ESPRIT-estimation of signal parameters via rotational invariance techniques. IEEE Trans Acoust Speech Signal Process 37(7):984–995
9. Alassery F, Ahmed WKM, Sarraf M, Lawrence V (2014) A low computational complexity statistical discrimination algorithm for collision detection in wireless sensor networks. IAENG Int J Comput Sci 41(3):204–211
10. Yue W, Franco J (2010) A new way to reduce computing in navigation algorithm. Eng Lett 18 (4):341–350
11. Chen F-J, Kwong S, Kok C-W (2010) ESPRIT-like two-dimensional DOA estimation for coherent signals. IEEE Trans Aerosp Electron Syst 46(3):1477–1484
12. Xu X, Ye Z (2012) Two-dimensional direction of arrival estimation by exploiting the symmetric configuration of uniform rectangular array. IET Radar Sonar Navig 6(5):307–313
13. Ho KC, Lu X, Kovavisaruch L (2007) Source localization using TDOA and FDOA measurements in the presence of receiver location errors: analysis and solution. IEEE Trans Signal Process 55(2):684–696

14. Tian Y, Tatematsu A, Tanabe K, Miyajima K (2014) Development of locating system of pulsed electromagnetic interference source based on advanced TDOA estimation method. IEEE Trans Electromagn Compat 56(6):1326–1334
15. Tantisattayakul T, Masugata K, Kitamura I, Kontani K (2005) Broadband VHF sources locating system using arrival-time differences for mapping of lightning discharge process. J Atmos Solar-Terr Phys 67:1031–1039
16. Zhang Y, Ye Z, Xu X, Cui J (2010) Estimation of two-dimensional direction-of-arrival for uncorrelated and coherent signals with low complexity. IET Radar Sonar Navig 4(4):507–519
17. Chen J, Li J, Li Peng, Zhu Y, Long W (2014) Joint DOD and DOA estimation for high speed target using bistatic MIMO radar. Int J Antennas Propag. doi:10.1155/2014/914327
18. Chen VC (2011) The Micro-Doppler effect in radar. Artech House
19. Yunhe C (2010) Joint estimation of angle and Doppler frequency for bistatic MIMO radar. Electron Lett 46(2):124–126
20. Zoltowski MD, Mathews CP (1994) Real-time frequency and 2-D angle estimation with sub nyquist spatio-temporal sampling. IEEE Trans Signal Process 42(10):2781–2794
21. Qian C, Huang L, So HC (2014) Computationally efficient ESPRIT algorithm for direction-of-arrival estimation based on Nyström method. Signal Process 94:74–80
22. Mathews CP, Haardt M, Zoltowski MD (1996) performance analysis of closed-form, ESPRIT based 2-D angle estimator for rectangular arrays. IEEE Signal Process Lett 3(4):124–126
23. Tao R, Zhang W-Q, Chen E-Q (2008) Two-stage method for joint time delay and Doppler shift estimation. IET Radar Sonar Navig 2(1):71–77

Chapter 10
Pathologies Segmentation in Eye Fundus Images Based on Frequency Domain Filters

Gonzalo Urcid, Luis David Lara-R and Elizabeth López-M

10.1 Introduction

Millions of people in North America live with varying degrees of irreversible vision loss because they have an untreatable, degenerative eye disorder, which affects the retina. In these conditions, the delicate layer of tissue that lines the inside back of the eye is damaged, affecting its ability to send light signals to the brain. When the blood vessels are damaged by high blood sugar levels and initially become defective, later they may become blocked off. The defective vessels can lead to hemorrhages (spots of bleeding), fluid and exudates (fats) to escape from the blood vessels over the retina. The blocked vessels can starve the retina from oxygen (ischaemia), leading to the growth of new abnormal vessels in the retina [1]. There are several screening eye exams that help to find any illness, among the exams are the Amsler grid, autofluorescence, dilated eye exam, fundoscopy or ophthalmoscopy, eye fundus photography, fluorescence angiography, optical coherence tomography (OCT), and tonometry [2, 3]. In recent researches, diverse techniques such as the Hough transform, mathematical morphology operators, illumination correction and histogram equalization [4–9] have worked in eye fundus photography to find blood vessels, exudates, and hemorrhages. Other works enhance the image using Gaussian filters as well as the watershed morphological transform [10–12]. In this work, exemplar eye fundus color images have been selected from the databases DIARETDB0 and DIARETDB1 (*Standard Diabetic Retinopathy Database Calibration Level* {0,1}) (Lappeenranta University, Finland). Images in these databases have size of $1,152 \times 1,500$ pixels. Since the area of the optic disk is approximately 2.47 mm^2 (radius = 0.88 mm), the corresponding spatial resolution is about 4.75 μm per pixel [13].

The purpose of this work is to extract exudates, blood vessels, and aneurysms from eye fundus color images. Basically, our proposal consists of the following steps.

G. Urcid (✉) · L.D. Lara-R · E. López-M
Optics Department, INAOE, Luis Enrique Erro No. 1,
72840 Tonantzintla, Puebla, Mexico
e-mail: gurcid@inaoep.mx

S.-I. Ao et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-981-10-2717-8_10

First, a homomorphic high-pass filter in the discrete frequency transform domain is applied to homogenize image illumination, after which a super-Gaussian frequency band-pass filter is defined and applied to distinguish foreground objects of interest from the background. Also, a binary image mask is created to obtain the boundary of the eye fundus in the acquired color image by clipping the effective area taken by the camera. Third, two similar procedures are provided for the different types of pathologies mentioned earlier. Thus, to determine exudates a gamma correction is applied to enhance objects contrast. Otsu's global thresholding method is then applied to binarize the image and a logical operation between a binary mask and the thresholded image gives the segmented exudates. Analogously, for blood vessels and aneurysms, the negative of the filtered image is first determined, then a median filter is applied to reduce noise or artifacts, a gamma correction is applied to enhance objects contrast, and image thresholding (Otsu's method) is performed using global statistics to obtain the desired object regions including their edges. Final segmentation of blood vessels and aneurysms is obtained after morphological closing and by masking the thresholded image.

The chapter is organized as follows: Sect. 10.2 provides the necessary mathematical background on frequency transforms and filters to understand the remaining sections. Section 10.3 explains in sufficient detail the illumination correction step and the frequency filtering approach to edge detection. In Sect. 10.4 a brief description is given of some elementary spatial operations used for image enhancement as well as the final steps for segmenting some pathologies by statistical thresholding and a masking operation. The proposed segmentation algorithm is presented in a mathematical pseudo-code format. Next, in Sect. 10.5 segmentation results of exudates and blood vessels are illustrated and a quantitative discussion of the validity of our technique is provided based on a confusion matrix elaborated from a finite set of eye fundus sample images. We close the paper with Sect. 10.6 of conclusions.

10.2 Mathematical Background

10.2.1 Frequency Transforms

From a physical point of view, various mathematical transforms such as the Fourier, the cosine or sine, and the Hartley transforms can be considered as "frequency transforms". The discrete version of these transforms is simply a finite sequence of data points expressed as a weighted finite sum of complex exponential, sine or cosine, and cas (cosine plus sin) functions oscillating at different frequencies and amplitudes. Each one of these frequency domain transforms is linear and invertible and can be taken to be equivalent to a rectangular matrix of $M \times N$ elements (complex or real). Frequency transforms are important for numerous applications in science and engineering. Table 10.1 summarizes the most common frequency two-dimensional (2D) transforms [14–16], where each row shows its name, acronym and symbol, a

Table 10.1 Two-dimensional forward and inverse discrete frequency transforms

Name and acronym	Symbol	Factor	Kernel functions
Discrete Fourier transform-DFT	$\mathscr{F}$	$\frac{1}{\sqrt{MN}}$	$\exp[-2\pi i(\frac{ux}{M}+\frac{vy}{N})]$
Inverse discrete Fourier transform-IDFT	$\mathscr{F}^{-1}$	$\frac{1}{\sqrt{MN}}$	$\exp[+2\pi i(\frac{ux}{M}+\frac{vy}{N})]$
Discrete Cosine transform-DCT	$\mathscr{C}$	$\frac{\varsigma(u)\varsigma(v)}{\sqrt{MN}}$	$\cos\frac{\pi(2x+1)u}{M}\cos\frac{\pi(2y+1)v}{N}$
Inverse discrete cosine transform-IDCT	$\mathscr{C}^{-1}$	$\frac{1}{\sqrt{MN}}$	$\varsigma(u)\varsigma(v)\cos\frac{\pi(2x+1)u}{M}\cos\frac{\pi(2y+1)v}{N}$
Discrete Sine transform-DST	$\mathscr{S}$	$\frac{2}{\sqrt{M+1}\sqrt{N+1}}$	$\sin\frac{\pi(x+1)(u+1)}{M+1}\sin\frac{\pi(y+1)(v+1)}{N+1}$
Inverse discrete Sine transform-IDST	$\mathscr{S}^{-1}$	$\frac{2}{\sqrt{M+1}\sqrt{N+1}}$	$\sin\frac{\pi(x+1)(u+1)}{M+1}\sin\frac{\pi(y+1)(v+1)}{N+1}$
Discrete Hartley transform-DHT	$\mathscr{H}$	$\frac{1}{\sqrt{MN}}$	$\mathrm{cas}[2\pi(\frac{ux}{M}+\frac{vy}{N})]$
Inverse discrete Hartley transform-IDHT	$\mathscr{H}^{-1}$	$\frac{1}{\sqrt{MN}}$	$\mathrm{cas}[2\pi(\frac{ux}{M}+\frac{vy}{N})]$

scaling factor, and the forward and inverse kernel or basis functions. Note that a scaling factor makes orthogonal the corresponding transform. Other authors assign to the forward or inverse transforms different though similar scaling factors.

In Table 10.1, function "cosine plus sine" is defined by, $\mathrm{cas}\theta = \cos\theta + \sin\theta$, and identical to the expression, $\sqrt{2}\cos(\theta - \pi/4)$. The frequency domain approach considers the design of specific frequency filters in order to perform certain operations such as contrast enhancement or edge detection [17]. Thus, if $\boldsymbol{\Phi}(u, v)$ denotes a frequency filter defined for $u = 0, \ldots, M-1$ and $v = 0, \ldots, N-1$, then the corresponding frequency domain filtering operation is given by $T_g(u, v) = T_f(u, v)\boldsymbol{\Phi}(u, v)$ where T_f and T_g are the transform spectra, respectively, of the input and output images, f and g. Clearly, the equivalent spatial domain filtering operation is given by $g(x, y) = f(x, y) * \varphi(x, y)$, where '$*$' denotes two-dimensional convolution. If $\mathscr{T}$ represents any transform then $T_f(u, v) = \mathscr{T}\{f(x, y)\}$; similarly, $g(x, y) = \mathscr{T}^{-1}\{T_f(u, v)\boldsymbol{\Phi}(u, v)\}$. For numerical computation, the functions g, T_f, and $\boldsymbol{\Phi}$, are matrices of the same size as the given image and forward as well as inverse transforms are computed with a *fast transform* algorithm. In this paper, we focus our attention to the *discrete cosine transform* (DCT) and the *inverse discrete cosine transform* (IDCT) that, written in full are given by,

$$C_f(u, v) = \frac{\varsigma(u)\varsigma(v)}{\sqrt{MN}} \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f(x, y) \cos\frac{\pi(2x+1)u}{2M} \cos\frac{\pi(2y+1)v}{2N}, \tag{10.1}$$

$$f(x, y) = \frac{1}{\sqrt{MN}} \sum_{u=0}^{M-1} \sum_{v=0}^{N-1} \varsigma(u)\varsigma(v) C_f(u, v) \cos\frac{\pi(2x+1)u}{2M} \cos\frac{\pi(2y+1)v}{2N}, \tag{10.2}$$

where, $C_f(u,v) = \mathscr{C}\{f(x,y)\}$, $f(x,y) = \mathscr{C}^{-1}\{C_f(u,v)\}$, $u = 0,\ldots,M-1$, $v = 0,\ldots,N-1$, $\varsigma(u) = \varsigma(v) = 1$ if $u, v = 0$, and $\varsigma(u) = \varsigma(v) = \sqrt{2}$ for $u, v \neq 0$. The ς auxiliary functions refer only to the DCT and IDCT entries in Table 10.1.

10.2.2 Spatial Frequency Filters

Given any frequency transform domain it is possible to modify a 2D spectrum by applying different types of frequency functions such as low-pass, high-pass, band-pass or band-reject filters. Besides the well-known ideal, Gaussian, and Butterworth filters to name a few, here we introduce the class of super-Gaussian frequency filters that generalizes the class of Gaussian filters and simultaneously enhances the frequency response of the class of Butterworth filters. Table 10.2 gives the filter types and class that are relevant to the present work. To distinguish between filters, labels are attached to $\boldsymbol{\Phi}$, e.g., $\boldsymbol{\Phi}_{\mathrm{B}}^{\mathrm{hp}}$ refers to a Butterworth (B) high-pass (hp) filter, and $\boldsymbol{\Phi}_{\mathrm{sG}}^{\mathrm{bp}}$ corresponds to a super-Gaussian (sG) band-pass (bp) filter.

The parameters of a high-pass frequency filter are the filter order (a positive integer), denoted by $n \in \mathbb{Z}^+$ and the cut-off spatial frequency specified by d_0. For a band-pass filter, its bandwidth w is an additional third parameter. In any case, $d(u,v) = \sqrt{u^2+v^2}$, is the Euclidean distance from a point (u,v) to the origin.

We point out that frequency filtering is the key step of the segmentation technique described in this chapter. Super-Gaussian distributions have been reported in studies of light propagation [18] and here, we extend and adapt them to the two-dimensional case in the context of digital image processing. As given in Table 10.2, super-Gaussian filters are an extension of the Gaussian class of filters by incorporating an "order" (n) parameter in a similar way as it is employed in a Butterworth filter. Clearly, if $n = 1$ a super-Gaussian filter is the same as a Gaussian one. Furthermore, the advantages of this novel class of filters is explained next in comparison to the Butterworth filter class. To begin with, Fig. 10.1, displays several super-Gaussian and Butterworth band-pass filter profiles of orders $n = 1, 2, 3$ defined in the DCT domain. For both class filters the parameters values are the same, a cut-off spatial frequency

Table 10.2 Two-dimensional frequency parametric filters

Filter type	Butterworth	super-Gaussian
High-pass	$\boldsymbol{\Phi}_{\mathrm{B}}^{\mathrm{hp}}(u,v\,;n,d_0) = \left(1+\left[\frac{d_0}{d(u,v)}\right]^{2n}\right)^{-1}$	$\boldsymbol{\Phi}_{\mathrm{sG}}^{\mathrm{hp}}(u,v\,;n,d_0) = 1-e^{-\left(\frac{d(u,v)}{\sqrt{2}d_0}\right)^{2n}}$
Band-pass	$\boldsymbol{\Phi}_{\mathrm{B}}^{\mathrm{bp}}(u,v\,;n,w,d_0) = \left(1+\left[\frac{d^2(u,v)-d_0^2}{w\,d(u,v)}\right]^{2n}\right)^{-1}$	$\boldsymbol{\Phi}_{\mathrm{sG}}^{\mathrm{bp}}(u,v\,;n,w,d_0) = e^{-\left(\frac{d^2(u,v)-d_0^2}{w\,d(u,v)}\right)^{2n}}$

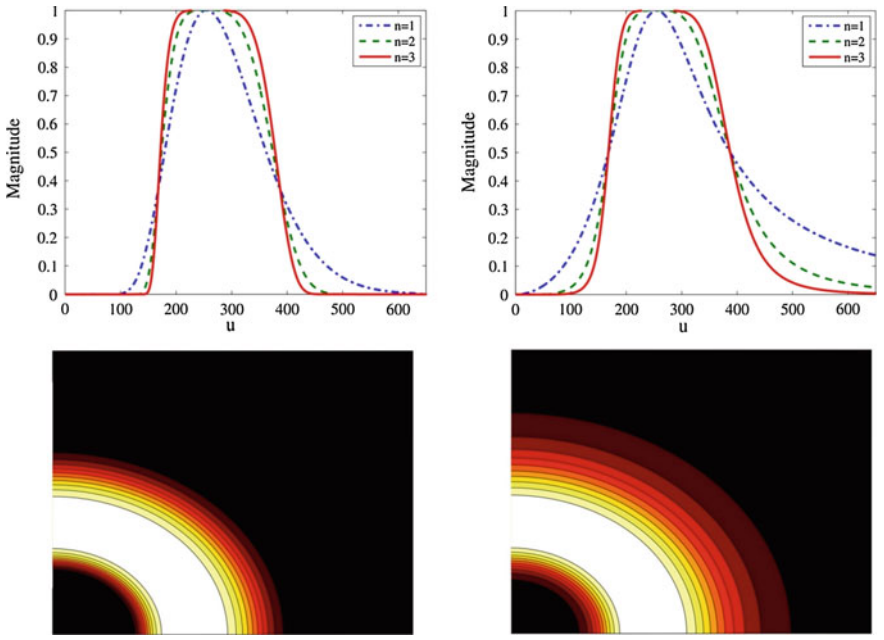


Fig. 10.1 *Left* super-Gaussian band-pass filter profiles and 2D contours ($n = 2$). *Right* Butterworth band-pass filter profiles and 2D contours ($n = 2$)

$d_0 = 256$ and the band-width $w = 220$. Observe that the overall band-width is better delimited in a super-Gaussian band-pass filter than in a Butterworth band-pass filter. Therefore, an sG-profile shape is more sharpened and consequently more selective than a B-profile filter of the same order n.

10.3 Illumination Correction and Edge Detection

Eye fundus color images available in public domain databases are usually stored as acquired by an eye fundus camera. A light source working in the visible range of the electromagnetic spectrum is normally used to illuminate the retina in a spherical wavefront optical system arrangement. Consequently, the retina is heterogeneously illuminated around the eye's optical axis and covering, usually, a 50° field of view. A modified high-pass filter of any type can be used for illumination correction by offsetting and scaling the filter magnitude in order to compensate for illumination variations by enforcing the reflectance component of the input image. Formally, the parametric *homomorphic* high-pass filter is generally defined by,

$$\Psi(u, v\,; n, d_0, \alpha_{\max}, \alpha_{\min}) = (\alpha_{\max} - \alpha_{\min})\Phi^{\mathrm{hp}}(u, v\,; n, d_0) + \alpha_{\min}, \tag{10.3}$$

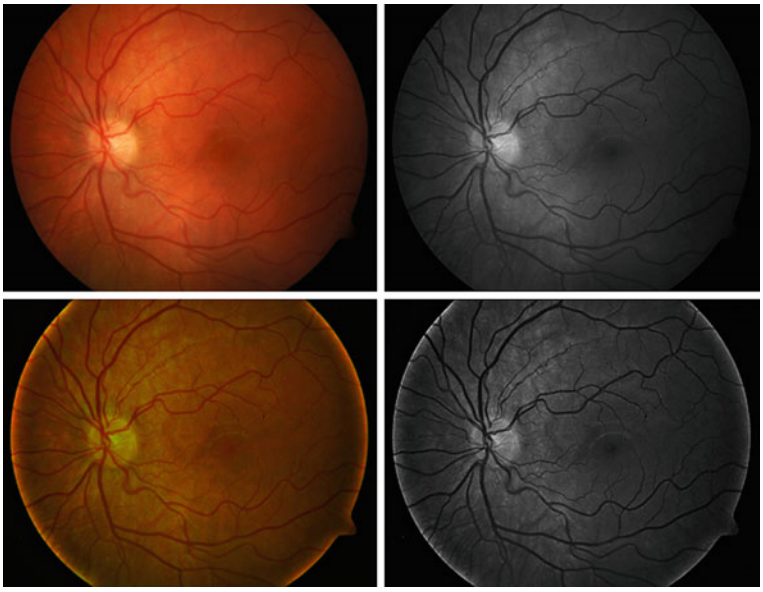


Fig. 10.2 *Top* Eye fundus color image and its *green* channel *grayscale* image. *Bottom* DCT homomorphic filtering illumination corrected eye fundus color image and its *green* channel image

where, $\alpha_{\max} > 1$ and $0 < \alpha_{\min} < 1$, are maximum and minimum magnitude bounds use to change the slope of the high-pass filter $\boldsymbol{\Phi}^{\text{hp}}$. In this work, all the eye fundus color images selected as samples from the DIARETDB0/1 databases were corrected for illumination variations as a pre-processing step. Also, for digital images, the filter parameter values d_0 and w will be given in pixels (except n) since these values can be associated with physical distances. In particular, the parameter values for the DCT homomorphic high-pass filter were set to $\alpha_{\max} = 1.75$, $\alpha_{\min} = 0.75$, $n = 2$, $d_0 = 20$, and $\boldsymbol{\Phi}^{\text{hp}} = \boldsymbol{\Phi}^{\text{hp}}_{\text{sG}}$. An example of an eye fundus illumination corrected image is shown in Fig. 10.2.

Recall that image segmentation makes possible to discriminate objects of interest from their background by dividing the image in regions that satisfy certain conditions [16]. After illumination has been homogenized, DCT band-pass filtering is applied to the eye fundus green channel image to intensify foreground objects against the surrounding background. Thus, edge detection can be controlled by choosing adequate band-pass filter parameters.

From Table 10.2, the proposed *super-Gaussian band-pass* filter for object edge detection in eye fundus illumination corrected images is specified by,

$$\boldsymbol{\Phi}^{\text{bp}}_{\text{sG}}(u, v\,; n, w, d_0) = e^{-\left[\dfrac{d^2(u,v) - d_0^2}{w\, d(u,v)}\right]^{2n}}, \tag{10.4}$$

where, $n = 2$ (filter order), $w = 275$ (band-pass width), $d_0 = 100$ (cutoff spatial frequency), and $d^2(u, v) = u^2 + v^2$. The resulting image after inversion with the IDCT is given by the equivalent spatial expression,

$$g_{\mathrm{sG}}^{\mathrm{bp}}(x, y) = \left[f_{\mathrm{G}}(x, y) * \psi(x, y)\right] * \varphi_{\mathrm{sG}}^{\mathrm{bp}}(x, y), \tag{10.5}$$

where, f_G denotes the green channel image G of an eye fundus color RGB image, ψ is the spatial filter version of the homomorphic frequency high-pass filter Ψ, and similarly, φ is the spatial filter version of the frequency band-pass filter Φ. The filtering stage based on the DCT or another frequency domain transform is the same for segmenting exudates as well as blood vessels or aneurysms which are localized blood filled balls of variable size in the wall of a blood vessel [20]. Details of the specific steps used to segment each type of pathology is described in the following section.

10.4 Segmentation of Eye Fundus Pathologies

Once the frequency filtering step is realized, the post-processing of the edge filtered image is subject to various though simple spatial operations to enhance it. In the case of hard and soft exudates work is done on the filtered image $g = g_{\mathrm{sG}}^{\mathrm{bp}}$. Otherwise, for blood vessels or aneurysms, the negative of the filtered image is required. Specifically, we recall that, $\bar{g}(x, y) = (L - 1) - g(x, y)$, where $L - 1$ is the maximum value of the grascale images dynamic range. In this second case, $L = 256$ and $\bar{g}$ is the intermediate image to be enhanced. First, a 3×3 median kernel is applied to g or $\bar{g}$ to reduce background noise while simultaneously preserving object edges. Second, a gamma correction ($\gamma = 2$) of $\mathrm{med}\{g(x, y)\}$ or $\mathrm{med}\{\bar{g}(x, y)\}$ is performed to emphasize contrast between foreground and background.

The output of the last two steps, denoted by g^{γ} or $\bar{g}^{\gamma}$, is a slightly smoothed and contrast enhanced grayscale spatial image that facilitates the application of an autonomous thresholding technique. Since the normalized histograms of the previously obtained grayscale image is bimodal, Otsu's method based on global and local statistics is a good choice to determine a global threshold value to binarize the image under study [19]. Assuming that the global variance, $\sigma_g^2 > 0$, and that the probability assigned to a foreground pixel $p(t)$ is such that $0 < p(t) < 1$, then the optimal threshold obtained by Otsu's method, denoted by τ, is the numerical value that satisfies the equality,

$$\sigma_b^2(\tau) = \max_{0 \le t \le < L-1} \{\sigma_b^2(t)\}, \tag{10.6}$$

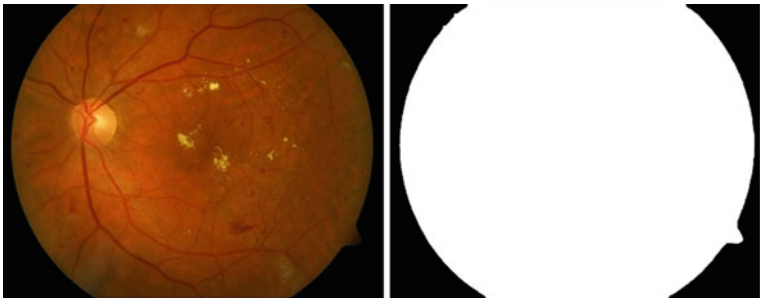


Fig. 10.3 *Left* eye fundus color image. *Right* Eye fundus effective area binary mask

where, $\sigma_b^2(t)$ is the *between-class* variance evaluated at gray level t. In other words, $\tau \in \{0, \ldots, L-1\}$ maximizes $\sigma_b^2(t)$ over the grayscale dynamic range. Hence, the binary output image at this stage is determined as follows,

$$B(x,y) = \begin{cases} 1 & \Leftrightarrow \quad g_{\mathrm{sG}}^{\mathrm{bp}}(x,y) \geq \tau, \\ 0 & \Leftrightarrow \quad g_{\mathrm{sG}}^{\mathrm{bp}}(x,y) < \tau\,. \end{cases} \tag{10.7}$$

The resulting binary image B can be used directly in the case of exudates segmentation. However, in the case of blood vessels, double thin edges or disconnected small regions might appear from the frequency filtering step. Therefore, a *binary morphological closing*, denoted by $B \bullet E$, with an isotropic structuring element E of size 3×3 is applied to join double edges or to fill small gaps or holes. We remind the reader that, $B \bullet E = (B \oplus E) \ominus E$, i.e., a dilation followed by an erosion with the same structuring element.

Before the final step in the segmentation process is accomplished, a binary mask M of the effective area taken by the eye fundus camera is computed (see Fig. 10.3). The final segmented image, containing exudates, blood vessels or aneurysms, is found by masking B with M. That is, $S = (L-1)(B \wedge M)$, where $\wedge$ is the logical 'And' operation. An illustrative example of exudates or blood vessels extraction is shown, respectively, in Figs. 10.4 and 10.5. Algorithm 1 resumes the steps of the frequency transform approach for segmenting certain pathologies in eye fundus color images.

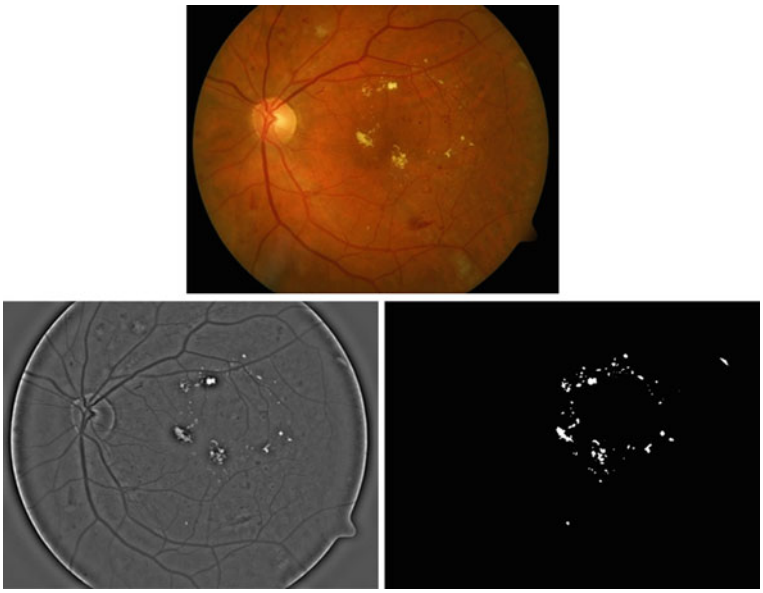


Fig. 10.4 *Top* Eye fundus image with hard and soft exudates (ddb1_v1_1/image015). *Bottom left* super-Gaussian (bp) filtered image. *Bottom right* binary image with segmented exudates

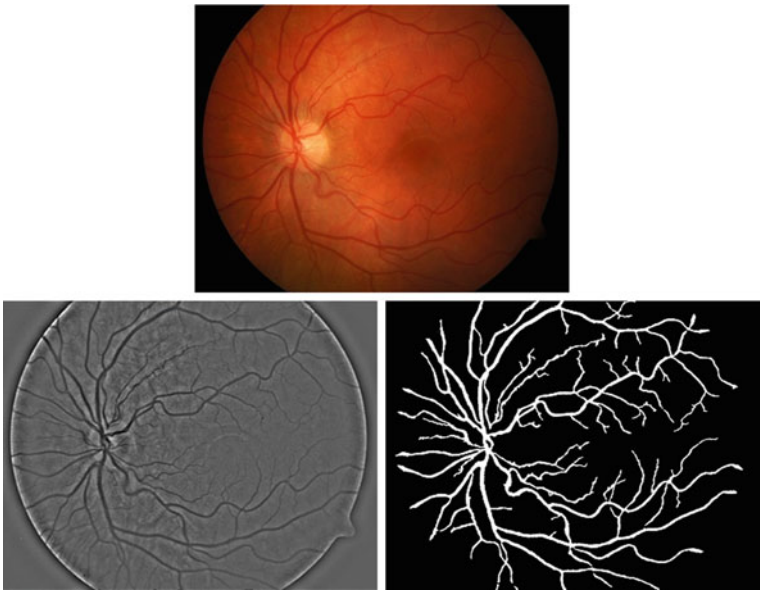


Fig. 10.5 *Top* Eye fundus image with blood vessels (ddb0_v1_1/image064). *Bottom left* super-Gaussian (bp) filtered image. *Bottom right* binary image with segmented blood vessels

Algorithm 1 - Frequency Transform Based Segmentation

Input: f, δ [eye fundus color image, pathology type]
$[\delta = 1 \mid 0$, exudates extraction | blood vessels extraction$]$
$n', d_0', \alpha_{\max}, \alpha_{\min}$ [high-pass filter parameters]
n, d_0, w [band-pass filter parameters]
K, E [3×3 median kernel, 3×3 isotropic structuring element]
Eye fundus mask
$R \leftarrow \text{Channel}(f, R)$ [extract red channel]
$G \leftarrow \text{Channel}(f, G)$ [extract green channel]
$Q \leftarrow R/(G + 1)$ [quotient image]
$Q \leftarrow \text{MedFilter}(Q)$ [remove artifacts]
$\tau \leftarrow \text{OtsuThresh}(Q)$ [find Otsu's optimal threshold]
$M \leftarrow \text{Bin}(Q, \tau)$ [working area image mask]
Illumination correction
$\Psi \leftarrow \Phi_{\text{HP}}(n', d_0', \alpha_{\max}, \alpha_{\min})$ [specify frequency high-pass filter]
$T_G \leftarrow \mathscr{T}(G)$ [green channel spectrum]
$h \leftarrow \mathscr{T}^{-1}(T_G \cdot \Psi)$ [illumination corrected image]
Edge Detection
$\Phi \leftarrow \Phi_{\text{BP}}(n, d_0, w)$ [specify frequency band-pass filter]
$T_h \leftarrow \mathscr{T}(h)$ [corrected image spectrum]
$g \leftarrow \mathscr{T}^{-1}(T_h \cdot \Phi)$ [edge filtered image]
Enhancement and Segmentation
$g \leftarrow \text{MedFilter}(\delta g + (1 - \delta)\bar{g})$ [reduce overall noise]
$g \leftarrow g^{\gamma}$ [gamma enhanced contrast]
$\tau \leftarrow \text{OtsuThresh}(g)$ [find Otsu's optimal threshold]
$B \leftarrow \text{Bin}(g, \tau)$ [background elimination]
$B \leftarrow \delta B + (1 - \delta)\,\text{Closing}(B, E)$ [detected objects of interest]
$S \leftarrow (L - 1)\text{And}(B, M)$ [working area masked image]
Output: S [pathology segmented binary image]

10.5 Pathologies Segmentation Results

To test our proposed segmentation technique, we use six eyes fundus color images taken from the free domain databases DIARETDB0 (ddb0) and DIARETDB1 (ddb1). The first database consists of 130 images, of which 20 are diagnosed as normal and 110 contain signs of *diabetic retinopathy* such as hard exudates, soft exudates, aneurysms, hemorrhages, and neovascularization. The second data bank has 89 images, from which 84 images contain at least mild non-proliferate signs of diabetic retinopathy. The other 5 images are considered normal since these do not contain any pathological signs according to the experts who participated in the evaluation. Eye fundus color images of both databases were captured using the same 50

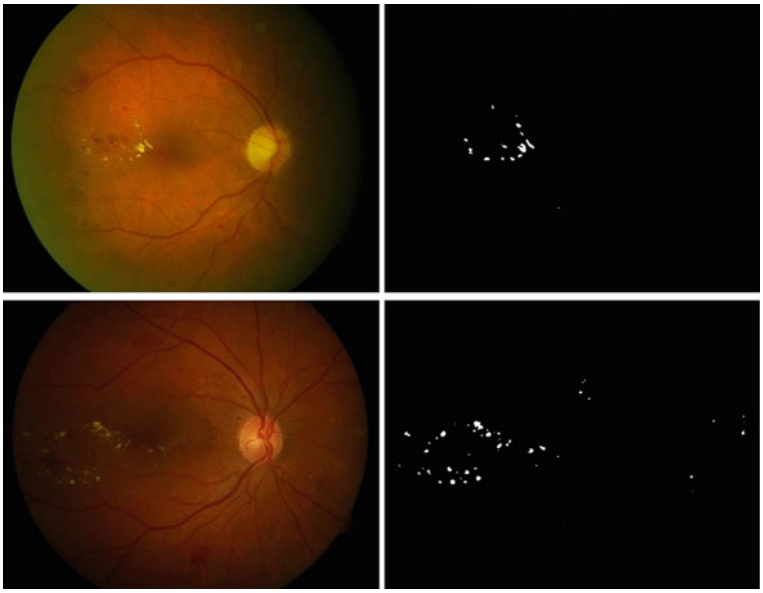


Fig. 10.6 *Left* Eye fundus color images (ddb0_v1_1/image003 & ddb1_v02_01/image064). *Right* Binary images with segmented hard exudates

degree field of view digital fundus camera with varying imaging settings [21]. As mentioned earlier, each image has a size of $1,152 \times 1,500$ pixels, with an approximately spatial resolution of 4.75 micrometers per pixel. For image contrast enhancement after frequency filtering, $\gamma = 2$ for exudates and $\gamma = 3$ for blood vessels. Also, the optimal threshold value obtained by applying Otsu's method is different for each image. Figures 10.6 and 10.7 show, respectively, illustrative examples of exudates and blood vessels segmentations and Table 10.3 lists the umbralization threshold values used for exudates and blood vessels.

In images with soft and hard exudates, the optic disk has similar colors but our technique does not segment it. Soft exudates are nerve fiber layer infarcts or precapillary arterial occlusions. On the other hand, hard exudates represent the accumulation of lipid in or under the retina due to the aqueous portion of the fluid that is absorbed more quickly than the lipid component,and becomes visible as *yellowish* deposits. Examples of both kinds of exudates are illustrated in Figs. 10.4 and 10.6.

In Fig. 10.7, in addition to blood vessels (temporal arcades), aneurysms are also extracted during segmentation and correspond to small amorphous regions not connected to vessels. Aneurysms look like islands and are easily seen where the macula is located.

A *confusion matrix* contains information about actual and predicted classifications over a finite number of classes made by a pattern recognition system [22]. In a medical context, a pattern recognition system can be considered a pre-diagnostic tool constrained to two classes that are interpreted as corresponding to a positive or

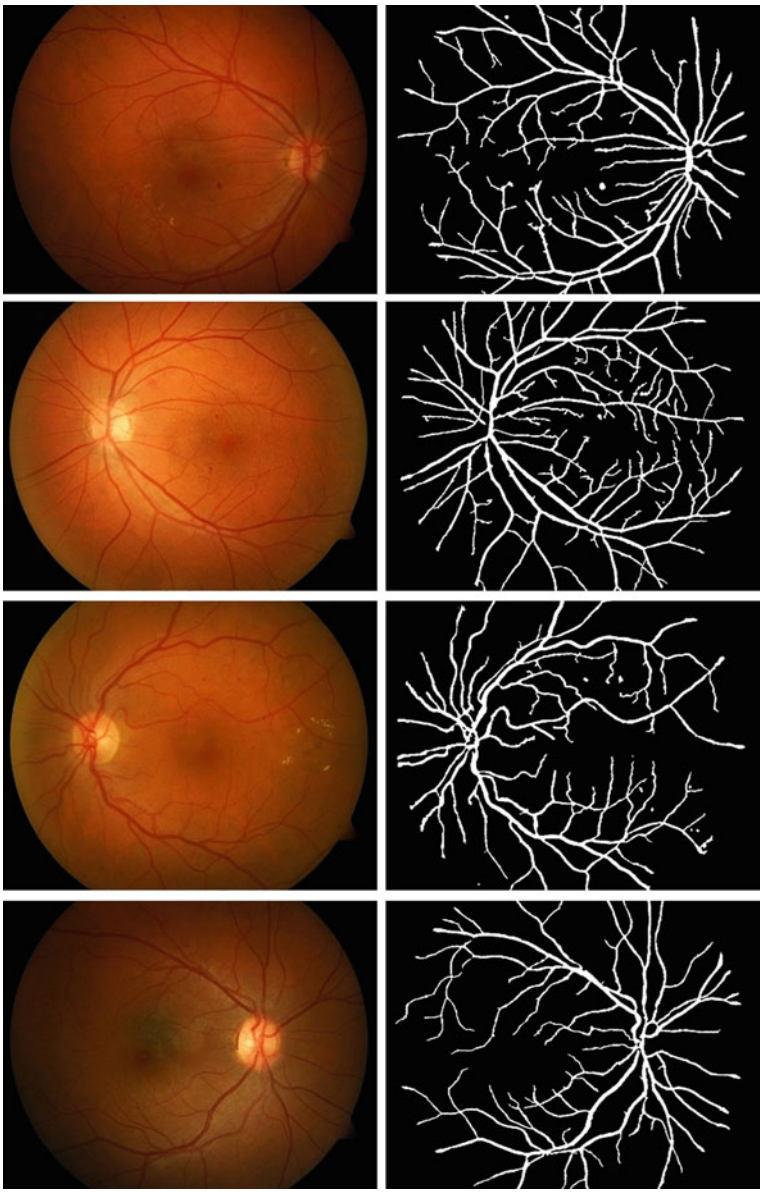


Fig. 10.7 *Left* Eye fundus color images (ddb0_v1_1/image005 & image006, ddb1_v1_1/image006 & image093). *Right* Binary images with segmented blood vessels and aneurysms

Table 10.3 Otsu's optimal threshold values for pathologies segmentation

Fig. no. and pathology	Otsu's τ	Fig. no. and pathology	Otsu's τ
Figure 10.4 hard and soft exudates	48	Figure 10.5 blood vessels	74
Figure 10.6 (1st row) hard exudates	51	Figure 10.7 (1st row) blood vessels and aneurysms	82
Figure 10.6 (2nd row) hard exudates	43	Figure 10.7 (2nd row) blood vessels and aneurysms	79
		Figure 10.7 (3rd row) blood vessels & aneurysms	92
		Figure 10.7 (4th row) blood vessels	82

Table 10.4 The four outcomes of a 2×2 confusion matrix

Actual condition	Predicted condition	Predicted condition
	Positive	Negative
Condition positive	True positive	False negative
	T_p	F_n
Condition negative	False positive	True negative
	F_p	T_n

negative *actual condition* (true pathology) delivered by a set of clinical tests, against a positive or negative *predicted condition* (segmented pathology) computed by the system. Performance and validation of such systems is commonly done using the data of the confusion matrix shown in Table 10.4. In this work, the confusion matrix entries are defined as follows: T_p is the number of correctly classified pathologies (conditions) from image samples tested as positive, F_p is the number of misclassified pathologies from image samples tested as positive, F_n is the number of misclassified pathologies from image samples tested as negative, and T_n is the number of correctly classified pathologies from image samples tested as negative.

Of all statistical quantities that can be derived from the confusion matrix (10.4) two of them stand out from others due to their usefulness for evaluating the capabilities of a classification system. More specifically, the *sensitivity* or *true positive rate*, given by $T_p^r = T_p/(T_p + F_n)$, is the proportion of pathological conditions that give positive test results. Similarly, the *specificity* or *true negative rate*, given by $T_n^r = T_n/(F_p + T_n)$, is the proportion of pathological conditions that give negative test results. In other words, T_p^r conveys all information relative to all "positive" cases and T_n^r refers to all "negative cases". In order to calculate the sensitivity and specificity rates of our test image samples, 20 hand-marked eye fundus color images of exudates and blood vessels were used as *ground-truth*. In Table 10.5 the values of sensitivity and specificity of the frequency transform based segmentation technique here described are compared to other techniques taken from the technical literature.

Table 10.5 Sensitivity (T_p^r) and specificity (T_n^r) rates for the segmentation of exudates and blood vessels in eye fundus images versus other segmentation techniques

Refs. for exudates	T_p^r	T_n^r	Refs. for blood vessels	T_p^r	T_n^r
Garaibeh [6]	0.921	0.990	Niemeijer [24]	0.690	0.970
Jaafar [4]	0.893	0.993	Rangayyan [25]	0.858	0.900
Kande [23]	0.860	0.980	Saleh [9]	0.842	0.966
Welfer [7]	0.705	0.988	Staal [26]	0.719	0.779
Lara-R [17]	0.941	0.991	Lara-R [17]	0.852	0.983

10.6 Conclusion

In this chapter the frequency transform based filtering approach to the problem of extracting exudates and blood vessels in eye fundus color images has been exposed [17]. An homomorphic high-pass filter is proposed for illumination correction of eye fundus color images. The key step before segmenting a specific pathology is the spatial frequency domain processing with a super-Gaussian band-pass filter. This novel type of filtering achieves adequate contrast between foreground objects against their background. We choose the DCT over other discrete frequency transforms since its computation works with real values and the results obtained with the homomorphic high-pass filter are visually better. The different stages required for the segmentation of exudates, blood vessels, and aneurysms are succinctly stated in Algorithm 1. Illustrative image samples taken from the DIARET databases were provided to demonstrate the capabilities of the frequency transform approach. High sensitivity and specificity rate values were obtained in our segmentation examples and compared with other reported segmentation techniques. Future work contemplates further tests with other databases such as DRIVE and STARE including unsupervised pathology recognition.

Acknowledgements Gonzalo Urcid thanks the National Research System (SNI-CONACYT) for partial financial support through grant No. 22036. Luis David Lara-Rodríguez and Elizabeth López-Meléndez are grateful with the National Council of Science and Technology (CONACYT) for doctoral scholarships CVU-332238 and CVU-332355, respectively.

References

1. Netdoctor: Diabetic retinopathy (eye disease), 2016. http://www.netdoctor.co.uk/conditions/diabetes/a833/diabetic-retinopathy-eye-disease/
2. Rowe S, MacLean CH, Shekelle PG (2004) Preventing visual loss from chronic eye disease in primary care: scientific review. J Am Med Assoc 291(12):1487–1495
3. Chou R, Dana T, Bougatsos C (2009) Screening older adults for impaired visual acuity: a review of the evidence for the US preventive services task force. Ann Intern Med 151(1):44–58

4. Jaafar HF, Nandi AK, Al-Nuaimy W (2011) Detection of exudates from digital fundus images using a region-based segmentation technique. In: Proceedings of the IEEE, 19th European signal processing conference 2011, pp 1020–1024
5. Budai A, Bock R, Maier A, Hornegger J, Michelson G (2013) Robust vessel segmentation in fundus images. Int J Biomed Imaging:1–11
6. Garaibeh NY (2014) Automatic exudate detection using eyes fundus image analysis due to diabetic retinopathy. Comput Inf Sci 7(2):48
7. Welfer D, Scharcanski J, Marinho DR (2010) A coarse-to-fine strategy for automatically detecting exudates in color eye fundus images. Comput Med Imaging Graph 34(3):228–235
8. Kaur J, Mittal D (2015) Segmentation and measurement of exudates in fundus images of the retina for detection of retinal disease. J Biomed Eng Med Imaging 2(1):27
9. Saleh MD, Eswaran C, Mueen A (2011) An automated blood vessel segmentation algorithm using histogram equalization and automatic threshold selection. J Digit Imaging 24(4):564–572
10. Zana F, Klein J-C (1999) A multimodal registration algorithm of eye fundus images using vessels detection and Hough transform. IEEE Trans Med Imaging 18(5):419–428
11. Walter T, Klein J-C (2002) Automatic detection of microaneurysms in color fundus images of the human retina by means of the bounding box closing. Med Data Anal Springer:210–220
12. Walter T, Klein J-C, Massin P, Erginay A (2002) A contribution of image processing to the diagnosis of diabetic retinopathy detection of exudates in color fundus images of the human retina. IEEE Trans Med Imaging 21(10):1236–1243
13. Kauppi T, Kalesnykiene V, Kamarainen J-K, Lensu L, Sorri I, Raninen A, Voutilainen R, Uusitalo H, Kälviäinen H, Pietilä J (2007) The DIARETDB1 diabetic retinopathy database and evaluation protocol. In: Proceedings of the British machine vision conference, BMVA Press, pp 15.1–15.10
14. Jain AK (1979) A sinusoidal family of unitary transforms. IEEE Trans Pattern Anal Mach Intell PAMI-1(4):356–365
15. Pratt WK (2007) Digital image processing, 4th edn. PIKS scientific inside. Wiley, Los Altos, California
16. Gonzalez RC, Woods RE (2008) Digital image processing, 3rd edn. Pearson, Prentice -Hall, Upper Saddle River
17. Lara RLD, López ME, Urcid G (2015) Blood vessels and exudates segmentation in eye fundus images based on Fourier filtering. Lecture notes in engineering and computer science: proceedings of the world congress on engineering and computer science 2015, WCECS 2015, 21–23 Oct 2015, San Francisco, USA, pp 503–507
18. Parent A, Morin M, Lavigne P (1992) Propagation of super-Gaussian field distributions. Opt Quant Electron 24(9):S1071–S1079
19. Otsu N (1979) A threshold selection method from gray-level histograms. IEEE Trans Syst Man Cybern 9(1):62–66
20. Dorland. Dorland's illustrated medical dictionary. Dorland's Medical Dictionary, Elsevier Health Sciences (2011)
21. Lappeenranta University: Standard Diabetic Retinopathy Database, 2011. http://www.it.lut.fi/project/imageret/diaretdb0/
22. Fawcett T (2006) An introduction to ROC analysis. Pat Rec Lett 27(8):861–874
23. Kande GB, Subbaiah PV, Savithri TS (2008) Segmentation of exudates and optic disk in retinal images. In: 6th Indian conference on computational visual, graphics and imaging proceedings 2008, pp 535–542
24. Niemeijer M, Staal J, van Ginneken B, Loog M, Abramoff MD (2004) Comparative study of retinal vessel segmentation methods on a new publicly available database. Med Imaging International Society for Optics and Photonics:648–656
25. Rangayyan RM, Ayres FJ, Oloumi F, Oloumi F, Eshghzadeh-Zanjani P (2008) Detection of blood vessels in the retina with multiscale Gabor filters. J Electr. Image 17(2):023018
26. Staal J, Abràmoff MD, Niemeijer M, Viergever M, Van Ginneken B et al (2004) Ridge-based vessel segmentation in color images of the retina. IEEE Trans Med Imaging 23(4):501–509

Chapter 11
The Distribution of HIV/AIDS Commodities: NPOS and Sustainability

Tatenda Talent Chingono, Sebonkile Cynthia Thaba and Charles Mbohwa

11.1 Introduction

Most research has focused on trying to implement and introduce humanitarian logistics and supply chain management to Humanitarian organizations in trying to make their efforts quicker, efficient and cost effective such researchers include, [1–7]. Some even further suggest Supply Chain Analytics for Humanitarian Logistics Transformation [8] focuses on the important role of humanitarian logistics, Networks for Africa in support for the implementation of the Millennium Development Goals. Some research has also focused on comparing the lessons learnt from both humanitarian and private sector logistics and supply chain management. [6, 9] Explore the way one such odd couple operates and learn from each other. Moving the World, a unique partnership between TNT, a global corporation specializing in transportation and logistics, and the U.N. World Food Program shows how two organizations can combine their core strengths to make a life-saving difference and increase a company's competitiveness and reputation. [10], focuses his study on

T.T. Chingono
Department of Quality and Operations Management, University of Johannesburg, Bunting Road Conbas, Auckland Park, South Africa
e-mail: ttchingono@uj.ac.za

S.C. Thaba (✉)
Department of Transport and Supply Chain Management, University of Johannesburg, Kingsway Campus, Auckland Park, South Africa
e-mail: scthaba@uj.ac.za

C. Mbohwa
Department of Quality and Operations Management, University of Johannesburg, Bunting Road Compus, Auckland Park, South Africa
e-mail: cmbohwa@uj.ac.za

S.-I. Ao et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-981-10-2717-8_11

Zimbabwe, discussing the challenges, difficulties and problems faced by humanitarian organizations in running logistic systems. Case studies of humanitarian organizations were conducted. Logistics Performance measurements were also done for the World Food Program, the International Red Cross Society and the Zimbabwe Red Cross Society, the World Health Organization, the United Nations Children's Fund and the Zimbabwean Civil Protection Organization in Zimbabwe are discussed. This research will take a different approach, it will focus on the distribution mechanisms of the supply chains so as to strengthen the much needed service delivery and to also try and make sure that goods arrive where they are needed and in time. United Nations Program on HIV/Aids UNAIDS (2002) report, since diagnosis of the first case of HIV infection in Zimbabwe in 1985, the epidemic has grown to about 2.3 million people infected of the country's total 12 million population. An estimated 34 % of sexually active adults aged 15–49 years are infected with HIV. It is estimated that about 600,000 people have full-blown AIDS. More than 3,800 people are dying every week due to this epidemic, which has become the top killer disease in the country. At least 70 % of hospital beds in medical wards are occupied by patients with AIDS-related conditions. Life expectancy has fallen to 43 years, while infant mortality has more than doubled to 130 per 1,000 live births, [11] the global project report the increased morbidity and illness related to AIDS is stretching scarce health resources at a time when the country is facing enormous economic hardships, [12]. Antiretroviral therapy has substantially reduced HIV related morbidity and mortality worldwide. Although it is as such, access to antiretroviral therapy is still quite limited in Zimbabwe, with over 300 000 people still in need of Antiretroviral therapy. Zimbabwe is one of the 20 Sub-Saharan African countries having the highest unmet needs for antiretroviral therapy [13]. Despite the good sign of decline, the double digit prevalence rate of 13, 7 % is still unacceptable and high, thus much still needs to be done, [14].

Logistics and supply chain management is increasingly becoming important for Humanitarian operartions in a similar way as it is for private sector operations. Not until these recent years, humanitarian logistics was neglected hence supplychain and logistics skills developed slowly. Change is being witnessed slowly, logistics and supply chain management are moving slowly to the center stage of relief operations, as absent them, Humanitarian relief is vitually imposible. This became more clear during the Indian Ocean Tsunami disaster [5]. Unfortunately, disaster relief is and will continue to be a growth market. Both natural and man-made disasters are expected to increase by five-fold over the next fifty years due to environmental degradation, rapid urbanization and the spread of HIV/AIDS in the developing world. According to the Munich Reinsurance group, the real annual economic losses have been growing steadily, averaging US$75.5 billion in the 1960s, US$138.4 billion in the 1970s, US$213.9 billion in the 1980s and US$659.9 billion in the 1990s [5]. Thus the logistics and supply chain management systems of humanitarian agencies such as Natpharm need to be constantly monitored and

strengthened. Stakeholders agreed that, notwithstanding the declining HIV and AIDS prevalence rate, the prevalence rate was still too high and Zimbabwe still faces enormous challenges in the fight against HIV and AIDS in all the key programme areas of prevention and care. As such, medical services are experiencing growing demand because of HIV and AIDS and hospital services are increasingly strained owing to escalating cases and continued budget reductions because of the unfavourable economic situation Zimbabwe is facing. It is therefore in against this background that Non-Governmental Organizations in support of the distribution.

11.2 Literature Review

Humanitarian supply chain refers to the network created through the flow of supplies, services, finances and information between donors, beneficiaries, suppliers and different units of humanitarian organizations for the purpose of providing physical aid to beneficiaries [15]. Humanitarian supply chains might include functionalities which may not typically fall into the field of humanitarian logistics. Managing relationships with donors, performing needs assessments, planning for supplies required and monitoring and evaluating the impact of distributed supplies, are usually the responsibility of non-logistics program units [5]. Humanitarian supply chains include units implementing programs, managing grants with donors, controlling budgets and monitoring activities which must coordinate with logistics units. Humanitarian supply chains cannot be built by solely increasing the capacity or responsibility of individual units, but are formed by the creation of stronger links between units within the supply chain. Humanitarian logistics information systems can improve the flow of information with other units, in a mutually constructive manner, improving the effectiveness of the humanitarian supply chain. Several Non-Governmental Organisations are in support to providing the ARVs in aid to the needing people of Zimbabwe. Some of the NGOs include Medicines Sans Frontiers (MSF), DART (Annexe) Parerinyatwa, to mention just a few. The distribution of ARVs by NGOs is generally governed by the Ministry of Health and Child Welfare and the Medicines Control Authority of Zimbabwe. Most of the NGOs distribute the drugs to the countries' district and provincial hospitals and clinics. Patients would then collect the drugs from these hospitals and clinics. But recently MSF has started providing the drugs to some needy individuals directly through the clinics and hospitals. MSF in particular is one of the leading organisations in support of the ARV distribution program with projects in Bulawayo, Tsholotsho, Buhera, Gweru, Epworth and Beitbridge. MSF is ensuring medical care to more than 35,000 HIV patients out of whom more than 16,000 are receiving Anti-retroviral therapy (Fig. 11.1)

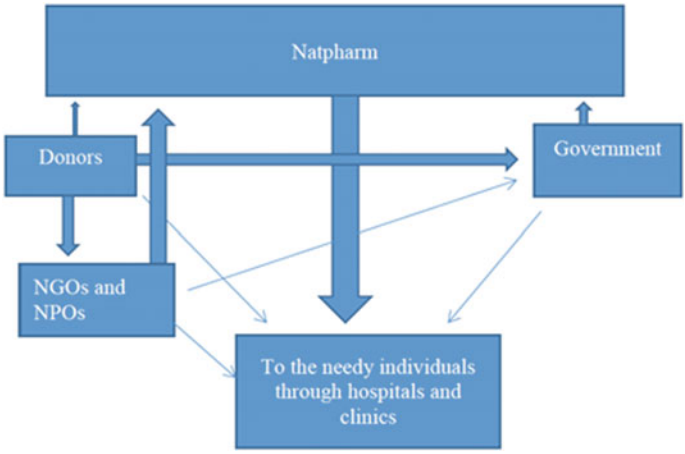


Fig. 11.1 Humanitarian supply chain of medicinal drugs

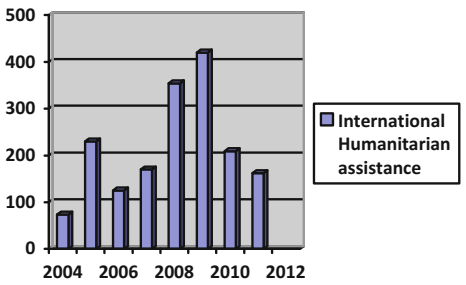
11.3 Methodology

The research was conducted in Zimbabwe, 150 questionnaires were distributed. 105 responses were recorded. At most locations short interviews were also held with various stakeholders including government, NGOs and Embassy officials that assisted in filling in the research instruments.

11.4 Results

Figure 11.2, Zimbabwe scored 'medium' in the 2012–13 ECHO Vulnerability Index thus funding started to deteriorate, hence Zimbabwe remains in Debt. Official Development excluding debt relief showed a steady constant rise to up close to a billion United States dollars. In 2015 the National AIDS council noted that the situation might even get worse as more than 350,000 might not get aid in the form

Fig. 11.2 International humanitarian to Zimbabwe



of ARVs as the Global fund which contributes more than 30 % of total Aid did not commit to any funding to Zimbabwe from the year 2014 (www.nac.org, accessed December 21, 5015). The major aid givers to the country include; The European Union (EU) and most of its member states, The United States, Canada, Australia and Japan. Also of importance to note is United Nations (UN) related agencies (UNIEF, WFP, UNDP, UNHCR, FAO, and the Global Fund). ZIMFUND coordinates and manages donor funds normally. Key committee members from the ZIMFUND meet regularly with representatives from the Zimbabwean government, mostly from the ministry of health and observers representing the World Bank and lastly the resident UN Coordinator.

Other donors general mainly contributes funds/commodities to UN agencies including UNIAIDS, UNFPA, ECHO, UNICEF and WHO, these will allocate the donations on their behalf to where they are needed most. Donor organisations linked to other country governments or continents such as the DfID, EU, US government related agency have chosen to be more involved and distribute commodities recourses and skills directly to where they are needed through committed organisations that get the tenders like the Elizabeth Glaser Paediatric AIDS Foundation (EGPAF) and Absolute Return for Kids (ARK).

Zimbabwe's governing body has contributed also and tried to exhibit loyalty and ownership of the crisis to its limited extent in the fight against the epidemic. Firstly it instated legislation through parliament, such that 3 % of every person's pay within the tax bracket was taxed. Cooperates were also liable to pay a tax as well towards the alleviation of the pain and suffering of those hampered by the devastating effects of the epidemic. Secondly it made sure HIV/AIDS programmes had priority votes in the annual budget, hence allowing the allocation of certain funds to go towards the HIV/AIDS programmes. In 2015 the government contributed a total of USD 830,000 to the procurement of ARVs. Government's counterpart or partnership contribution has been estimated to 24 % after all its efforts have been summed up. Government's contribution is linked to economic performance thus when it goes down so will the funds, of late it's been growing at 5 % (www.globalfund.org). This is still not good enough as a large number of people are still in desperate need of assistance.

Even though it froze most post in the health sector due to lack of funds to recruit new staff, the government tried and committed to supporting and maintaining the existing staff to the best of its abilities even in extreme financial situations. Payment of salaries was could sometimes be late on rare occasions, but the workers would eventually receive it so as to motivate the keep their jobs and tend to the patients. It also facilitated payment of a retention allowance to health staff with the assistance of partner organisations. There is also evidence of facilities infrastructure maintaining, but little of upgrading. The government also assist in making and improving the skills of the staff through in service training at major health institutions and health facilities. Thus it plays a pivotal role in making sure that all acquired and donated material eventually reaches its intended beneficiaries. Thus it successfully helps the major supply chain of HIV, commodities that starts from various donors and NGOs to NatPharm then to government's health facilities.

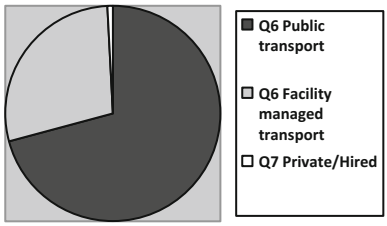


Fig. 11.3 Mode of transport

Government also supported and facilitated Natpharm's involvement in the overall supply chain of the National ARVs programme. Many partnerships with various NGOs and aid agencies were also forged (www.globalfund.org).

Figure 11.3 Pie Chart showing proportionate percentage distribution of respective types of transport that are used. It is evident that most facilities use their own transport which they can easily manage and thus reliable. Some to the lesser extend (more than a quarter) relay on private or hired transport. This can bring about problems as it cannot be fully managed/controlled and terms/conditions can easily change, thus reliability is questionable. This was evident in some of the clinics/hospitals visited (mostly government, and a few private).

Figure 11.4, show the relative percentage of carriers used per organisation. It is complemented by Table 11.1, showing that most organisation relay on an average of 4 carriers. This can be a challenge when operating countrywide operations (4 carriers cannot cover all 7 provinces, thus the probability of delay in service provision is also higher). Hence there is a need for to equip these organisations with more vehicles.

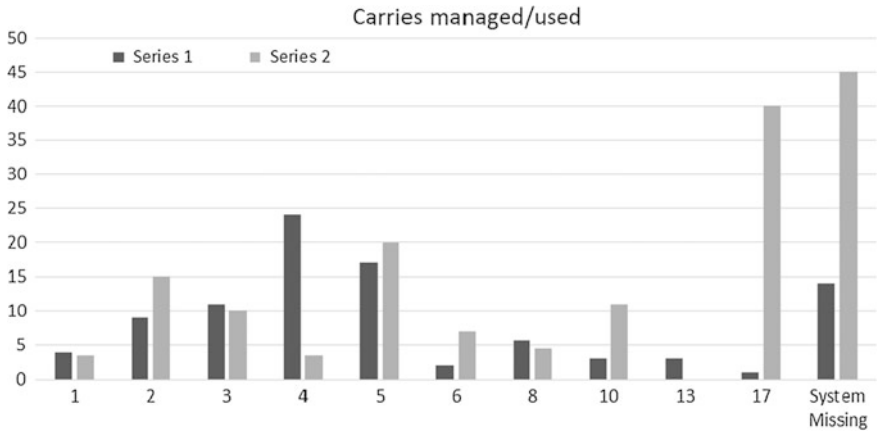


Fig. 11.4 Bar graph depicting number of carriers managed in relation to percentage

Table 11.1 Statistics for number of carriers used

Mean	5.529411765
Media	4
Std. deviation	4.25907336
Range	16
Minimum	1
Maximum	17

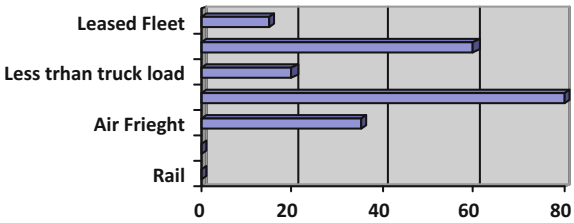


Fig. 11.5 Percentage of relative transport used

The bar graph in Fig. 11.5 depicts that most organisations use their own fleet and there also deliver a truckload. Rail and Ocean transport are not used at all as they are usually slower when compared to the others.

Figure 11.6 shows that more than 40 % of the study sample use up to ten or more contact points to manage their supply chain. This is mainly so because the supply chains are vast and complex, they mostly start in foreign overseas countries. Implementation and Distribution candidates that participated in the survey were mainly from NGOs and most were qualified as supply chain/logistics officers. This is very important as NGOs and donors do not want to look bad after a relief operation and especially in front of the media if the operation goes wrong. Unqualified personnel as well as fame seeking celebrities who might want to do the work should not be employed. Organisations should try to recruit employees with the necessary skills, knowledge and experience. These will help make the sure the project/program is successful. Volunteers with their own interests should also be avoided. Even when recruiting, recruits should be placed in familiar places with roughly the same culture such that they do not get lonely frustrated and stressed

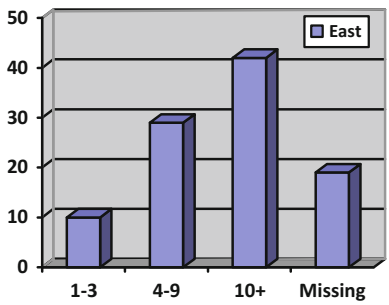


Fig. 11.6 Supply chain contact points

because if the change in environment and the way things are done. Workers should also be united and have the same basic positive perspectives. Negative perspectives can hinder and disrupt relief efforts as they often lead to miscommunication and impaired collaborative effort. It has been noted that its mostly inexperienced aid workers who normally indulge in this sought of behaviour hence it was encouraging to note that most organisation had qualified experienced and matured respondents who can steer humanitarian operation in the right way.

Some organisations that did not have professionals with management and supply chain management are encouraged to recruit these to lead relief operations as they can also ensure that the operation is effective and timely so as to save thousands of lives. Delivery of quality care and completion of the many involved tasks also requires experienced manager. Managers are also required to establish good relations with government, create partnership with it and other donor agencies, multi-task and facilitate diplomatic relations. They should also be able to assess community needs and come up with culturally acceptable solutions as well. Donors also need to be satisfied. The researcher also complimented the survey result with an internet search of the respondents' names to check their experiences and qualifications on site such as LinkedIn.

11.4.1 Distribution

Zimbabwe has adequate internal transportation and electrical power networks, however maintenance has been neglected over several years due to harsh economic conditions. Poorly paved road full of pot-holes link the major urban and industrial centres and cities flows down distribution efforts and damages delivery vehicles hence un-expected delays and disruptions in the supply chain can be experienced. The roads are also filled with mostly corrupt police officers which can also be a problem. Rail lines managed by the National Railways of Zimbabwe tie it into an extensive central African railroad network with all its neighbours this can be helpful in distributing large quantities of medicinal relief commodities, but NatPharm does not use it as it has an adequate vehicle fleet and trucks. Most of the imports are brought though ships even though Zimbabwe has no Coast, trucks transport it from Neighbouring countries. Some fly it from overseas through the Harare international airport.

Donor funded projects should focus more on increasing distribution towards key populations most affected by HIV and AIDS. These include heterosexual couples in stable unions, orphans and vulnerable children, mobile workers (e.g. truckers, small scale miners, informal traders) and young people mostly at the periphery of society in the rural areas. People living in hard to reach geographical locations are also underserved and have to adopt various strategies to access services including group collection of periodic medicines. Most-at-risk populations include young orphaned sex workers, people living with disabilities, prisoners and rarely men who have sex with men.

Access to services is geographically homogeneous due to the decentralised Scale up of services through health and community structures at all levels. However some hard to reach populations remain including resettlement and illegal settlement areas, mobile populations such as small scale and informal miners, and areas distant from health facilities with geographical barriers to access varying on a seasonal basis e.g. restricted access due to rains which might result in flooding.

11.4.2 Storage

In general, all commodities need procedures for safe storage that increase their shelf life and make them readily available for distribution. A considerable amount of organisations surveyed did not have their own storage facilities, most acquire services from NatPharm. NatPharm has two major Regional warehouses and offices in Harare and Bulawayo and 4 relatively smaller branches (Mutare, Chinhoyi, Masvingo and Gweru). Each Regional Store is linked with two branches and health providers in its region. Harare regional ware house serves Greater Harare, Mashonaland East, Mashonaland Central and Harare and Parirenyatwa Central Hospitals. The Harare Regional Store also serves Mutare and Chinhoyi branches. Bulawayo Regional warehouse supplies all hospitals and clinics that cater for the general public at Bulawayo urban and Matabeleland North and South Provinces. The Bulawayo Regional Store also supplies to Gweru and Masvingo branches which serve health facilities in Midland and Masvingo provinces.

NatPharm six warehouses nationwide have a total storage capacity of more than 12812 m^2. The regional stores also act as the major international receiving warehouses and together they have a capacity of 6621 m^2 NatPharm is currently under stocked on vital HIV/AIDS commodities, and current space and organisation is not sufficient. The space is also not sufficient for reasonable expansion. Calculations carried out to estimate the storage space required for HIV and AIDS commodities projected a total requirement of about 4008 m^2.

In general, at all NatPharm Stores, all commodities are stored in accordance with good storage practices and laid down standard operating procedures. Commodities are stored at recommended temperatures and issued out on first in first out (FIFO) basis. Cartons and Products are in good condition and not crushed due to mishandling. Products are also always protected from direct sunlight, water and humidity during all seasons. Storage area was also visually free from Rodents. The storage area was also safe and secure under lock and key. Al hazardous waste was properly disposed and non-accessible to non-medical personnel. Theft/loss indicators are also being continuously monitored.

The current practice is for NatPharm to include a 4–6 % charge on the value of the drugs of which government and donor organisations take care of this. This money is intended to cover covers storage, handling, and distribution of the donated commodities to victims. If this model is followed, it is important that the charge is carefully discussed because of the high cost of ARVs.

From the survey as well, most donor agencies as well do not have adequate storage space as well but they also noted that they maintain and keep their facilities up to recommended storage and safety measures. Some of them even have up to date security systems way advanced and better than NatPharms.

11.5 Recommendation

NatPharm should provide additional secured storage space for the anticipated increased volume of HIV and AIDS commodities. The space to be secured should be calculated taking into consideration the interim logistics subsystem design including the maximum stock level at the central level and the refurbishment of existing cold rooms to take care of increasing volume of commodities. It is recommended that financing for proper storage and distribution be sought and secured as soon as possible and that the charges are carefully negotiated because of the high cost of ARVs.

It is also recommended that any price staggering by NatPharm based on value should also consider possible devaluations in the price of ARVs over time. Payment terms should be strictly adhered to and guaranteed through MOUs. Extra measures that NatPharm will have to institute to ensure security of the supply should be also taken into consideration.

If the capacity of other participants in the survey could also increase it would ease the pressure and burden from NatPharm. Natpharm can also try and adopt some of the security measures used by fellow industry players that participated. These include the use of bio-metric entrance system and also security cameras.

Most organisations that participated in the survey did not have a good number of vehicles to serve the nation as a whole but NatPharm did. It had a record 23 vehicles that service the nation. A good number of good functioning vehicles are available and are efficiently used for routine and emergency distribution hence in general, orders are delivered on time between NatPharm Stores and Provincial and District Hospitals and other health facilities. The distribution capacity was strengthened mainly by the European Union who donated of six seven tone trucks. This fleet has made it possible for NatPharm to supply the much needed HIV/AIDS heath commodities to all health facilities in the country a using robust IT scheduling system. The efficiency of distribution system is sometimes disrupted by the erratic availability of fuel, bad roads and extreme weather. The USAID-funded DELIVER project also donated further three 7-tonne trucks dedicated to delivery under the new provisional HIV and AIDS Ordering and Distribution System. The support also includes provision of salary for drivers, fuel and vehicle maintenance. In addition, the support project can also acquire additional distribution to private courier where applicable and necessary. Although Zimbabwe has one of the better road systems in Africa it is being poorly maintained and during the rainy season

some clinics, hospitals, and community-based distribution workers become inaccessible due to the bad condition of roads, thus leading to increased stock outs. In light of the scaling up of ART services, it is recommended that more detailed assessment of NatPharm's distribution system with specific focus on future transport needs, needs to be conducted. Collaborations with other industry players is also encouraged. If the government cannot assist donor agencies are also recommended to assist in road maintenance in inaccessible areas.

The study shows that there is a good number of sufficiently functioning vehicles that are available and are efficiently used for routine and emergency distribution hence in general, orders are delivered on time. The distribution capacity is strengthened mainly by the donation from the Western Countries. Those with a good fleet are able to distribute essential supplies to all health facilities in the country using a robust scheduling system. The efficiency of distribution system was often hampered by erratic availability of fuel. The support from donors also includes provision of trucks, salary for drivers, fuel and vehicle maintenance. In addition, most agencies/organizations have support projects that can also outsource distribution to private courier if need arises.

Although Zimbabwe has one of the better road systems in Africa outside South Africa, during the rainy season some clinics, hospitals, and community-based distribution workers become inaccessible due to the bad condition of roads, this leads to increased stock outs. The government or interested stake holders are urged to address this.

11.5.1 Non-governmental Organisations (NGOs)

There are windows of opportunities that exist that sometimes NGOs fail to exploit due to lack of awareness that the opportunities exist. Government funds available were not accessed by civil society organizations due to inadequate awareness about the availability and the procedures required to access the funds.

11.5.1.1 Governance

Many organisations fail in the two central pillars of good governance that is Transparency and Accountability. Many organisations lack sound systems for financial management, programme monitoring and evaluation, and managing overall programme performance that ensure they consistently earn stakeholders' trust. The Founder or Founders tend to control and manage the affairs of the organisation with minimal participation from other members.

11.5.1.2 Lack of Governance Structures

No strategic plans that guide the organisation to know what the objectives are and enable the organisation to identify he resources needed. Most strategic plans developed are for donor purposes only and do not reflect the actual needs to be addressed by the communities.

11.5.1.3 Efficacy of Boards

NGOs in many countries do not have effective governance structures and where a board exists, they are rarely effective in providing strategic leadership in ensuring resources are mobilised Boards are supposed to provide guidance and oversight to the operations of the NGO. However, many boards are not aware of their role in resource mobilisation. The presence and extent of involvement of management and governance structures in NGOs influenced their ability to mobilise resources. Many times NGOs did not have governance instruments such as constitutions, policies and guidelines and this tended to is care off potential donors.

11.5.1.4 Benefits of Merging

For NPOS to leverage their chances of success when it comes to funding and ongoing operations, there is a need of merging as organisation. The organisations will benefit in terms of human resources, expertise, positive reputation and strengthening of programmes.

11.6 Conclusion

Surveyed Organisations, including Natpham encounter a number of challenges, which have a considerable impact on their ability to deliver and perform fully and effectively. Some of the common challenges noted from unstructured interviews that also disrupts or tent to limit the success of the supply chains include:

- Payments can be made late by the ministry of health.
- Salaries of key staff is not consistent with Market.
- Some donor agencies not fully committing after 2014, including the global fund.
- Increased stock outs in some difficult to access areas during the rainy season.
- Increased government control and grip on NGO operations.
- Operating in an environment with a very high demand for HIV/AIDS commodities.

- High employee turnover rates due to harsh economic and sometimes political environment.
- Additionally, as more and more patients are enrolled, the main ART distributing sites initiating patients on ART are quickly reaching their maximum capacity.

For future research the study have to look at the non-profit organisations (NPO) within the Southern African Developing Communities (SADC), looking into the humanitarian distributing, for comparison and coming with effective model effecting for all developing countries.

References

1. Chingono TT, Thaba SC, Mbohwa C (2015) An overview of the distribution of HIV/AIDS commodities in Zimbabwe. In: Lecture notes in engineering and computer science: proceedings of the world congress on engineering and computer science 2015, WCECS 2015, 21–23 Oct 2015, San Francisco, USA, pp 760–763
2. Clark A, Culkin B (2007) A network transshipment model for planning humanitarian relief operations after a natural disaster. Presented at EURO XXII–22nd European conference on operational research, Prague
3. Thomas A (2003) Humanitarian logistics: enabling disaster response. The Fritz Institute
4. Kleindorfer PR, Van Wassenhove LN (2004) Managing risk in global supply chains in Strategies for building successful global businesses, Chapter 12. In: Gatignon H, Kimberley JR (eds) Cambridge University Press, Cambridge, pp 288–305
5. Thomas A, Kopczak L (2005) From logistics to supply chain management: the path forward in the humanitarian sector. Technical Report, Fritz Institute, San Francisco, CA, USA
6. Tomasini R, Van Wassenhove LN (2004) Genetically modified food donations and the cost of neutrality, logistics response to the 2002 Southern Africa food crisis. INSEAD Case 03/2004-5169
7. Thompson (2008) Supply chain analytics for humanitarian logistics transformation. In: Humanitarian logistics: network for Africa Rockefeller foundation Bellagio center conference, Bellagio, lake Como, Italy
8. Nyaguthie (2008) The important role of humanitarian logistics. Oxfam-GB, Pretoria
9. Tomasini R, Van Wassenhove LN (2009) From preparedness to partnerships: case study research on humanitarian logistics. Int Trans Oper Res 16:549–559
10. Mbohwa C (2008) Identifying challenges and collaboration areas in humanitarian logistics: Southern African perspective. In: Humanitarian logistics: network for Africa Rockefeller foundation Bellagio center conference, Bellagio, lake Como, Italy
11. UNIDO (2011) Pharmacuitical sector profile; Zimbabwe, global UNIDO project: strengthening the local production of essential generic drugs in least developed and developing countries
12. Takang E, Dragana V, Celestine K, Nyeriwa J (2006) Management of HIV and AIDS commodities in Zimbabwe—A capacity assessment of NatPharm and Ministry of Health and Child Welfare
13. Global AIDS Responds Progress Report (2012) Zimbabwe Country Report, Reporting period: January 2010–Dec 2011
14. Kerina D, Stray-Pedersen B, Muller F (2013) HIV/AIDS: the Zimbabwean situation and trends. Am J Clin Med Res 1(1):15–22
15. Mentzer JT, DeWitt W, Keebler JS, Min S, Nix NW, Smith CD, Zacharia ZG (2001) Defining supply chain management. J Bus Logist 22:2

Chapter 12
Representation of a DNA Sequence by a Substring of Its Genetic Information

Bacem Saada and Jing Zhang

12.1 Introduction

To determine the affiliation of a strain to a given specie, biologists compare it with a known sequence of reference of the specie to which it is presumed to belong. If the similarity percentage is very large, we conclude that this sequence belongs to a well-defined specie. The comparison between sequences can also allow the comparison between different species. These comparisons lead to the conclusion that two species have a common ancestor or not.

In order to properly analyze the results of alignment methods and comparison of DNA sequences, we assign weights to the various pairs of the sequence to calculate the degree of similarity and the costs of non-similarity between sequences. This operation allows us to infer relationships between the sequences. This relationship is described as the degree of similarity between sequences. This degree of similarity is quantified by a score. The most commonly used alignment algorithms between sequences are the Smith-Waterman algorithm [1] which determines local alignment between DNA sequences and the algorithm of Smith and Waterman [2] which determines a global alignment between DNA sequences.

B. Saada (✉) · J. Zhang
College of Computer Science and Technology, Harbin Engineering University, Harbin 150001, China
e-mail: basssoum@gmail.com

J. Zhang
e-mail: zhangjing@hrbeu.edu.cn

S.-I. Ao et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-981-10-2717-8_12

12.2 State of the Art

The process of alignment and comparison of DNA sequences presents several problems:

Today, there are several open access DNA sequences databases. These banks continue to grow at an exponential rate. In 2006, GenBank, for example, created within the framework of international collaboration on nucleotide sequencing, contained over 65 billion nucleotide bases [3]. In 2013 it grew to 154.2 billion bases. Nowadays, the quantity of information can reach petabytes in size. In this case, applying a treatment on a large number of sequences to infer which sequence belongs to a given specie, is very costly in terms of execution time and needed resources. To decrease the amount of stored information, researchers are trying to reduce the number of DNA sequences stored in their databases and keep only the DNA sequences that best characterize each specie.

Storage of such alignment is also a problem. Thereafter, any analysis or interpretation of this alignment would be strenuous. If the researcher decides to use a portion of the sequence, no current algorithm allows him to optimally choose the desired length to extract from the original chain.

To solve the problems described above and to optimize the use and performance of DNA sequences alignment algorithms, several researches were conducted.

Furthermore, the growth of the new DNA sequences alignment technologies has enabled the study of human genome. The size of those genomes reaches 3 billion bases. It can even reach more than 100 billion bases in some amphibian species. It is not possible to use conventional algorithms for the alignment of DNA sequences. Indeed, the result of an alignment between entire genomes would be an alignment of millions of base pairs. Taking into consideration the execution time, the application of such an operation is impossible for usual microcomputers.

The collection, analysis and understanding of this enormous amount of information became a challenge for taxonomic researches. This has led to the development of DNA compression algorithms. Based on the English text compression of the four bases {A, C, G, T}, those algorithms try to reduce the ratio "bits per base" [4, 5].

As a conclusion, research themes were therefore based on the parallelization of classical DNA sequences alignment algorithms to reduce the execution time of these algorithms or to compress DNA information. No research addresses the reduction of the size of the DNA bases to be stored.

12.3 New Approaches to DNA Sequences Alignment

In this section, we propose an approach to DNA sequences alignment that can represent DNA sequences with a sub chain of their genetic information. First, we list in the first instance, the motivations of our approach. After that, we will present our approach while offering a study of its complexity.

12.3.1 Motivations

To overcome the problems described above, we tried to propose a new approach that attempts to combine the performance of algorithms for DNA sequences alignment and significantly reduce the size of stored genetic information.

Our approach will essentially provide an algorithm that is:

- **Able to determine an optimal alignment for a length requested by the researcher**: Usually alignment resulting from the implementation of the Smith-Waterman algorithm has a length of 1500 base pairs. We will try to present our approach through an algorithm able to give an optimal alignment for a length less than half the size of the sequences to be aligned.
- **Able to represent a DNA sequence, not by its full genetic information but by a smaller sub chain**: It is highly desirable that a DNA sequence is represented not by its full genetic information but by some of its DNA only. Our approach will also try to represent a set of DNA sequences by a sub chain.
- **Able to reduce the amount of data stored in databases**: By reducing the size of the genetic information representative of a DNA sequence, the amount of data stored in the database will be reduced. And thus all data can be stored in the same storage media.

12.3.2 Algorithm for Determining a Best Local Optimal Alignment

12.3.2.1 Introducing the Approach

The major reason that led researchers to use heuristics approaches is that with a large number of sequences, the dynamic programming algorithms are quite expensive in terms of execution time. If further treatment is added, the total execution time will grow more. For that an alignment would be well analyzed. It is desirable to scan the entire alignment. Therefore, the approach that we present seeks to analyze the similarity score of the entire alignment to extract the sub chain with the highest percentage of similarity between the two input data sequences. Indeed, this algorithm, after building the entire sequence alignment, saves the values of the matrix score. Thus, it is possible to find the alignment portion with the highest similarity score to better represent the sequences. This part will be determined after calculating the score of similarity between the different regions of the alignment.

12.3.2.2 Algorithm

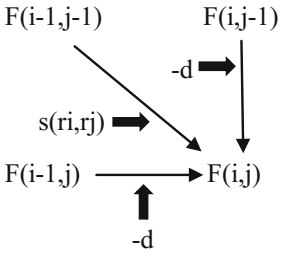
As well as the alignment algorithms between DNA sequences, we perform computing of a score matrix. We will use the following recurrence formula:

$$\mathrm{F(i,\ j)}\ = \max \begin{cases} 0, \\ F(i-1,\ j-1) + s(si,\ yj), \\ F(i-1, j) + d, \\ F(i,\ j-1) + d. \end{cases}$$

For the energy costs of each alignment operation, we will use the following values:

Del = −1; ins = −1; sub = 1; id = 1.

The selection of these values as parameters was made so that energy costs would be uniform and equal in absolute value. This choice will exclude the consideration of these energy costs in any subsequent treatments. We keep track of the highest value in the matrix of the score. The alignment will be built starting with this value following this rule:



For a good alignment analysis, it is desirable to scan the entire alignment built. Therefore, the algorithm starts, like any classical algorithm for DNA sequences alignment, by calculating the score matrix over the entire sequences.

Algorithm 1 Build Local ALignment

Require: $S \geq 0\ I \leq 0\ D \leq 0\ A \neq \varnothing\ B \neq \varnothing$

```
{Compute Matrix}
maxScore ← 0
maxRow ← 0
maxCol ← 0
AlignmentA ← ""
AlignmentB ← ""
i ← maxRow
j ← maxCol
for i ≤length(A) do
  F(i,0) ← 0
end for
for j ≤length(B) do
  F(0,j) ← 0
end for
```

```
for i ≤length(A) do
  F(i,0) ← 0
end for
for j ≤length(B) do
  F(0,j) ← 0
end for
for i ≤length(A) do
  for j ≤length(B) do
    Match ← F(i − 1, j − 1) + M
    Delete ← F(i − 1, j) + D
    Insert ← F(i, j − 1) + I
    F(i,j) ← max(Match, Insert, Delete)
    if F(i,j) > maxScore then
      maxScore ← F(i,j)
      maxRow ← i
      maxCol ← j
    end if
  end for
end for
```

The algorithm keeps track of the matrix cell that contains the highest similarity score. The construction of the alignment starts from the box. After building this alignment, the algorithm seeks to determine the alignment region best suited, in terms of similarity score, to represent the sequences. This part will be determined after calculating the average of the similarity scores of the alignment's different regions.

```
{Function : Build Optimal Alignment}
while ((i ≥ 0 or j ≥ 0) and F(i,j) ≥ 0) do
  Score ← F(i,j)
  ScoreDiag ← F(i − 1, j − 1)
  ScoreUp ← F(i, j − 1)
  ScoreLeft ← F(i − 1, j)
  if Score == ScoreDiag + S(Ai, Bj) then
    AlignmentA ← Ai + AlignmentA
    AlignmentB ← Bj + AlignmentB
    i ← i − 1
    j ← j − 1
  else
    if Score == ScoreLeft + d then
      AlignmentA ← Ai + AlignmentA
      AlignmentB ← " − " + AlignmentB
      i ← i − 1
    end if
```

```
    else
      if Score == ScoreLeft + d then
        AlignmentA ← Ai + AlignmentA
        AlignmentB ← " − " + AlignmentB
        i ← i − 1
      end if
    else
      AlignmentA ← " − " + AlignmentA
      AlignmentB ← Bj + AlignmentB
      j ← j − 1
    end if
    AlignmentB ← Score
  end while
  if F(i, j) ≥ 0 then
    while i ≥ 0 and F(i, j) ≥ 0 do
      AlignmentA ← Ai + AlignmentA
      AlignmentB ← " − " + AlignmentB
      i ← i − 1
      AlignmentB ← F(i, j)
    end while
    while j ≥ 0andF(i, j) ≥ 0 do
      AlignmentA ← " − " + AlignmentA
      AlignmentB ← Bj + AlignmentB
      j ← j − 1
      AlignmentB ← F(i, j)
    end while
  end if
  region ← ""
  for all AliRegion ⊂ ALig do
    if AVGscore(AliRegion) > AVGscore(region) then
      region ← AliRegion

    end if
  end for
```

12.3.2.3 Complexity of the Approach

Consider two sequences seq_1 seq_2. Let $l_1 l_2 l_3$ the respective lengths of the first sequence, the second sequence and of the alignment result; and consider l the length of the chain requested by the researcher.

The complexity of the filling phase is O ($3l_1l_2$). Indeed, to fill each cell of the matrix, we must realize three arithmetic operations. The complexity of the optimal alignment of the construction phase for the required length of O (3l); Indeed, to detect the region with the highest similarity score, our algorithm will perform l arithmetic operations. It would calculate the average of the similarity scores of this region in (l3 − l) total regions.

The total complexity of this approach is O ($3l_1l_2 + 3l + [l_3 - l]\ l$). The complexity of this approach is polynomial of order 2.

12.4 Experimental Results

In this section, we will illustrate the experimental results of our approach over the Smith and Waterman algorithm. In the first part, we will interpret the results of experiments obtained on the similarities percentages. Subsequently, we will present a comparative study based on the execution time of our approach compared to the classical approach.

12.4.1 Species of the Experiments

To measure our approach's performance, we used a set of DNA sequences of different genres. The size of the DNA sequences varies between 1300 and 1550 base pairs.

The diversity of the classification of these sequences allowed us to conduct a comparative study presented in three steps:

- Experimental results for species of a same genus.
- Experimental results for a set of species of the genera of the phylum Firmicutes.
- Experimental results for random species.

12.4.1.1 Experimental Results for Species of the Same Genus

We analyzed the experimental results for 11 species of the genus Bacillus. The species used are amyloliquefaciens Anthracis, Azotoformans, Badius, Cereus, Circulans, coagulans, licheniformis, megaterium mycoides Psychrosaccharolyticus, pumilus. This experiment would analyze the percentages of similarity of alignment operations between these DNA sequences and infer relations of similarities between species of the same genus.

12.4.1.2 Experimental Results for a Set of Species of the Phylum Firmicutes

We analyzed the performance results of our approach on a set made of 33 different species. These species are from different genera but are in the same phylum. All the species used are species of the genera: Alicyclobacillus, Anoxybacillus, Bacillus, Geobacillus, Lactobacillus, Lysinibacillus, Paenibacillus, Sporosarcina. These experiments would determine similarity relationships between the genera in terms of execution time, similarity percentage, and if our approach could build a hierarchical classification according to the taxonomic classification of species.

12.4.1.3 Experimental Results for Random Species

In this part, we analyzed the experiments conducted on any specie regardless of any hierarchical classification. The number of species used to make the experiments was 500 species.

12.4.2 Results in Term of Percentages of Similarity

12.4.2.1 Sequences from the Same Genus

We note that the percentages of similarity of our approach are higher than those of the Smith-Waterman algorithm. For lengths equal to or less than 500 base pairs, the percentage similarities are higher than or equal to 95 %. These similarity percentages describe, at best, in this case, regions with high similarity between species of the same genus (Fig. 12.1).

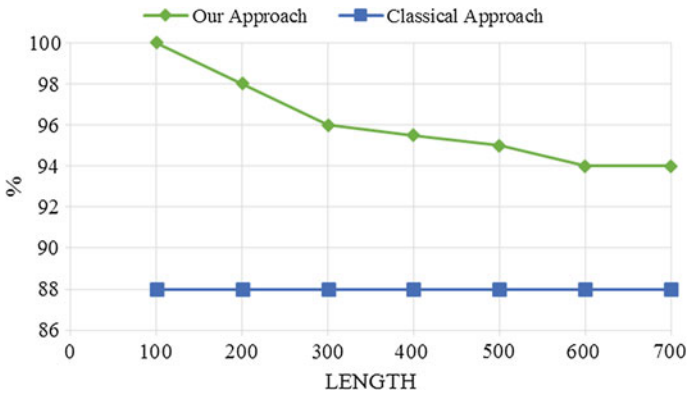


Fig. 12.1 Experiments on sequences from same gender

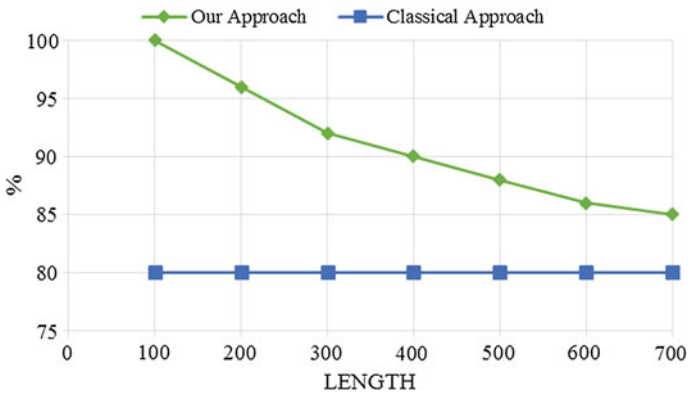


Fig. 12.2 Experiments on sequences from same phylum

12.4.2.2 Sequences from the Same Phylum

The percentages of similarity of our approach reach 92 % for a length of 300 base pairs. Still for shorter lengths, percentages of similarity are around 88 %. These similarity percentages are by 7 % better than the algorithm of Smith and Waterman (Fig. 12.2).

12.4.2.3 Sequences from Random Species

We note that the percentages of similarity have significantly decreased compared to our previous experiments. Indeed, the species used are not similar in taxonomic classification. We also note that for small lengths, lower than 400, the percentages of similarity are higher than 70 %. While for longer lengths, percentages are around 64 % but remain higher than those of the Smith-Waterman algorithm which is less than 57 % (Fig. 12.3).

12.4.3 Experiments in Terms of Execution Time

In this section, we will present a comparative study between our approach and the classical approach of determining an optimal alignment in terms of execution time.

The difference in execution time between the two algorithms is not very important and does not exceed, in the worst case, 1 min and 30 s for Tests with 4000 constructed alignments. This similarity in execution time favors, at most, the use of our optimal approach (Fig. 12.4).

We can therefore conclude that it is desirable to use our approach that determines optimal local alignment not only because it has a higher similarity percentage compared to the other two approaches, but also because in terms of execution time, the difference between the approaches is not quite significant (Fig. 12.5).

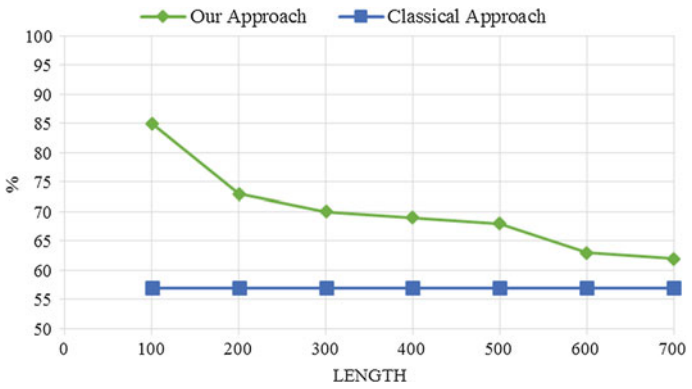


Fig. 12.3 Experiments on random sequences

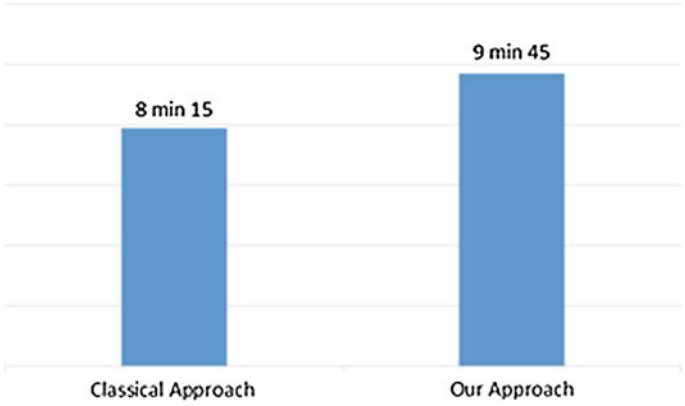


Fig. 12.4 Execution time of our approach for 4000 constructed alignments

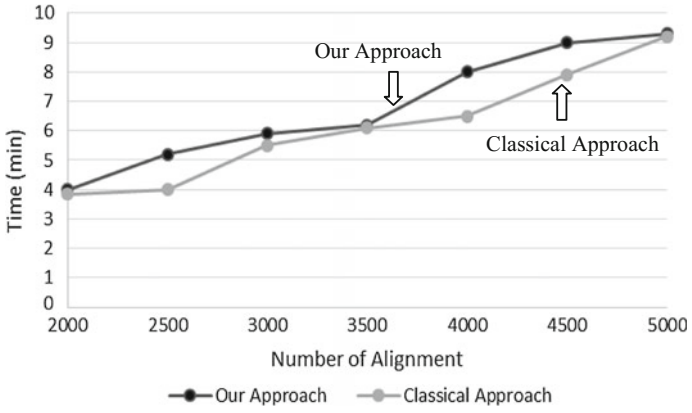


Fig. 12.5 Execution time of our approach and the Smith and Waterman algorithm

12.5 Conclusion and Future Work

Our optimal approach allows researchers to find a sub string called representative of a given DNA sequence [6]. The percentages of similarity are better than the algorithm of Smith and Waterman, and reach 100 % for small lengths. Our approach, then, allows them to reduce the amount of information stored in their databases. This considerable reduction in the size of DNA sequence alignments can reduce the size of databases by a factor of 2.

Nevertheless, we try to do other research in this area to:

- **Propose an algorithm for the compression of DNA sequences representation**: as in the networks, we will try to develop an algorithm for compressing DNA sequences information and reduce its representation, which will reduce the size of the data in databases.
- **Represent a set of DNA sequences by a unique string**: based on our approach, we will try to group multiple species and represent them by a unique string.
- **Find a new representation of DNA information**: it is true that our contribution proposes a decrease in the size of the sequences alignment comparison. Thus, the analysis and the treatment of large number of sequences present big challenges to biologists. We may have a new representation of DNA sequences.

Acknowledgements This work was funded by the International Exchange Program of Harbin Engineering University for Innovation-oriented Talents Cultivation.

References

1. Needleman SB, Wunsch CDA (1970) General method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 48:443–453
2. Smith TF, Waterman MS (1981) Identification of common molecular subsequences. J Mol Biol 147:195–197
3. Genbank. http://www.ncbi.nlm.nih.gov/genbank/
4. Saada B, Zhang J (2015) Vertical DNA sequences compression algorithm based on hexadecimal representation. Lecture notes in engineering and computer science: proceedings of the world congress on engineering and computer science 2015, WCECS 2015, 21–23 Oct 2015, San Francisco, USA, pp 570–574
5. Saada B, Zhang J (2015) DNA sequences compression algorithm based on extended-ASCII representation. Lecture notes in engineering and computer science: proceedings of the world congress on engineering and computer science 2015, WCECS 2015, 21–23 Oct 2015, San Francisco, USA, pp 556–560
6. Saada B, Zhang J (2015) Representation of a DNA sequence by a subchain of its genetic information. Lecture notes in engineering and computer science: proceedings of the world congress on engineering and computer science 2015, WCECS 2015, 21–23 Oct 2015, San Francisco, USA, pp 536–540

Chapter 13
Fatal Flaws in Norwich's "Mystery of Loudness Adaptation"

Lance Nizami

13.1 Introduction

Professor K.H. Norwich offers us "A mathematical exploration of the mystery of loudness adaptation" [1]. The "mystery" involves "Simultaneous Dichotic Loudness Balance", abbreviated SDLB. In SDLB, a procedure done in the psychology laboratory, one of the listener's ears (Dr. Norwich calls it the "adapting ear") is exposed to a continuous single-frequency tone, while the opposite ear (the "control ear") is intermittently exposed to a shorter same-frequency tone. The listener adjusts the control-ear tone's physical magnitude, within sessions of 5–20 s (the particular value varies by study), until it is as loud as the adapting-ear tone. This is "loudness balance". Generally, a tone's loudness increases with its magnitude.

Figure 13.1 shows the listening conditions [2]. "Stickman" sits within a soundproof chamber. Electrical cables (dashed lines) extend to headphones from outside of the chamber. Here, Stickman's right-ear headphone is the "control" headphone, whose cable passes through an attenuator, which Stickman can adjust to alter control-tone magnitude (and hence loudness).

Figure 13.2 shows the time-schedule of the tones, in which "The on-off markers do not show the variations in intensity of the comparison [i.e., control] stimulus [here, a tone] during a loudness balance" [3]. Over successive sessions of loudness-balance, the magnitude of the control-ear tone which is required for loudness-balance decreases. Figure 13.3 shows this change. The change is traditionally assumed [2] to indicate diminishing loudness of the tone in the *adapting* ear, "adaptation". However, when a tone is given to only *one* ear, intermittently or continuously, the tone's loudness does *not* adapt. Sixty years of work has confirmed these phenomena.

L. Nizami (✉)
Independent Research Scholar, Palo Alto, CA 94306, USA
e-mail: nizamii2@att.net

S.-I. Ao et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-981-10-2717-8_13

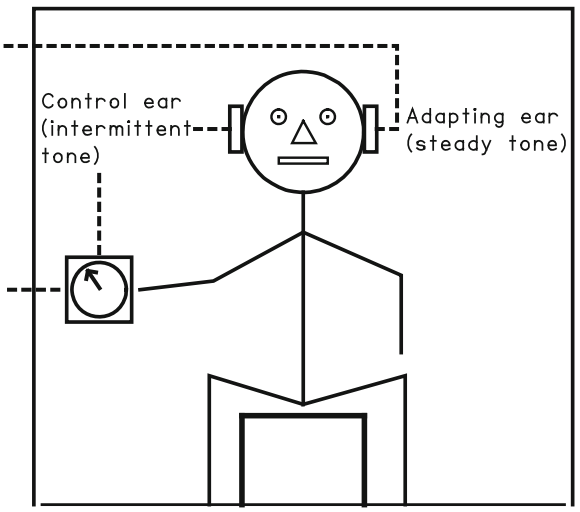


Fig. 13.1 The listener in the soundproof chamber

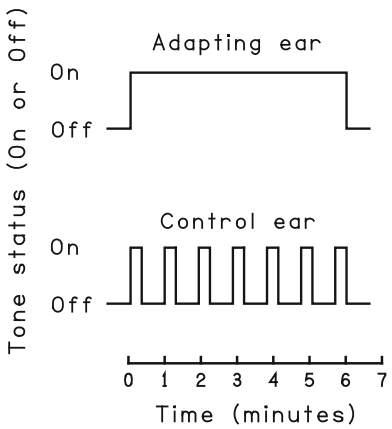


Fig. 13.2 A typical schedule of tone presentation during SDLB

There is a contradiction, then. Norwich [1] calls this "the mystery of loudness adaptation". He proposes to solve it, and his method is to "analyze mathematically" [1, p. 298] the monaural (one ear) and binaural (two ears) loudness adaptation. Unfortunately, Norwich's analysis proves to contain considerable confusion. He makes one particularly egregious error, namely, to assume that two different unitless measures are mathematically relatable merely because they are unitless. Such errors can cause havoc; unitless quantities are ubiquitous in science. Let us examine Norwich's analyses.

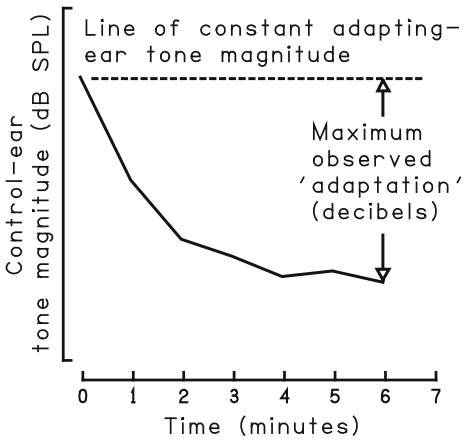


Fig. 13.3 For the tones of Fig. 13.2. The vertical axis shows the magnitude of a control-ear tone of equal loudness to the adapting-ear tone [3]. That adapting-ear tone has constant magnitude (horizontal dashed line). The vertical axis is unnumbered, because adaptation (the drop in decibels below the *dashed line*) can vary from 0 to 50

13.2 Two "Loudness Laws" into One

13.2.1 The "Entropy Equation" for Loudness Growth

Professor Norwich begins with loudness growth as a function of sound-pressure-level (SPL). This growth has traditionally been described by either the Weber-Fechner Law, $y = a \ln x + b$, or Stevens' Law, $y = ax^n$. These are coarse approximations. Norwich nonetheless calls them "forms of the *psychophysical law*" ([1, p. 300]; original italics), so that, to Norwich, "the two laws can be merged into a single loudness law" [1, p. 300].

Norwich now denotes L for loudness, and γ for "a parameter whose value is greater than zero" [1, p. 300]. He denotes φ for the magnitude of a tone. He then states that

$$L = \ln (1 + \gamma \varphi^n). \tag{13.1}$$

As sources for this equation, Norwich cites two of his publications [4, 5]. But the respective equation, also found in at least 45 of Norwich's publications on his Entropy Theory of Perception (1975—present), is the "Entropy Equation". It differs from Eq. (13.1), by containing a parameter k as a multiplier at the immediate right-hand side of the equals sign. k's magnitude is not pre-determined, but is nonetheless said to be crucial [4, 5]; hence it *cannot* be arbitrarily set to unity. But Norwich fails to explain its absence from Eq. (13.1).

Norwich [1, p. 301] notes that "when $\gamma\varphi^n$ becomes substantially less than one" (i.e., when φ is "small", for fixed γ), then Eq. (13.1) "reduces to Stevens' law". (Consider the Taylor-series expansion of the logarithm.) Conversely, when $\gamma\varphi^n$ "is substantially greater than one" (i.e., when φ is "large", for fixed γ), a logarithmic

law ensues, which Norwich will eventually denote the Weber-Fechner law. This notion of two-laws-in-one is ubiquitous within Norwich's Entropy Theory. But generally, the Weber-Fechner law and Stevens' law respectively refer to logarithmic or power equations applied to *the entire dynamic pressure range of hearing* [6], not to some portion of it. Professor Norwich seems willing to bend definitions, and this will continue.

13.2.2 Does Loudness Actually Follow the Entropy Equation?

Of course, the tone magnitude for which "$\gamma\varphi^n$ becomes substantially less than one" [1, p. 301] is something that Dr. Norwich does not know *a priori*. That is, the values of the Entropy Equation's parameters can only be obtained by regression-fitting that equation to plots of loudness versus tone magnitude. Norwich admits this [4, 5]. However, he and his colleagues have had more than 40 years [7] of opportunities to handle hundreds of others' published data-sets, but only show a handful of regression-fits. Without *fitted* parameters, any approximation to Eq. (13.1) is speculation.

Along this line of reasoning, the present author has evaluated numerous examples of the fit of the Entropy Equation to loudness estimates from the peer-reviewed literature [6]—including the loudnesses from Luce and Mo [8], which Norwich [1] displays as evidence that loudness growth is described (and with "near-equivalence"; [1], p. 300) by "Stevens' law" and the "Weber-Fechner law". However, the present author's evaluations [6] reveal that (1) the Entropy Equation may reduce to a power law for most of a loudness plot, with only the *highest* loudnesses following the actual Entropy Equation; or that (2) the Entropy Equation may reduce to a logarithmic law for most of a loudness plot, with only the *lowest* loudnesses following the Entropy Equation; or finally that (3) the Entropy Equation may apply to most of a loudness plot, reducing to a power law at the lowest loudnesses and reducing to a logarithmic law at the highest loudnesses. Altogether, then, loudness growth is adequately approximated without need for an Entropy Equation. And, in seeming confirmation, Norwich will soon focus on the "Weber-Fechner law", but without explaining why.

But it is now necessary to address another issue that relates to equations, an issue that is clearly fundamental to Norwich [1] at both the earlier and later points in his arguments. That issue is the *units* of the tone magnitude. Such units can normally be taken for granted, but not here.

13.3 The Units of Tone Magnitude

13.3.1 "Decibels Sensation Level"

Let us return, as Dr. Norwich does, to the equation presently called Eq. (13.1). Norwich continues: "γ is dimensionless, since φ has been defined as a pure number" [1, p. 300]. Norwich does not explain the term "pure number". Formally, it is a number having no physical dimensions (i.e., meters, kilograms, or seconds).

To comprehend γ and φ requires going back to Norwich's p. 298. There, he supplies a list of Abbreviations. He includes "dB SL", which he defines as "decibels sensation level; 10 $\log_{10}\varphi$". This a problem. Traditionally,

$$\begin{aligned} &dB\,SL\,(decibels\,sensation\,level) \\ &\quad = 20\ \log_{10}(p/p_0),\ for\ p \geq p_0 \end{aligned} \tag{13.2}$$

(e.g., [9]), where p is the root-mean-square sound-pressure-level of a tone. $p = p_0$ when just-detectable by the listener (here, Stickman; Fig. 13.1). Note well that Norwich uses "10" rather than "20" in 10 $\log_{10}\varphi$, and that he apparently sets $p_0 = 1$; once again, Norwich has arbitrarily set a parameter equal to unity. But we may express tone magnitude by intensity I rather than by pressure p; the square of pressure is proportional to intensity [9]. Using I_0 rather than p_0,

$$\begin{aligned} &dB\,SL\,(decibels\,sensation\,level) \\ &\quad = 10\ \log_{10}(I/I_0),\ for\ I \geq I_0. \end{aligned} \tag{13.3}$$

This may explain the "10" in Norwich's 10 $\log_{10}\varphi$. If so, $\varphi = I/I_0$. Thus φ would indeed be a pure number, with φ=1 at 0 dB SL, the just-detectable tone magnitude. These issues of definition will prove crucial.

13.3.2 Loudness-Equation Parameters

In [1] p. 302, Dr. Norwich states that "the decibel measures in the present study (e.g. A and B) will be expressed in units of sensation level, or dB SL (that is, dB SL = 10 $\log_{10}\varphi$)". This is the first appearance of the terms A and B, but they are not elaborated-on until Norwich's p. 305 (see below). Regarding "units of sensation level", Norwich here uses "units" for "dimensions", a usage that will be presently continued. However, on Norwich's next line [1, p. 302], he declares that "the units of sound intensity, φ, may be either units of pressure (newton m^{-2}), or units of power (watt.m^{-2}) [*sic*]". Read this carefully; it means that φ now has physical units, and is therefore no longer a pure number.

Henceforth, φ apparently keeps physical units; on Norwich's p. 309, for example, he speaks of "φ (in units of sound pressure)". This change is important, as follows. Note well that in Eq. (13.1), $\gamma\varphi^n$ is added to the number 1, which is not assigned units in Norwich's publications. Therefore, $\gamma\varphi^n$ must likewise be unitless. If so, then γ must have the same units as $1/\varphi^n$; however, the latter is not itself unitless if φ has units of sound pressure or sound intensity. Nonetheless, "γ is dimensionless, since φ has been defined as a pure number" [1, p. 300]. This is all very odd.

And from this point onwards, it remains unclear what the units of γ really are. This is disturbing, given the following declaration: "The parameter, γ, will be seen to govern loudness adaptation" [1, p. 301]. But adaptation is broadly defined as change (usually a monotonic decrease) happening over a period of time, whereas Norwich's paper [1] proves no time-dependence for γ. It might occur elsewhere in his Entropy Theory of Perception (1975—present), but that would be irrelevant; it is not proven in [1].

13.4 Algebra for Loudness-Matching

Recall the SDLB procedure. The listener adjusts the loudness of the control-ear tone to equal the loudness of the adapting-ear tone. Dr. Norwich declares [1, p. 303] that "Clearly, the [SDLB] matching process can be described by"

$$L_{control} = L_{adapting}. \tag{13.4}$$

He further states that "Fechner's law [*sic*]" applies to "the incompletely adapted control ear which has larger values of $\gamma\varphi^n$" [1, p. 303]. He then states that "Stevens' law" applies to "the adapting ear which has smaller values of $\gamma\varphi^n$" [1, p. 304]. But there is a problem. If Eq. (13.4) holds (with loudness L being given by Eq. (13.1)), then the two ears *cannot* have different values of $\gamma\varphi^n$ during SDLB. This contradicts Norwich's statement that the incompletely adapted control ear has larger values of $\gamma\varphi^n$. Also, the Weber-Fechner logarithmic law is not the same thing as "Fechner's law", which is an approximation *underlying* the Weber-Fechner law.

13.5 Algebra for Loudness Adaptation

13.5.1 Loudness: Its Alleged Time-Dependence

Let us continue to follow Dr. Norwich's explanations: "The reason for assigning Fechner's law [*sic*] in one case and Stevens' law in the other is because the parameter, γ, decreases with adaptation" [1, p. 304]. But, as noted above, γ was not proven to be time-dependent. Continuing with Norwich [1, p. 304]: "Both the

adapting ear *and* the control ear adapt to the tones introduced to them (although we expect the latter to adapt less, since it is stimulated for a shorter time)" (original italics). Altogether, Dr. Norwich implies that the (alleged) time-dependence of γ is identical for both ears. What, then, of φ? By design, it remains constant in the *adapting* ear during adaptation. But its magnitude in the *control* ear, as set by Stickman to achieve loudness balance (Fig. 13.1), empirically decreases over time.

Hence, if (for example) γ decreases equally quickly in both ears, and if (after Norwich) Eq. (13.1) applies separately to each ear, then during SDLB the control ear's contribution to loudness will always be lower than the adapting ear's contribution. Parameters-wise, $\gamma\varphi^n$ (Eq. (13.1)) will differ at each ear. However, as explained above, the two ears *cannot* have different values of $\gamma\varphi^n$ when loudnesses are matched. Hence, γ *could not* decrease equally quickly in both ears (assuming that it can change at all).

The one remaining parameter of Eq. (13.1), n, does not contribute to adaptation, because it remains constant [1, p. 304].

13.5.2 Loudness as a Logarithmic Equation: The Heart of Norwich's Argument

At this point [1, p. 304], Norwich reaches the algebraic and conceptual heart of his argument. He re-introduces the Weber-Fechner law, as

$$\begin{aligned} L &= \ln(\gamma\varphi^n) = (\ln 10)\log_{10}(\gamma\varphi^n) \\ &= (\ln 10)[\log_{10}(\gamma) + n\log_{10}(\varphi)]. \end{aligned} \tag{13.5}$$

Norwich has hitherto implied that this particular equation applies to the *control* ear. Henceforth, however, he will refer to *either* ear, or to *both* ears. Regardless, Norwich rationalizes the Weber-Fechner law, as follows: "It will be useful" [1, p. 304]. Is that all? Norwich further explains: "We shall make use of the experimental data measured by Jerger (1957)" (here as ref. [10]). Norwich does not say why he uses such old data. A great mass of newer data is available.

Norwich [1, p. 303] makes plots of sets of Jerger's data, and claims that they are "corroborated" by Weiler et al. [11] and by D'Alessandro and Norwich [12]. Norwich's plots [1] show adaptation in decibels, *which is not a loudness measure* (see Fig. 13.3), versus sensation level in decibels (dB SL) at the *adapting* ear, for a variety of Jerger's tone waveform frequencies. The data-points very roughly follow straight lines. But how does this justify Eq. (13.5)? The latter expresses *loudness* dependence as a linear equation in dB SL, if $10\log_{10}\varphi$ is indeed dB SL.

13.5.3 Adaptation to a Fixed-Intensity Auditory Stimulus

Norwich continues [1, p. 304]: "Our interest initially is in the extent of adaptation produced by an auditory stimulus of fixed intensity, φ_b, corresponding to a steady tone at the adapting ear". During a steady tone, intensity φ_b and the exponent n remain constant. Norwich states that "γ is the parameter that encodes the process of adaptation" [1, p. 304]. He explains: "We shall deal with γ only at two distinct times: at $t = 0$, before any adaptation has taken place, it will have the constant value γ_0; and at large times it will have the smaller value γ, which will govern the maximum degree of adaptation" [1, p. 304]. "Large times" is unexplained. Norwich then assigns L_0 to γ_0 and L to γ, and then subtracts L from the larger term L_0. That is,

$$\begin{aligned}
&(10/\ln 10)L_0 - (10/\ln 10)L_\infty \\
&\quad = (10/\ln 10)(\ln 10)\left[\log_{10}\left(\gamma_0\varphi_b^n\right) - \log_{10}\left(\gamma_\infty\varphi_b^n\right)\right] \\
&\quad = 10\,\left[\log_{10}\gamma_0 - \log_{10}\gamma_\infty\right] \\
&\quad = 10\log_{10}(\gamma_0/\gamma_\infty) \\
&\quad = \text{decibels of adaptation.}
\end{aligned} \tag{13.6}$$

"Decibels of adaptation" is Dr. Norwich's own unique interpretation. Dr. Norwich further states that "decibels of adaptation can also be expressed as the difference between the basal sound level of the sustained tone administered to the adapting ear and the balancing sound level of the adjusted tone at the control ear" [1, p. 305]. That is (Norwich's Eq. (12)),

$$10\,\log_{10}(\gamma_0/\gamma_\infty) = B - A. \tag{13.7}$$

Here $B - A$ is described as "the laboratory measure" of "the magnitude of adaptation to any steady tone" [1, p. 305]. The term B is the dB SL of the "sustained tone" [1, p. 305]. The term A, in contrast, is not explicitly defined at this juncture. It seems to be the final magnitude set by Stickman for the control-ear loudness-matching tone, as adaptation reaches asymptote.

In Eq. (13.6), $(10/\ln 10)L_0 - (10/\ln 10)L_\infty$ is equated to $10\,[\log_{10}\gamma_0 - \log_{10}\gamma_\infty]$. This depends upon φ_b^n being in both L_0 and L (from Eq. (13.5)). This, in turn, means that the present Eqs. (13.6) and (13.7) either (a) refer to the adapting ear, for which $\varphi = \varphi_b$ even at $t = \infty$, or (b) refer to both ears, but at $t = 0$, i.e., before any possible adaptation! Neither (a) nor (b) is logical; the tone sensation-level B is at the adapting ear, and the tone sensation-level A is at the control ear. Note now what Eq. (13.7) actually means: Norwich takes two things lacking physical units—a difference in levels of a sensation (loudness difference) and a sensation-level difference (a logarithm of ratios of tone magnitudes)—and simply declares them to be equal. Perhaps he has problems comprehending the differences.

13.6 Why no Monaural Adaptation?

Professor Norwich's paper now momentarily transitions to several pages of derivatives [1, pp. 305–307], which prove to be aimless and unsuitable for synopsis. Norwich then declares that $10 \log_{10} \gamma_0 - 10 \log_{10} \gamma_\infty$ (which is actually the left-hand side of the present Eq. (13.7)) "is equal to the difference in sound level at the adapting ear" [1, p. 311]. "Difference" is an unusual word, because the sound level at the adapting ear is φ_b, which is a constant, as Norwich already notes [1, p. 305]. Regardless, Norwich then declares that $B - A$ (the right-hand side of Eq. (13.7)) is "the difference between the sound level of the adapting tone (applied to the adapting ear) and the sound level of the matching tone (applied to the control ear)" [1, p. 311].

Any confusion, by this point, can hardly be blamed on the reader. To review: Norwich says that the left-hand side of Eq. (13.7) applies to the adapting ear, whereas the right-hand side of Eq. (13.7) depends upon *both* ears. If loudness adaptation was not a mystery before, it certainly is now. To continue: Dr. Norwich [1, p. 311] then states that

> At zero time—i.e. at the instant the adapting tone was begun – there was, of course, no adaptation whatever. If a balancing tone could have been invoked at the control ear at zero time, this tone would also have carried the sound level *B*. *Therefore, let us regard the quantity B – A on the right-hand side as representing the difference in balancing tone level at the control ear, between zero and some large time* (5 to 7 minutes).

(Original italics from Norwich.) In other words, the right-hand side of Eq. (13.7) now describes the *control* ear.

Norwich then declares that $A = B = 0$ for the "special case" [1, p. 312] of no control-ear tone, "since the control ear is always maintained in silence" [1, p. 312]. That is, now the adapting ear is the only one to have received a tone. But Norwich's declaration that $A = B = 0$ is hence illogical; if B is indeed the magnitude of the steady tone at the adapting ear, then B cannot be set to zero when an adapting tone is present! Also, A is the final magnitude set by Stickman for the control-ear tone; that is, *the term A is not independent of SDLB*. Overall, then, there is *no "special case" of no control-ear tone*.

Norwich then notes [1, pp. 311–312] that if $A = B = 0$, then $10 \log_{10} \gamma_0 - 10 \log_{10} \gamma_\infty = 0$ and hence $\gamma_\infty = \gamma_0$, altogether (to Norwich) "implying that no adaptation takes place, which resolves the paradox mathematically" [1, p. 312]. This begs the question of how a *physical* paradox can be resolved *mathematically*. Where, for example, are the necessary physical insights, obtained through measurement?

13.7 Summary: Norwich's "Mystery of Loudness Adaptation"

Professor K.H. Norwich describes what he calls the "mystery of loudness adaptation" [1]. The "mystery" concerns Simultaneous Dichotic Loudness Balance (SDLB), a laboratory procedure in which a listener adjusts the loudness of an intermittent single-frequency tone played to one ear (the "control ear") to match the loudness of a steady same-frequency tone played to the other ear (the "adapting ear"). Empirically, the loudness-matching tone-magnitude in the control ear needs to be decreased over time. This implies that the loudness at the "adapting ear" adapts. However, when there is no tone to the control ear, only the steady tone to the adapting ear, the latter tone's loudness does not adapt. Why?

Dr. Norwich offers an answer. He starts by assuming that the empirical growth in loudness with tone magnitude follows the Entropy Equation from his publications. However, for reasons unexplained, he omits one of the usual three parameters of his equation, called k, leaving the two remaining parameters, γ and n. The equation nonetheless has one variable, the "sound intensity" φ. Norwich assumes that the equation splits into two cases, a power equation ("Stevens' law") for small tone magnitudes, and a logarithmic equation ("the Weber-Fechner law") for large tone magnitudes.

The logarithmic equation forms the core of Norwich's mathematical interpretation of SDLB. Initially, though, Norwich discusses another issue crucial to his model, namely, the units of measurement of the "sound intensity", φ. He states that φ is unitless. He then contradicts himself, by implying that φ has physical units of tone magnitude. Regardless, within Norwich's equation the units of φ are actually inseparable from the units of the parameter γ. The parameter γ is crucial to Norwich, insofar as he states that it governs adaptation. However, he provides no proof that γ involves *time*, making it difficult to imagine how a time-dependent process, such as adaptation, would depend upon γ.

Norwich then notes that, when the loudness in the control ear is matched to the loudness in the adapting ear, his equation should give the same loudness for the two ears. Norwich then states that "the Weber-Fechner law" applies to the control ear, and that "Stevens' law" applies to the adapting ear. This implies altogether *different* loudness values at the two ears. The different "laws" are allegedly due to a decrease in γ during adaptation—a decrease in γ which has *equal* speed in both ears.

At this juncture, Norwich adopts "the Weber-Fechner law" entirely. He rationalizes his choice by plotting some data from the literature, but his analysis transpires to justify nothing. Norwich then takes the loudness at the start of the adapting tone, and a later loudness, the matching loudness in the control ear when such loudness has reached an asymptote, and subtracts that later loudness from the starting loudness. This quantity Norwich sets directly proportional to the empirical magnitude of adaptation (*in decibels*) over the time-interval in question.

Professor Norwich then declares that the theoretical loudness-difference describes the adapting ear, whereas the empirical adaptation depends upon both of the ears, but *then* depends upon the control ear. All of this is truly extraordinary.

Dr. Norwich finally reveals his resolution to the "mystery of loudness adaptation": that the observed constancy of loudness of an ongoing tone when given only to one ear (adapting ear) is a special case of SDLB, a case in which the control (i.e., other) ear receives no stimulus.

13.8 Discussion: Norwich's Use of Units

Professor Norwich's paper contains numerous flaws. Indeed, there are so many that it is difficult to comprehend how the manuscript passed peer review at an established mathematics journal (*Bulletin of Mathematical Biology*; Editor Philip Kumar Maini). There is nonetheless one particular error that stands out as fatal: Norwich sets a loudness difference, which lacks physical units, equal to a *sensation-level* difference, which lacks physical units but is nonetheless not equivalent to a loudness difference. Things cannot be equated simply because they lack physical units. Do two apples equal two oranges? Norwich may think so.

The correctness of units is crucial, but is usually taken for granted. This allows great potential for abuse. Further remarks may therefore be justified. It transpires that units have been used incorrectly elsewhere by Norwich. For example, he claims to derive the Stevens' Law exponent for the sense of taste directly from principles of thermodynamics [13]. However, he uses quantities having the wrong units, thereby deriving a hypothetical Stevens'-law exponent which is *negative* (see [14]) —in contrast to all empirical Stevens'-law exponents, which are positive.

That's not all. McConville et al. [15] use Norwich's Entropy Equation to mathematicize two-interval forced-choice (2IFC) auditory intensity discrimination, a well-known laboratory procedure. In particular, they derive an equation in five unknowns for the Weber fraction, the just-noticeable-increase in tone intensity divided by its base intensity, which altogether is unitless. McConville et al. list their unknowns (their Table 1), describing the units of an unknown called β as s^{-1} (i.e., inverse-seconds). Using t for "time", they list β/t as unitless. Further, the McConville et al. [15] equations also include the same n used in [1], said to be Stevens' exponent. The McConville et al. n is obliged to be unitless, to make their Weber fraction unitless [16]. The units of β must therefore be (time) $\cdot$ (intensity)$^{-n}$, hence β/t has units of (intensity)$^{-n}$, hardly unitless.

13.9 Conclusions

In "A mathematical exploration of the mystery of loudness adaptation" [1], Professor K.H. Norwich offers a *mathematical* resolution to a conundrum: that the loudness of a tone given to just one ear (i.e., by headphones) will not decrease, whereas that tone's loudness *will* apparently decrease when matched by the loudness of a same-frequency tone presented intermittently to the other ear ("loudness adaptation"). However, Norwich's "mathematical exploration" proves to be a mass of inconsistencies and unsupportable stipulations, through which he fails to explain the "mystery of loudness adaptation". His arguments look confused, perhaps circular. Indeed, scrutiny elsewhere [6, 14, 16–20] reveals circular logic throughout Norwich's previous work on his Entropy Theory of Perception, the evident source of his starting equations. Norwich's closing proposition, that the observed constancy of loudness of an ongoing tone when given only to one ear is a special case of SDLB, is absurd.

The flaws of Dr. Norwich's model were first identified elsewhere by the present author [21], but they deserve a second (and clearer) explication, as a warning against attempts to do mathematical biology as "mathematical explorations" lacking adequate biology. The missing biology can in fact be identified, and the "mystery of loudness adaptation" can then be solved, in an alternative model which is phenomena-based rather than algebra-based [21]; for more depth see [22]. The alternative model accounts for something that Norwich does not, namely, the interaction of auditory physiology with the experimental-psychology protocol in SDLB. That is, the alternative model recognizes the likely role of the auditory efferent feedback system, a system that has been known for 65 years [23]. There, an auditory stimulus in either ear affects the other ear. When such interaction is recognized, ten established results of SDLB can be explained, one of them the "mystery of loudness adaptation". Oddly, Norwich (MD, PhD) is unaware of auditory efferents [24]:

> Neuroanatomical pathways from sensory receptors through the spinal cord, brain stem, midbrain to the cortex of the brain are well known for the senses of taste (tongue to cortex), hearing (cochlea to cortex) etc. However, these pathways are purely afferent in nature.

Acknowledgements My thanks to Claire S. Barnes PhD for her valuable insights.

References

1. Norwich KH (2010) A mathematical exploration of the mystery of loudness adaptation. Bull Math Biol 72:298–313
2. Hood JD (1950) Per-stimulatory fatigue. Acta Oto-Laryngologica Supplementum 92:26–57
3. Egan JP (1955) Perstimulatory fatigue as measured by heterophonic loudness balances. J Acoust Soc Am 27:111–120

4. Norwich KH (1977) On the information received by sensory receptors. Bull Math Biol 39:453–461
5. Norwich KH (1993) Information, Sensation, and Perception. Academic Press, Toronto
6. Nizami L (2009) A computational test of the information-theory based Entropy Theory of Perception: does it actually generate the Stevens and Weber-Fechner Laws of sensation? Lecture Notes in Engineering and Computer Science: Proceedings of the World Congress on Engineering 2009, WCE 2009, London, UK, pp 1853–1858
7. Norwich KH (1975) Information, memory, and perception. Institute of Biomedical Engineering University of Toronto, Vol. 17
8. Luce RD, Mo SS (1965) Magnitude estimation of heaviness and loudness by individual subjects: a test of a probabilistic response theory. Br J Math Stat Psychol 18:159–174
9. Hartmann WM (1998) Signals, Sound, and Sensation. Springer, New York
10. Jerger JF (1957) Auditory adaptation. J Acoust Soc Am 29:357–363
11. Weiler EM, Loeb M, Alluisi EA (1972) Auditory adaptation and its relationship to a model for loudness. J Acoust Soc Am 51:638–643
12. D'Alessandro LM, Norwich KH (2009) Loudness adaptation measured by the simultaneous dichotic loudness balance technique differs between genders. Hear Res 247:122–127
13. Norwich KH (1984) The psychophysics of taste from the entropy of the stimulus. Percept Psychophys 35:269–278
14. Nizami L (2010) Fundamental flaws in the derivation of Stevens' Law for taste within Norwich's Entropy Theory of Perception. In: Korsunsky AM (ed) Current themes in engineering science 2009: selected presentations at the world congress on engineering-2009 (AIP Conference Proceedings 1220). American Institute of Physics, Melville, NY, USA, pp 150–164
15. McConville KMV, Norwich KH, Abel SM (1991) Application of the entropy theory of perception to auditory intensity discrimination. Int J Biomed Comput 27:157–173
16. Nizami L (2008) Is auditory intensity discrimination a comparison of entropy changes? In: Proceedings of the 155th Meeting of the Acoustical Society of America, 5th Forum Acusticum of the EA, 9e Congrès Français d'Acoustique of the SFA, 2nd ASA-EAA Joint Conference, Paris, France, 2008, pp 5739–5744
17. Nizami L (2008) Does Norwich's Entropy Theory of Perception avoid the use of mechanisms, as required of an information-theoretic model of auditory primary-afferent firing? Proceedings of the 155th Meeting of the Acoustical Society of America, 5th Forum Acusticum of the EA, 9e Congrès Français d'Acoustique of the SFA, 2nd ASA-EAA Joint Conference, Paris, France, 2008, pp 5745–5750
18. Nizami L (2009) On the hazards of deriving sensory laws from first-order cybernetics: Norwich's Entropy Theory of Perception does not derive the Weber fraction for audition. In: Proceedings of the 13th World Multi-Conference on Systemics, Cybernetics and Informatics (The 3rd International Symposium on Bio- and Medical Informatics and Cybernetics: BMIC 2009 of WMSCI 2009), Orlando, FL, USA, pp 235–241
19. Nizami L (2009) Sensory systems as cybernetic systems that require awareness of alternatives to interact with the world: analysis of the brain-receptor loop in Norwich's Entropy Theory of Perception. In: Proceedings of the 2009 IEEE International Conference on Systems, Man, and Cybernetics, San Antonio, TX, USA, pp 3477–3482
20. Nizami L (2011) A warning to the human-factors engineer: false derivations of Riesz's Weber fraction, Piéron's Law, and others within Norwich et al.'s Entropy Theory of Perception. In: Schmidt M (ed) Advances in Computer Science and Engineering, Rijeka, Croatia: InTech, pp 391–406
21. Nizami L (2015) On mathematical biology without the biology: a refutation of K.H. Norwich's Mystery of Loudness Adaptation and a physiology-based replacement. In: Lecture Notes in Engineering and Computer Science: Proceedings of the World Congress on Engineering and Computer Science 2015, WCECS 2015, San Francisco, CA, USA, pp 561–569

22. Nizami L (2012) Loudness 'fatigue' with two ears but not with one: simultaneous dichotic loudness balance (SDLB) explained. In: Fechner Day 2012: Proceedings of the 28th Annual Meeting of the International Society for Psychophysics, Ottawa, Ontario, Canada, pp 6–11
23. Portmann M-R (1952) Les Fibres Nerveuses Efférentes Cochléaires. Delmas, Bordeaux
24. Norwich KH (2011) Neurophysiological encoding of the apprehension of physiological events. In: Abstracts of Decision Making: Bridging Psychophysics and Neurophysiology, a workshop held at University of North Texas, Denton, TX, USA, March 18–20, 2011

Chapter 14
Wrong Limits and Wrong Assumptions: Kenny Norwich and Willy Wong Fail to Derive Equal-Loudness Contours

Lance Nizami

14.1 Introduction

What is a loudness contour? It is a plotted curve, which is obtained for a given person by having that research subject adjust the intensity of a "comparison" tone of a given waveform frequency until it seems as loud as a "reference" tone of a constant intensity and constant frequency. This method has been used in psychology laboratories for decades. For each different reference-tone intensity, a different loudness contour ensues.

Kenneth Howard (Kenny) Norwich and Willy Wong claim to derive equal-loudness contours mathematically, from theoretical first-principles [1, 2]. This is the only occurrence of such a feat, to the present author's knowledge. Further, full comprehension of what Norwich and Wong did requires reading *two* papers [1, 2], of which the more recent one [1] gives a backward derivation of what appears in the older one [2]. Indeed, further reading proves to be necessary [3–6]. All of the needed papers are examined here, to provide a synopsis that is not found in the literature. The synopsis reveals wrong limits and wrong assumptions underlying the Norwich and Wong derivations.

14.2 Norwich and Wong: A Loudness Equation

Norwich and Wong [1] introduce an "Entropy Equation", in which loudness is denoted L. Acoustical stimulus intensity, in units of power, is denoted I. Norwich and Wong ([1], Eq. (11)) write

L. Nizami (✉)
Independent Research Scholar, Palo Alto, CA 94306, USA
e-mail: nizamii2@att.net

S.-I. Ao et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-981-10-2717-8_14

$$L = \frac{k}{2} \ln \left(1 + \gamma \left(\frac{I}{I_{th}}\right)^n\right). \tag{14.1}$$

Investigation reveals that this is also Eq. (9) of [2], but that has k instead of k/2 (i.e., the ½ was presumably absorbed into "k"). "k" appears in earlier Norwich papers, as "a proportionality constant" [3], its value "determined by the arbitrary scale units of the experimenter" ([4], p. 269). Norwich and Wong assume that k is *independent of intensity and frequency*. The term I_{th} in Eq. (14.1) is the "threshold intensity". Norwich and Wong ([1], p. 931) now name L_{th}, "the loudness threshold" (actually the *threshold loudness*); from Eq. (14.1),

$$\begin{aligned} L(I_{th}) = L_{th} &= \frac{k}{2} \ln\left(1 + \gamma \left(\frac{I_{th}}{I_{th}}\right)^n\right) \\ &= \frac{k}{2} \ln (1 + \gamma). \end{aligned} \tag{14.2}$$

They then state that ([1], Eq. 12)

$$L = \begin{bmatrix} L - L_{th}, & L > L_{th} \\ 0, & otherwise \end{bmatrix}. \tag{14.3}$$

(This equation will be explored soon.) Altogether, from Eqs. (14.1) and (14.2) and $L = L - L_{th}$ in Eq. (14.3),

$$\begin{aligned} L &= \frac{k}{2} \ln\left(1 + \gamma \left(\frac{I}{I_{th}}\right)^n\right) - \frac{k}{2} \ln (1 + \gamma) \\ &= \frac{k}{2} \ln\left(\frac{1 + \gamma \left(\frac{I}{I_{th}}\right)^n}{1 + \gamma}\right) \quad for\ I > I_{th}. \end{aligned} \tag{14.4}$$

This appears in [1] as Eq. (13), with the γ misprinted as "y". Equation (14.4) also appears in [2], as Eq. (10) but with k instead of k/2.

14.3 Norwich and Wong: The "Weber Fraction"

Norwich and Wong [1] now differentiate Eq. (14.4). They get ([1], Eq. 14)

$$\frac{dL}{dI} = \frac{k}{2} \frac{n\gamma I^{n-1} I_{th}^{-n}}{1 + \gamma \left(\frac{I}{I_{th}}\right)^n} \tag{14.5}$$

which does not depend on which of Eq. (14.1) or Eq. (14.4) is used for loudness. Norwich and Wong [1] then reorder Eq. (14.5) to obtain dI/I. They then replace dL by ΔL, the difference limen (i.e., the subjectively just-noticeable difference) in loudness. Likewise, they replace dI by the intensity change corresponding to ΔL, namely, ΔI ([1], p. 932). Altogether,

$$\frac{\Delta I}{I} = \frac{2\Delta L}{nk}\left(1 + \frac{1}{\gamma}\left(\frac{I_{th}}{I}\right)^{n}\right), \tag{14.6}$$

which they name the Weber fraction ([1], Eq. (15)).

The *empirical* values of the Weber fraction had already been studied by many psychologists. Early-on, there was Riesz [7]. He mentioned an equation that "can be made to represent ΔE/E [his notation for ΔI/I] as a function of intensity at any frequency", by adjusting three equation parameters ([7], p. 873), each parameter itself being an empirical equation of frequency. Norwich and Wong [1] transformed Eq. (14.6) into the empirical equation used by Riesz [7], first by assuming (after Fechner [8]) that ΔL is constant with intensity, and then by defining two new positive quantities, S_∞ and S_0–S_∞:

$$S_\infty = \frac{2\Delta L}{nk}, \tag{14.7a}$$

$$S_0 - S_\infty = \frac{2\Delta L}{nk\gamma}. \tag{14.7b}$$

Hence ([1], Eq. (22))

$$\gamma = \frac{S_\infty}{S_0 - S_\infty}. \tag{14.7c}$$

Altogether, then,

$$\frac{\Delta I}{I} = S_\infty + (S_0 - S_\infty)\left(\frac{I_{th}}{I}\right)^{n} \tag{14.8}$$

where ΔI/I > 0. Equation (14.8) has the same general form as Eq. (2) of Riesz [7]. Equation (14.8) is also Eq. (16) of [1] and Eq. (1) of [2].

14.4 Norwich and Wong: Derivation of Equal-Loudness Contours

Norwich and Wong [1] then proceed to derive equal-loudness contours. First is the theoretical intensity of the comparison tone, as a function of its frequency, found by equating comparison-tone loudness to that of a constant-intensity 1 kHz reference tone. With the parameters of the reference tone denoted by ^ (caret),

$$L(I,f) = L\left(\hat{I},\ 1\ kHz\right) \tag{14.9}$$

(in [1] as Eq. (37)). Substituting appropriate terms from Eq. (14.4) gives

$$\frac{k}{2}\ln\left(\frac{1+\gamma\left(\frac{I}{I_{th}}\right)^{n}}{1+\gamma}\right) = \frac{\hat{k}}{2}\ln\left(\frac{1+\hat{\gamma}\left(\frac{\hat{I}}{\hat{I}_{th}}\right)^{\hat{n}}}{1+\hat{\gamma}}\right) \tag{14.10}$$

which is Eq. (12) of [2], but with k there instead of k/2, and with "prime" in place of "caret", and with a script I in place of caret-I. Equation (14.10) can be re-arranged to solve for I/I_{th}; ten times its logarithm to base 10 gives the intensity of the matching comparison tone in decibels *sensation level* (dB SL):

$$\begin{aligned} &10\log_{10}\left(\frac{I}{I_{th}}\right) \\ &= 10\log_{10}\left(\left(\frac{1}{\gamma}\right)^{\frac{1}{n}}\left[(1+\gamma)\left(\frac{1+\hat{\gamma}\left(\frac{\hat{I}}{\hat{I}_{th}}\right)^{\hat{n}}}{1+\hat{\gamma}}\right)^{\frac{\hat{k}}{k}} - 1\right]^{\frac{1}{n}}\right). \end{aligned} \tag{14.11}$$

It transpires that Eq. (13) of [2] is the term within the largest brackets here, *but* with k instead of k/2, and with "prime" instead of "caret", and with a script I in place of caret-I.

In Norwich and Wong [1], the frequency for the ^ parameters is taken to be 1 kHz. $\hat{I}_{th}$ is taken from Wegel [9]. S_0 and S_∞ are evaluated from Riesz [7], allowing values for γ (Eq. (14.7c)). k is assumed to be independent of frequency and intensity ([1], p. 931). k and $\hat{k}$ are hence assumed equal; $\hat{k}/k$ therefore disappears from Eq. (14.11).

The reference tone for making loudness contours is here 1 kHz. *Empirically*, each intensity of the 1 kHz tone would have an associated plot of the points {tone frequency, tone sensation level} that subjectively match the loudness of the 1 kHz reference tone. Equation (14.11) gives the respective *theoretical* equal-loudness contours; they appear in Fig. 5 of Norwich and Wong [1] and in Fig. 4 of Wong

and Norwich [2]. They are bowl-shaped, being lowest in-between the lowest and highest waveform frequencies.

Norwich and Wong note elsewhere [5] that Fletcher and Munson [10] had produced equal-loudness contours, likewise using an SL scale. Unlike the theoretical curves of Norwich and Wong [1] and of Wong and Norwich [2], however, the Fletcher and Munson [10] contours *increase* in-between the lowest and highest waveform frequencies, forming hill-shaped plots. Norwich and Wong [1] hence replaced Riesz's "n" by an equation for the Stevens exponent as a function of frequency. They employed ([1], p. 935) "a function similar to the one suggested by Marks" [11]. Their theoretical equal-loudness contours were now hill-shaped like those of Fletcher and Munson [10].

14.5 Examining Norwich and Wong (1): The Loudness Equation at the Loudness Threshold

Let us carefully scrutinize what Norwich and Wong [1] did. According to them, Eq. (14.4) above derives from Eq. (14.1) above. That is, loudness allegedly obeys *both* Eqs. (14.1) *and* (14.4). *But this is not possible*. First, note that the upper line of Eq. (14.3) should read $L \neq L–L_{th}$, not $L = L–L_{th}$. Now consider loudness limits. As $I \rightarrow I_{th}$ from $I > I_{th}$, $L \rightarrow L_{th}$ (following Eq. (14.1)); but according to Eq. (14.4), $L \rightarrow 0$. Thus, we have $L = 0$ *and* $L = L_{th}$ at the same intensity, I_{th}. Note also that as $I \rightarrow 0$ from $I > 0$, then after Eq. (14.1), $L \rightarrow 0$; but after Eq. (14.4), $L \rightarrow (–k/2)\ln(1 + \gamma)$, a *negative* loudness ($k, \gamma > 0$). Of course, a negative loudness is an impossibility.

14.6 Examining Norwich and Wong (2): The Constancy of K

Crucially, Norwich and Wong [1] assume that the parameter k is independent of intensity and frequency. *But Norwich and his co-authors never test that assumption, even up to the present day.* Therefore, it is tested here, as follows. k can only be obtained by fitting Eq. (14.1) or Eq. (14.4) to empirical plots showing loudness growth with intensity. To do so, first the term γ/I_{th}^{n} in Eq. (14.1) was replaced by a symbol γ'. Equation (14.1) was then fitted to thirty-seven loudness-intensity plots from the peer-reviewed experimental literature. Following that literature, Eq. (14.1) was put into logarithmic form in {*ln* I, *ln* L} coordinates, before fitting to logarithms of loudnesses.

Figure 14.1 shows the fitted values of k, versus the respective tone frequency, or for white noise. The empirical loudness-growths were taken from: [12], white noise (geometric means of magnitude estimates: series 1–3); [13], 1 kHz tone (Figs. 2, 3,

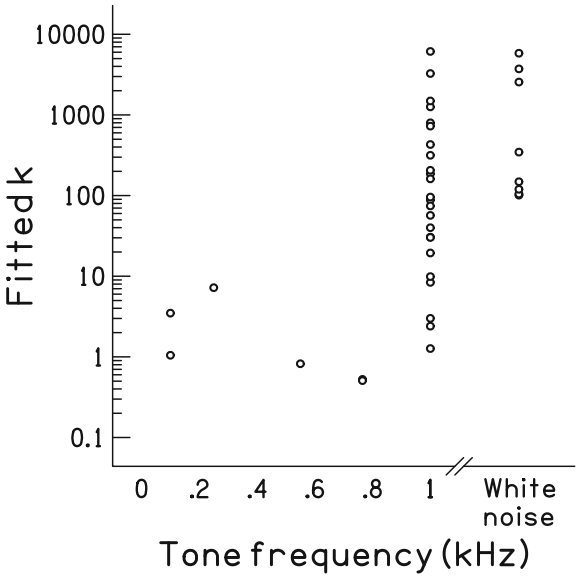


Fig. 14.1 The fitted value of the free parameter k of Eq. (14.1) as a function of tone frequency or for white noise (see text). The 37 dots represent more loudness-growth plots than in all of Norwich et al.'s relevant publications, from the very first in 1975 up to the present day

6, 7, 8, and 10); [14], 0.1 kHz tone (Fig. 2, crosses; Fig. 2, circles), 0.250 kHz tone (Fig. 3, geometric means of circles); [15], 1 kHz tone (subjects # 8, 9, 10, 11, 12, 13); [16], 1 kHz tone (curves 1–7); [17], 0.550 kHz tone (subject AWS), 0.765 kHz tone (subjects EWB, RSM); [18], white noise (Fig. 2: binaural, magnitude production; binaural, magnitude estimation; monaural, magnitude

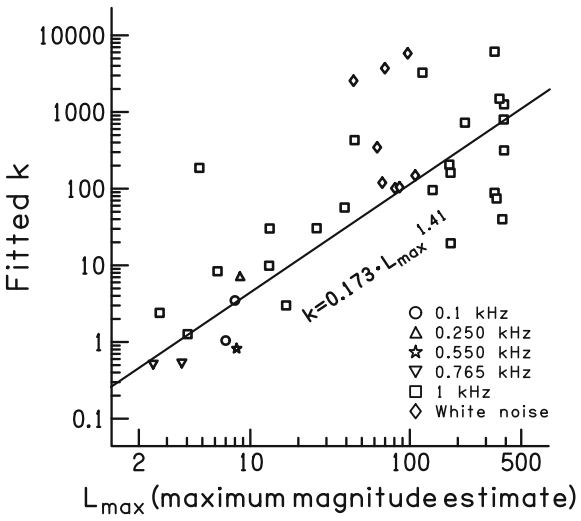


Fig. 14.2 The fitted value of the free parameter k of Eq. (14.1) (values of Fig. 14.1) versus the corresponding maximum magnitude estimate from each empirical loudness-growth plot. The line $k = 0.173 \cdot L_{max}^{1.41}$ is fitted to the data points (see text)

production; monaural, magnitude estimation); [19], 1 kHz tone (Fig. 1, circles; Fig. 1, squares), white noise (Fig. 7, circles and crosses); [20], binaural 1 kHz tone (cross-modality-matching, high range day 2; low range day 2).

Regarding Fig. 14.1: k is *not* constant. For 1 kHz tones and for white noise, multiple k values are evident. Figure 14.2 shows k versus the maximum loudness available from each empirical loudness-growth plot (i.e., the loudness at the highest respective applied stimulus intensity). A power function, converted to logarithmic form (as per the literature), was fitted using sum-of-squares-of-residuals. The latter favors higher loudnesses, hence the data points were weighted by the square root of the absolute value of each loudness. From this, $k = 0.173 \cdot L_{max}^{1.41}$. Norwich and co-authors never mention such a relation. Regardless, k is not constant with L_{max}. Hence, k cannot generally be constant, although Norwich and Wong [1] think otherwise.

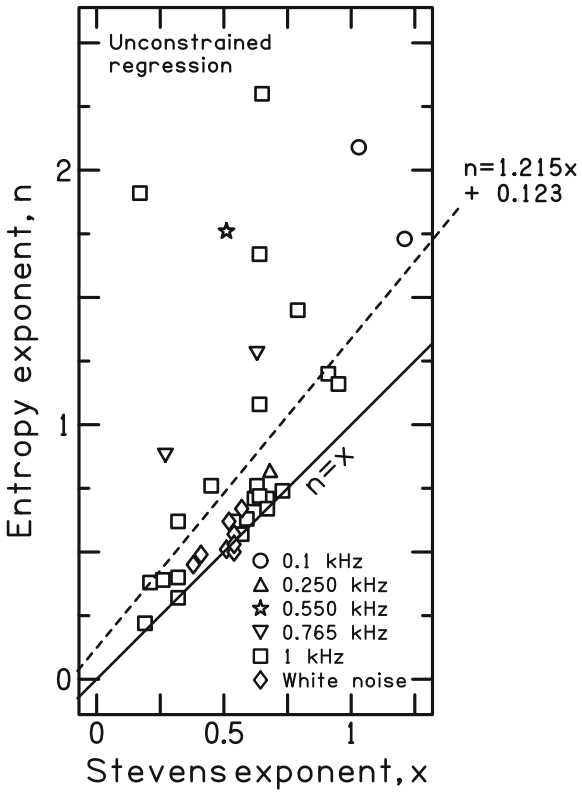


Fig. 14.3 The fitted value of the free parameter n (the Entropy exponent) of Eq. (14.1), versus the fitted value of the Stevens exponent, x, for the loudnesses used in Fig. 14.1. The line n = x indicates putative equality of n and x. The line n = 0.123 + 1.215x is fitted to the data points (see text)

14.7 Examining Norwich and Wong (3): Entropy Exponent, Stevens Exponent

Recall that Norwich and Wong [1] replace n, the exponent of Eq. (14.1) and of Eq. (14.4), with x, the exponent of Stevens' power law: they assume that n = x. But, just as for k, *Norwich and his co-authors never test that assumption, even up to the present day.* Values of n arise from the same curvefitting described just above. Therefore, there is just one parameter remaining to be quantified, namely x. It is obtained through the same procedures described above, but with $L = aI^x$ as the fitted equation. Having thereby obtained values of n and of x, we may examine the notion that n = x. Figure 14.3 shows n versus x.

A straight line fitted to the data points gives n = 0.123 + 1.215x. The reasons for the numbers 0.123 and 1.215 are not known. Nonetheless, n does *not* equal x, contrary to Norwich and Wong [1].

To thoroughly establish any actual relation between n and x, further analysis was done, as follows. Norwich's "Entropy Theory" (e.g., [6]) specifies that the maximum transmitted information during the perception of a stimulus, called $I_{t,\,max}$ units

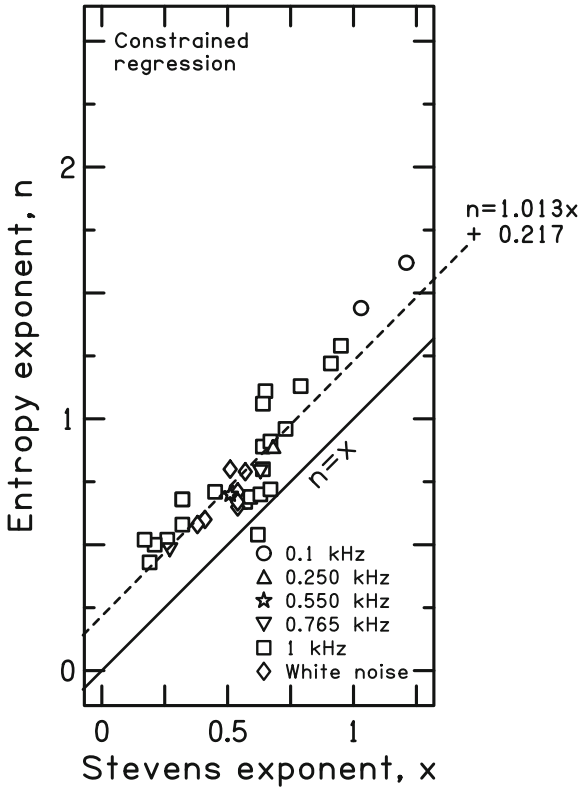


Fig. 14.4 n versus x, for the magnitude estimates used in Fig. 14.1, when the fit of Eq. (14.1) is constrained to produce $I_{t,\,max}$ = 2.5 bits/stimulus (see text). The Entropy exponents n may therefore differ from those in Fig. 14.3. The line n = 0.217 + 1.013x is fitted to the data points, as described in the text

of information, is related to maximum loudness L_{max} and minimum loudness L_{min} as

$$I_{t,\ max} = \frac{L_{max} - L_{min}}{k}. \tag{14.12}$$

k in the denominator is the same k as investigated above. Empirically $L_{max} \gg L_{min}$, allowing L_{min} to be ignored. The fitting of the Entropy equation to loudness growth can be stopped for a value of k yielding $I_{t,\ max} = 2.5$ bits/stimulus, which is the average value found in the literature [6]. Figure 14.4 shows the relation of n to x in such circumstances. A line fitted to the data points in the same manner described above is $n = 0.217 + 1.013x$. The reasons for the values 0.217 and 1.013 are unknown. Again, however, n does *not* equal x, contrary to Norwich and Wong [1]. The difference may seem trivial, but it is well-known that empirical values of x can be as low as 0.3 (e.g., [11]), in which case the difference between n and x is not, in fact, trivial.

14.8 Examining Norwich and Wong (4): Frequency-Dependence of the Exponent

Recall from above that Norwich and Wong ([1], p. 935) employed "a function similar to the one suggested by Marks" [11] to make their theoretical equal-loudness contours hill-shaped like those of Fletcher and Munson [10]. The

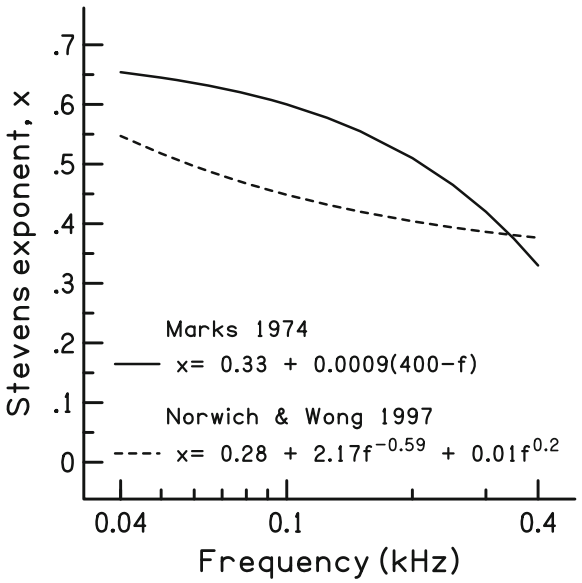


Fig. 14.5 x as a function of frequency, from the equation of Marks and the equation of Norwich and Wong. (The curves are only shown for frequencies below 400 Hz, Marks' limit of validity for his equation.)

respective equation of Marks [11] describes the Stevens exponent x as a function of frequency f (in Hertz), and is

$$x = 0.33 + 0.0009 \cdot (400 - f) \quad where \quad f < 400Hz, \tag{14.13}$$

which applies "over low frequencies (f) and not too high sound pressure levels" ([11], p. 74). Compare Eq. (14.13) to the Norwich and Wong equation "similar to the one suggested by Marks" [1], Eq. (40), but with no restrictions of sound pressure level or frequency:

$$x = 0.28 + 2.17f^{-0.59} + 0.01f^{0.2}. \tag{14.14}$$

Equations (14.13) and (14.14) are not the same. Indeed, we do not know the units of the constants in Eqs. (14.13) and (14.14), hence we do not even know whether the x's represented by the two equations have the same units! Figure 14.5 shows the curves generated by Eqs. (14.13) and (14.14). The curves clearly differ, and in fact, they intersect only once.

14.9 Conclusions

Norwich and Wong [1] present a backwards derivation of arguments made in Wong and Norwich [2]. Wrong arguments in one paper prove to be wrong in the other. In particular, Norwich and Wong [1] make several assumptions which prove to be unjustified. All of this is somewhat surprising, given that Norwich and Wong [1] identify several prominent tenured professors (Lawrence Marks, Lester Krueger, Lawrence Ward) as reviewers of their paper. It is even more alarming that Wong and Norwich [2] was reviewed by a *theorist*, William Hellman.

The present conclusions reveal that embarrassing and needless errors occur consequent to the embarrassing and needless errors of inappropriate limits, unjustified assumptions, and quantities having incompatible units. Indeed, Norwich's entire Entropy Theory is riddled with mistakes (see [21–30]). Professor Norwich evidently tolerates a culture of errors in his laboratory; papers authored independently by his co-authors may prove similarly faulty.

This paper follows upon a contribution [31] to WCECS 2015 (San Francisco, USA).

Acknowledgement My thanks to Claire S. Barnes PhD for her valuable insights.

References

1. Norwich KH, Wong W (1997) Unification of psychophysical phenomena: the complete form of Fechner's law. Percept Psychophys 59:929–940
2. Wong W, Norwich KH (1995) Obtaining equal loudness contours from Weber fractions. J Acoust Soc Am 97:3761–3767
3. Norwich KH (1979) The information content of a steady sensory stimulus, in Proceedings of the 12th International Conference on Medical and Biological Engineering and 5th International Conference on Medical Physics, Jerusalem, Israel, section 19.2
4. Norwich KH (1984) The psychophysics of taste from the entropy of the stimulus. Percept Psychophys 35:269–278
5. Norwich KH, Wong W (1995) A universal model of single-unit sensory receptor action. Math Biosci 125:83–108
6. Norwich KH (1981) The magical number seven: making a 'bit' of 'sense'. Percept Psychophys 29:409–422
7. Riesz RR (1928) Differential intensity sensitivity of the ear for pure tones. Phys Rev Ser 2, 31, 867–875
8. Fechner GT (1860) Element der Psychophysik. Leipzig, Germany: Breitkopf und Hartel. (Translated by H.E. Adler, *Elements of Psychophysics*. New York, USA: Holt, 1966)
9. Wegel RL (1932) Physical data and physiology of excitation of the auditory nerve. Ann Otol Rhinol Laryngol 41:740–779
10. Fletcher H, Munson WA (1933) Loudness, its definition, measurement and calculation. J Acoust Soc Am 5:82–108
11. Marks LE (1974) Sensory processes: the new psychophysics. New York, USA: Academic
12. Eisler H (1962) Empirical test of a model relating magnitude and sensory scales. Scand J Psychol 3:88–96
13. Hellman RP, Zwislocki JJ (1963) Monaural loudness function at 1,000 cps and interaural summation. J Acoust Soc Am 35:856–865
14. Hellman RP, Zwislocki JJ (1968) Loudness determination at low sound frequencies. J Acoust Soc Am 43:60–64
15. Luce RD, Mo SS (1965) Magnitude estimation of heaviness and loudness by individual subjects: a test of a probabilistic response theory. Br J Math Stat Psychol, 18 (part 2), 159–174
16. McGill W (1960) The slope of the loudness function: a puzzle. In: Gulliksen H, Messick S (eds) Psychophysical Scaling: Theory and Applications. Wiley, New York, NY, USA, pp 67–81
17. Richardson LF, Ross JS (1930) Loudness and telephone current. J Gen Psychol 3:288–306
18. Scharf B, Fishken D (1970) Binaural summation of loudness: reconsidered. J Exp Psychol 86:374–379
19. Stevens SS (1956) The direct estimation of sensory magnitudes—loudness. Am J Psychol 69:1–25
20. Ward LM (1987) Remembrance of sounds past: memory and psychophysical scaling. J Exp Psychol Hum Percept Perform 13:216–227
21. Nizami L (2010) Fundamental flaws in the derivation of Stevens' Law for taste within Norwich's Entropy Theory of Perception. In: Korsunsky AM (ed) Current Themes In Engineering Science 2009: Selected Presentations at the World Congress on Engineering-2009 (AIP Conference Proceedings 1220). American Institute of Physics, Melville, NY, USA, pp 150–164
22. Nizami L (2011) A warning to the human-factors engineer: false derivations of Riesz's Weber fraction, Piéron's Law, and others within Norwich et al.'s Entropy Theory of Perception. In: Schmidt M (ed) Advances in Computer Science and Engineering. InTech, Rijeka, Croatia, pp 391–406
23. Nizami L (2009) A computational test of the information-theory based Entropy Theory of Perception: does it actually generate the Stevens and Weber-Fechner Laws of sensation? In:

Lecture Notes in Engineering and Computer Science: Proceedings of the World Congress on Engineering 2009, WCE 2009, London, UK, pp 1853–1858
24. Nizami L (2011) Norwich's Entropy Theory: how not to go from abstract to actual. Kybernetes 40:1102–1118
25. Nizami L (2009) Sensory systems as cybernetic systems that require awareness of alternatives to interact with the world: analysis of the brain-receptor loop in Norwich's Entropy Theory of Perception. In: Proceedings of the 2009 IEEE International Conference on Systems, Man, and Cybernetics, San Antonio, TX, USA, pp 3477–3482
26. Nizami L (2009) A lesson in the limitations of applying cybernetics to sensory systems: hidden drawbacks in Norwich's model of transmitted Shannon information. IIAS-Trans Syst Res Cybern (Int J Int Inst Adv Stud Syst Res Cybern) 9:1–9
27. Nizami L (2008) Is auditory intensity discrimination a comparison of entropy changes? In: Proceedings of the 155th Meeting of the Acoustical Society of America, 5th Forum Acusticum of the EA, 9e Congrès Français d'Acoustique of the SFA, 2nd ASA-EAA Joint Conference, Paris, France, pp 5739–5744
28. Nizami L (2008) Does Norwich's Entropy Theory of Perception avoid the use of mechanisms, as required of an information-theoretic model of auditory primary-afferent firing? In: Proceedings of the 155th Meeting of the Acoustical Society of America, 5th Forum Acusticum of the EA, 9e Congrès Français d'Acoustique of the SFA, 2nd ASA-EAA Joint Conference, Paris, France, pp 5745–5750
29. Nizami L (2009) On the hazards of deriving sensory laws from first-order cybernetics: Norwich's Entropy Theory of Perception does not derive the Weber fraction for audition. In: Proceedings of the 13th World Multi-Conference on Systemics, Cybernetics and Informatics (The 3rd International Symposium on Bio- and Medical Informatics and Cybernetics: BMIC 2009 of WMSCI 2009), Orlando, FL, USA, pp 235–241
30. Nizami L (2015) On mathematical biology without the biology: a refutation of K.H. Norwich's Mystery of Loudness Adaptation and a physiology-based replacement. In: Lecture Notes in Engineering and Computer Science: Proceedings of the World Congress on Engineering and Computer Science 2015, WCECS 2015, San Francisco, CA, USA, pp 561–569
31. Nizami L (2015) Mathematical biology under wrong limits and wrong assumptions: Norwich and Wong (the Entropy Theory) fail to derive equal-loudness contours. In: Lecture Notes in Engineering and Computer Science: Proceedings of the World Congress on Engineering and Computer Science 2015, WCECS 2015, San Francisco, CA, USA, pp 546–551

Chapter 15
High Assurance Asynchronous Messaging Methods

William R. Simpson and Kevin Foltz

15.1 Introduction

This paper is based in part on a paper published by WCECS [1]. Asynchronous messaging describes communication that takes place between one or more applications or systems, in which the sender does not receive feedback from the receiver during transmission of a message. This is in contrast to synchronous communication, in which the sender of a message waits for acknowledgement or a response from the receiver before completing the transmission.

There is no assumption about which layers asynchronous and synchronous communication take place in or how these relate to each other. It is possible to implement synchronous communication using an asynchronous messaging service or an asynchronous messaging service using synchronous communication channels. In practice, asynchronous messaging often uses an underlying synchronous channel.

A common asynchronous messaging design involves one system placing a message in a message queue and continuing its processing. At the completion of message transmission, the sender does not know when or whether the receiver received it. The message queuing system is responsible for delivering the message to the recipient. Some systems use two or more queues or intermediaries.

W.R. Simpson (✉) · K. Foltz
Institute for Defense Analyses, 4850 Mark Center Drive, Alexandria, VA 22311, USA
e-mail: rsimpson@ida.org

K. Foltz
e-mail: kfoltz@ida.org

S.-I. Ao et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-981-10-2717-8_15

15.1.1 Some Advantages of Asynchronous Communication

Asynchronous messaging solves the problem of intermittent connectivity. If the receiving equipment fails or is unavailable, the message remains in a message queue and is delivered after the failure is corrected. This is especially useful for transmission of large data files, in which failures are more likely and retransmissions more costly.

An asynchronous messaging system with built-in intelligence may transform the content and/or format of the message automatically to conform to the receiving system's requirements or needed protocol but still successfully deliver the message to the recipient. This intelligence is used to provide a higher level of understanding of the content, which allows translation into other formats and protocols. Complicated transformations are better suited to asynchronous communication than synchronous communication because they may increase latency and cause connectivity problems or other underlying protocol failures for synchronous systems.

15.1.2 Some Disadvantages of Asynchronous Communication

The disadvantages of asynchronous messaging include the additional component of a message broker or transfer agent to ensure the message is received. This may affect both performance and reliability. Another disadvantage is the response time, which may be inconvenient and not consistent with normal dialog communication.

15.2 High Assurance Issues

There are several high assurance security principles that must be maintained for asynchronous communication.

1. Know the players—in synchronous (two way communication) this is done by enforcing bi-lateral end-to-end authentication. For synchronous communication Enterprise Level Security the certificate is PKI.
2. Maintain Confidentiality—in synchronous Communications this entails end-to-end encryption using Transport Layer Security (TLS).
3. Enforce Access and Privilege—in synchronous (two way communication) this is done by the use of an authorization credential. For synchronous communication Enterprise Level Security the certificate is SAML.
4. Maintain Integrity—know that you received exactly what was sent—know that content has not been modified—For synchronous communication and enterprise level end-to-end TLS encryption the use of message authentication codes is enforced. Packages are signed and signatures are verified and validated. Credentials of signers are verified and validated.

15.3 Prior Work

A proliferation of standards [2–12] for asynchronous messaging has caused interoperability problems, with each major vendor having its own implementations, interface, and management tools. Java EE systems are not interoperable, and Microsoft's MSMQ (Microsoft Message Queuing) does not support Java EE. Many of these are reviewed and compared in [13]. A few of the numerous standard protocols used for asynchronous communication as defined in the Internet Assigned Numbers Authority (IANA) protocol registries [14] are in the Table 15.1.

Table 15.1 Messaging ports

Port	TCP/UDP	Messaging Protocol and Description	Status
18	TCP and UDP	The **Message Send Protocol (MSP)**, more precisely referred to as Message Send Protocol 2, is an application layer protocol used to send a short message between nodes on a network. Defined in RFC 1312	Official
110	TCP	**Post Office Protocol v3 (POP3)** is an email retrieval protocol	Official
119	TCP	The **Network News Transfer Protocol (NNTP)** is an application protocol used for transporting Usenet news articles (*netnews*) between news servers and for reading and posting articles by end user client applications. Defined in RFC 3977	Official
143	TCP	**Internet Message Access Protocol (IMAP)** is a protocol for e-mail retrieval and storage as an alternative to POP. IMAP, unlike POP, specifically allows multiple clients simultaneously connected to the same mailbox and through flags stored on the server; different clients accessing the same mailbox at the same or different times can detect state changes made by other clients. Defined in RFC 3501	Official
161	UDP	**Simple Network Management Protocol (SNMP)** is an "Internet-standard protocol for managing devices on IP networks." Devices that typically support SNMP include routers, switches, servers, workstations, printers, modem racks, and more. Defined in RFC 3411-3418	Official
218	TCP and UDP	**Message Posting Protocol (MPP)** is a network protocol used for posting messages from a computer to a mail service host	Official
319	UDP	Event Messages for The **Precision Time Protocol (PTP)** is a protocol used to synchronize clocks throughout a computer network. On a local area network, it achieves clock accuracy in the sub-microsecond range, making it suitable for measurement and control systems. Defined in **IEEE 1588-2008**	Official
587	TCP	**Simple Mail Transfer Protocol (SMTP)**, as specified in RFC 6409	Official

(continued)

Table 15.1 (continued)

Port	TCP/UDP	Messaging Protocol and Description	Status
1801	TCP and UDP	**Microsoft Message Queuing** or **MSMQ** is a message queue implementation developed by Microsoft and deployed in its Windows Server operating systems	Official
1863	TCP	**MSNP (Microsoft Notification Protocol)**, used by the Microsoft Messenger service and a number of Instant Messaging clients	Official
1935	TCP	Adobe Systems Macromedia Flash Real Time Messaging Protocol (RTMP) "plain" protocol	Official
2195	TCP	Apple Push Notification service Link	Unofficial
2948	TCP and UDP	**Multimedia Messaging Service (MMS)** is a standard way to send messages that include multimedia content to and from mobile phones	Official
4486	TCP & UDP	**Integrated Client Message Service (ICMS). Defined in RFC 6335**	Official
5010	TCP	**IBM WebSphere MQ Workflow**	Official

15.3.1 Java Standard Messaging Protocol

Java Messaging System (JMS) is a message-oriented middleware API for communication between Java clients. It is part of the Java Platform Enterprise Edition. It supports point-to-point communication as well as publish-subscribe.

15.3.2 De-Facto Standard Microsoft Message Queuing

Microsoft Message Queuing (MSMQ) allows applications running on separate servers/processes to communicate in a failsafe manner. A queue is a temporary storage location from which messages can be sent and received reliably, as and when conditions permit. This enables communication across networks and between computers running Windows, which may not always be connected. By contrast, sockets and other network protocols require permanent direct connections.

15.3.3 Open Source Messaging Protocols

In addition to Java and Microsoft, different open source solutions exist [2–12]. RabbitMQ is an open source messaging solution that runs on multiple platforms and multiple languages. It implements Advanced Message Queueing Protocol (AMQP), in which messages are queued on a central node before being sent to

clients. It is easy to deploy, but having all traffic pass through a single central node can hinder scalability.

ZeroMQ is another cross-platform, cross-language messaging solution that can use different carrier protocols to send messages. It can support publish-subscribe, push-pull, and router-dealer communication patterns. It can be more difficult to set up, but it provides more control and granularity at the lower levels to tune performance.

ActiveMQ is a compromise between the ease of use of Rabbit MQ and the performance of ZeroMQ. All three support multiple platforms and have client APIs for C++, Java,.Net, Python, and others. They also have documentation and active community support. There are many other implementations, including Sparrow, Starling, Kestrel, Beanstalkd, Amazon Simple Queue Service (SQS), Kafka, Eagle MQ, and IronMQ.

15.3.4 Emerging Standard Advanced Message Queuing Protocol

Advanced Message Queuing Protocol (AMQP) is an open standard application layer protocol for message-oriented middleware [15]. It is an emerging technology addressing the standardization problem. Implementations are interoperable. It includes flexible routing and common message paradigms like publish/subscribe, point-to-point, request-response, and fan-out.

The defining features of AMQP are message orientation, queuing, routing (including point-to-point and publish-and-subscribe), reliability, and security. AMQP mandates the behavior of the messaging provider and client to the extent that implementations from different vendors are truly interoperable, in the same way as SMTP, HTTP, FTP, and others have created interoperable systems.

15.4 Asynchronous Messaging Security

Asynchronous messaging can provide authentication of the sender and receiver identities and the integrity and confidentiality of the message content if the holder of the queue is trusted. One key challenge in asynchronous messaging systems is that a third party is often involved in the transaction, which may or may not be trusted to speak for the sending or receiving entities or view or modify content in transit. As a result, security models often require a trusted third party, which restricts deployment options. In contrast, synchronous web traffic relies on routers and other infrastructure to deliver messages, but the use of TLS provides end-to-end security without the need to trust these intermediate nodes.

15.4.1 Security for Server Brokered Invocation

Server brokered invocation uses web server middleware to manage message queues. The sender and receiver both communicate directly through secure synchronous channels to the server to send and receive messages. This model is shown in Fig. 15.1. Asynchronous message security must be from sender to receiver, not just from sender to server and server to receiver. The latter fails to provide end-to-end authentication, integrity, and confidentiality, which are required for a high assurance environment.

In order for the parties involved in the transaction to provide accountability, integrity, and confidentiality, the service requester must authenticate itself to the receiver, encrypt the message so only the service provider can receive this message, and provide verifiable integrity checks on the full message content. The service provider must confirm that the message is from a known identity, decrypt the content with a valid key, and verify the integrity checks before that entity can take action on the message.

This is accomplished by invoking two cryptographic techniques. The first is the use of a digital signature by the sender. When the message signature is verified, the service provider knows the identity of the sender and that the content has not been altered by another entity after it was signed. The second is the encryption of the message using the public key of the service provider. This requires that the requester know the public key of the target. A response to the requester must similarly be signed and encrypted using the public key of the requester.

The use of asymmetric encryption is paired with more efficient symmetric encryption, where content is encrypted with a random symmetric key, which is itself encrypted using the receiver's public key. Additional security can be provided by message expiration deadlines within queues and central auditing of all messages sent and received.

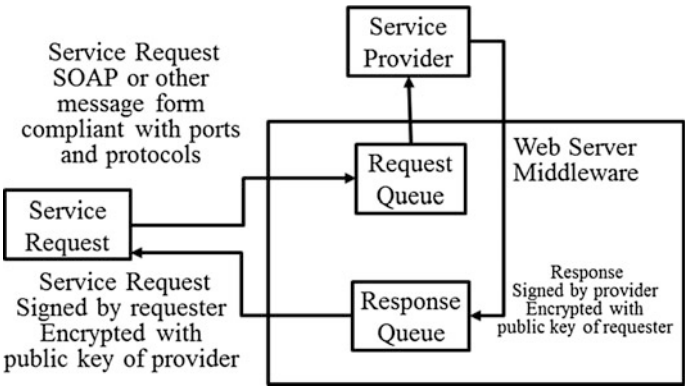


Fig. 15.1 Security considerations for server brokered invocation

15.4.2 Security for Publish Subscribe Systems

In a Publish Subscribe System (PSS) the queue server acts as an intermediary between sender and receiver to manage many-to-many instead of just many-to-one communications. Senders and receivers communicate with the PSS through a secure synchronous channel. The PSS collects messages and makes them available to entities based on subscriptions. This model is shown in Fig. 15.2.

The PSS is an active entity and registered in the Enterprise Service Directory. Active entities act on their own behalf and are not a proxy. To preserve the end-to-end accountability chain for messages, the original publisher signs the message. However, unlike server-brokered invocation, no single public key can be used for all potential receivers. One solution to address this is for the PSS to encrypt the content to the receivers. The sender's signature remains intact, preserving integrity, but end-to-end confidentiality is not guaranteed.

A PSS may use the web server broker as shown in Fig. 15.3. The web server broker is used only for notification messages, so it does not require security like the main channel. The transmission of the actual message is still done through the secure synchronous channel. The storage queue must be encrypted using the PSS's public key. This is piecemeal confidentiality, because the sender encrypts to the PSS, and the PSS encrypts to the receiver. This relies on trust of the PSS.

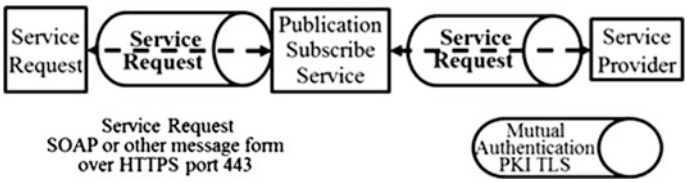


Fig. 15.2 Publish-subscribe push model

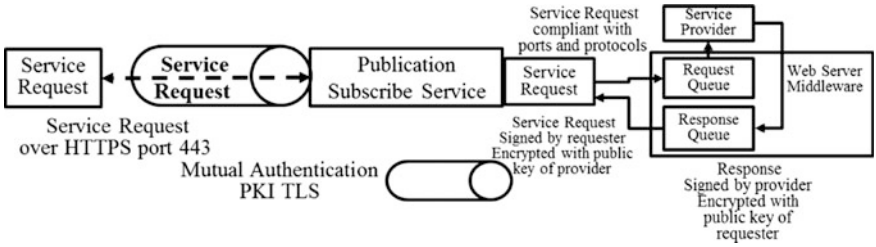


Fig. 15.3 Publish-subscribe pull model

15.5 Pss Rock and Jewel

The following is an approach developed to maintain high security assurances with the use of an untrusted PSS. In this formulation, the sender and receiver maintain end-to-end security because the PSS is unable to impersonate either endpoint or view or modify the content. The key concepts are the use of "rocks" and "jewels" to provide security guarantees. The "rocks" are encrypted content blocks, and the "jewels" are the decryption keys for these rocks, encrypted using public keys for the intended recipients.

15.5.1 Claims for Targeted Content (PSS)

After authentication through TLS v1.2 or later versions and authorization based on SAML claims, the sender accesses PSS services. The PSS will offer either publish or retrieve based on the values in the SAML content claim. If there are no SAML content claims, the subscriber will only receive basic services based on identity.

Publishing of content for a targeted list, as used by software publishers, is based upon registered delivery. The targeted list requires the following steps:

0. Publisher does a bi-lateral authentication and establishes a TLS 1.2 session with SAML authorizations for session establishment with the PSS. The PSS identifies him as a publisher. He may also be a subscriber, or he may be modifying a previous publish or he may be retrieving messages, so the PSS ascertains the reason for his session.
1. Content to be published will be digitally signed by the publisher.
2. The publisher will generate an AES-256 encryption key and encrypt the content.
3. Encrypted Content is placed in a queue based on an access claim and list name. The publisher will keep such lists. The PSS will assist in developing claims.
4. Access is based on a list of targets and claims. A target may be an individual subscriber or a group queue. The publisher may establish a new queue based on claims and the list for retrieval. This new queue requires an identity and a claims establishment for retrieval (see 3 above). Additional content may be published as needed.
5. Expiration time of targeted content is determined by the publisher or the messaging system.
6. The PSS will provide PKI certificates for each of the targets for the content (if the publisher needs them and they are already registered in the PSS). The publisher should check all certificates on the list for currency and revocation. If invalid certificates are discovered, the list should be pruned.
7. The publisher will prepare encrypted key sets (jewels) by wrapping the AES encryption key in each target's public key.

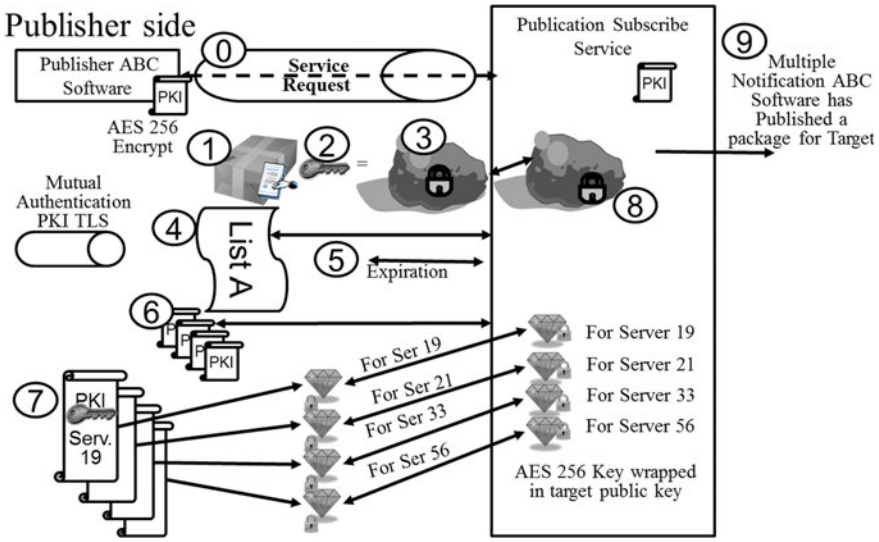


Fig. 15.4 Publishing of targeted content

8. The publisher will publish the encrypted material (rocks) and the encrypted key sets (jewels) for the targets. The PSS will link these to the encrypted material and the target(s).
9. The PSS will provide notification, if desired, to the subscriber list. The PSS will assist with message selection and target details, or the publisher may script his own.
10. The publisher closes the session.

Note: the target must be on the list and have authorization to view content. The steps are shown in Fig. 15.4.

15.5.2 Retrieving Content for Known Claimants

Retrieval of targeted content may be achieved without the targeted identities contacting the publisher. The following steps are followed:

0. Subscriber does a full bi-lateral authentication using TLS 1.2 with SAML authorizations for session establishment with the PSS. The claims identify him as a subscriber. He may also be a publisher, so the PSS ascertains the reason for his session.
1. The PSS offers subscriber content available for the claims in queues for which the claimant has an encrypted key available, and the subscriber chooses and retrieves the encrypted content (rock).

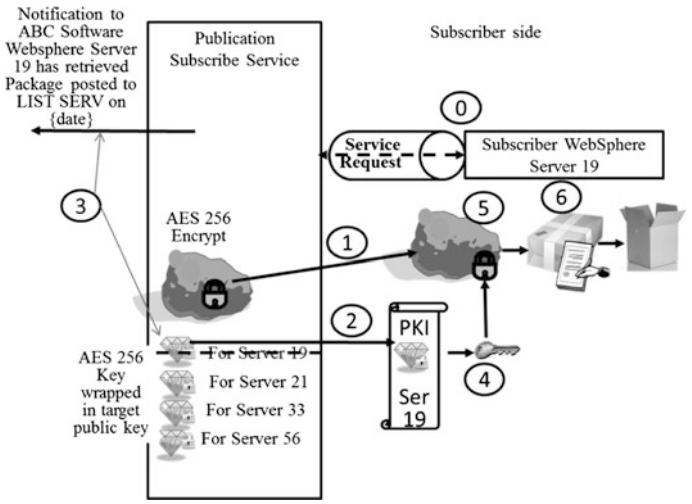


Fig. 15.5 Subscriber retrieval(s) from a known target

2. The PSS provides the encrypted key package (jewel).
3. The PSS notifies the publisher. When expiration time occurs, the server deletes the packages and notifies the publisher which packages were not delivered. The publisher may republish to that list if desired.
4. The subscriber decrypts the content encryption key (jewel) with his private key and accesses the content (rock) decryption key.
5. The subscriber decrypts the content.
6. The subscriber verifies and validates signature.
7. The subscriber closes the session or retrieves additional content.

Note: the target must be on the list and have a content claim. The steps are shown in Fig. 15.5.

15.5.3 Retrieving Content for Unknown Claimants

Unknown claimants cannot retrieve the content until registering with the content provider. The steps in that process are described below:

0. The subscriber does a full bi-lateral authentication TLS 1.2 with SAML authorizations for session establishment with the PSS. The authentication identifies him as a subscriber. He may also be a publisher, so the PSS ascertains the reason for his session.
1. The PSS checks the content claims available and the subscriber chooses and retrieves the content for which full packages exist.

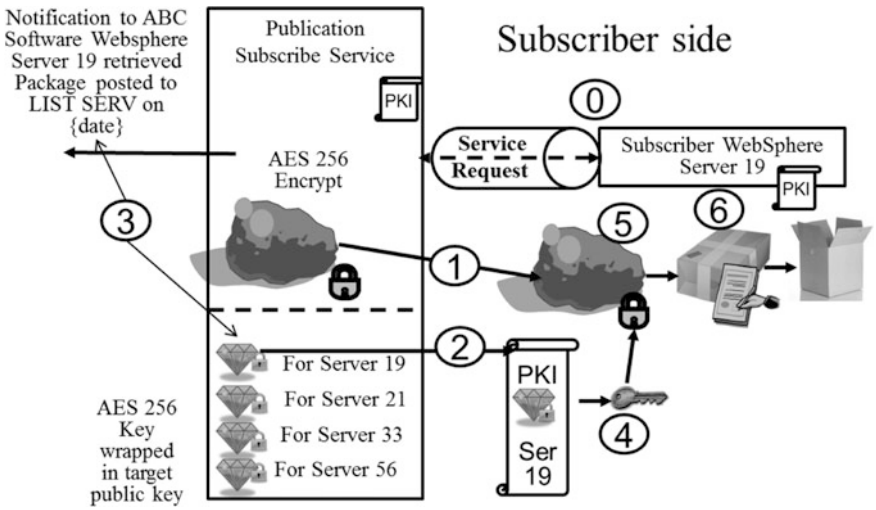


Fig. 15.6 Subscriber retrieval(s) from an unknown target

2. For the unknown list, the encrypted key package is not available. The PSS replies "the publisher has no record of your membership. I need to contact the publisher. I will send you a notice if the publisher agrees."
3. The PSS stores a message for the publisher and notifies him that he has a message.
4. The PSS and subscriber await publisher action.
5. The subscriber closes the session or retrieves additional content.

Note: the target has a content claim, but is not on the list. The steps are shown in the Fig. 15.6.

15.5.4 Adjusting Publishing Targets (Untrusted PSS)

Publishers must add receivers to the distribution list before they can be provided messages. The steps in that process are described below:

0. The publisher does a full bi-lateral authentication through TLS 1.2 with SAML authorizations for session establishment with the PSS. The authorization process identifies him as a publisher. He may also be a subscriber, or he may be modifying a previous publish or he may be retrieving messages, so the PSS ascertains the reason for his session.
1. Retrieve messages. These are retrieved one by one with action taken (or not) and deletion of the message.

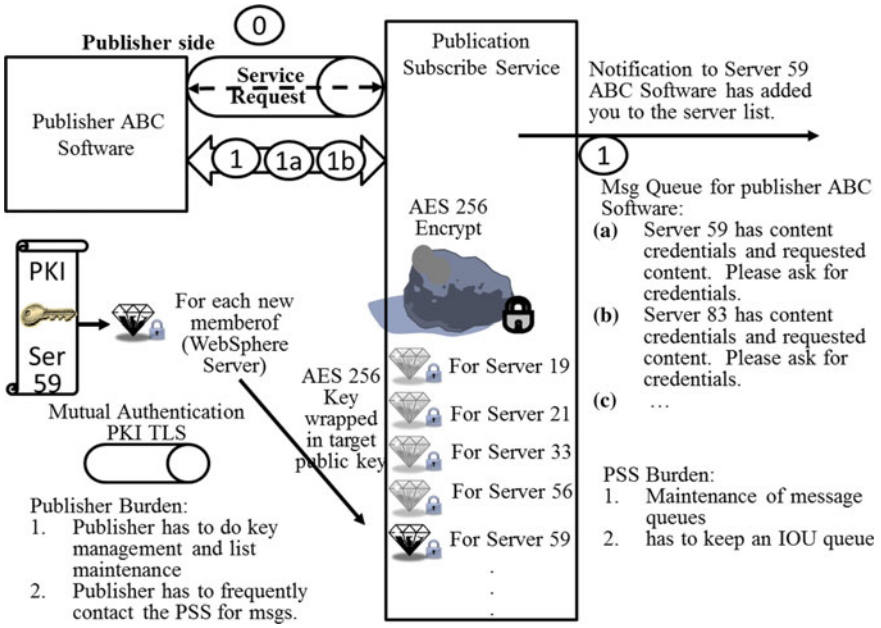


Fig. 15.7 Publisher message retrieval and subsequent actions

- The publisher asks for credentials of previously unknown claimants he wishes to add to his lists.
  - The publisher may add claimants to the publisher's list
  - The publisher computes jewels.
  - The publisher posts jewels.
  - The PSS notifies the subscriber that he has content available. This makes the entity a known target and Sect. 15.5.2 applies.
- PSS provides messages to requester.

2. The publisher closes the session.

The steps are shown in Fig. 15.7.

15.5.5 Distribution of Burdens

Several burdens are incurred in this high security mode. The publisher has to do key management and list maintenance. The publisher has to frequently contact the PSS for messages for publishers. The PSS must maintain message queues for publishers.

The PSS has to keep a linked wrapped key package by target with published content. The PSS is responsible for additional notifications that are sent out. The unknown claimant may have a delay in receiving content to which he has claims.

15.6 Summary

We have reviewed the basic approaches to asynchronous communication in computing environments. We have also described high assurance approaches to the process. The proliferation of standards in this area has created a problem with high assurance. In many instances the high assurance elements require additional steps in the asynchronous process, but they provide a way to proceed when some intermediaries are untrusted. This work is part of a body of work for high assurance enterprise computing using web services. Elements of this work are described in [16–29].

References

1. Foltz K, Simpson WR (2015) Maintaining high assurance in asynchronous messaging. In: Proceedings World Congress on Engineering and Computer Science 2015, WCECS2015. Lecture Notes in Engineering and Computer Science. San Francisco, USA, pp 187–192
2. World Wide Web Consortium (W3C): XML Encryption Syntax and processing. http://www.w3.org/
3. Organization for the Advancement of Structured Information Standards (OASIS) open set of Standards: Web Services (WS) standards and Security Assertion Markup Language (SAML) standards. https://www.oasis-open.org/standards
4. National Institute of Standards, Gaithersburg, MD: Encryption related standards. http://www.nist.gov/srm/
5. PKCS#1: RSA Cryptography Standard: ASN Module for PKCS#1 v2.1, June 14, 2002
6. MSMQ—http://msdn.microsoft.com/en-us/library/ms711472(v=vs.85).aspx
7. RabbitMQ—http://www.rabbitmq.com/
8. Apache ActiveMQ—http://activemq.apache.org/
9. Apache Qpid—http://qpid.apache.org/
10. JBoss HornetQ—http://www.jboss.org/hornetq
11. ZeroMQ—http://www.zeromq.org/
12. WebSphereWMQ–http://www-01.ibm.com/software/integration/wmq/
13. Message Broker Comparison– http://lifecorporatedev.blogspot.com/2012/07/recently-i-have-been-given-task-to-find.html
14. Internet Engineering Task Force (IETF) Standards: IANA and Protocol standards including The Transport Layer Security (TLS). https://www.ietf.org/
15. Advanced Message Queuing Protocol (AMQP)—https://www.amqp.org/
16. Simpson WR, Chandersekaran C, Trice A (2008) A persona-based framework for flexible delegation and least privilege. In: Electronic Digest of the 2008 System and Software Technology Conference, Las Vegas, Nevada, May 2008

17. Simpson WR, Chandersekaran C, Trice A (2008) Cross-domain solutions in an era of information sharing. In: The 1st International Multi-Conference on Engineering and Technological Innovation: IMET2008, vol I, Orlando, FL, June 2008, pp 313–318
18. Chandersekaran C, Simpson WR (2008) The case for bi-lateral end-to-end strong authentication. In: World Wide Web Consortium (W3C) Workshop on Security Models for Device APIs. London, England, December 2008, pp 4
19. Simpson WR, Chandersekaran C (2009) Information sharing and federation. In: The 2nd International Multi-Conference on Engineering and Technological Innovation: IMETI2009, vol I, Orlando, FL, July 2009, pp 300–305
20. Chandersekaran C, Simpson WR, (2010) A SAML framework for delegation, attribution and least privilege. In: The 3rd International Multi-Conference on Engineering and Technological Innovation: IMETI2010, vol 2, Orlando, FL, July 2010, pp 303–308
21. Simpson WR, Chandersekaran C (2010) Use Case Based Access Control. In: The 3rd International Multi-Conference on Engineering and Technological Innovation: IMETI2010, vol 2, Orlando, FL, July 2010, pp 297–302
22. Chandersekaran C, Simpson WR (2011) A model for delegation based on authentication and authorization. In: The First International Conference on Computer Science and Information Technology (CCSIT-2011). Lecture Notes in Computer Science. Springer Verlag Berlin-Heildleberg, pp 20
23. Simpson WR, Chandersekaran C (2011) An agent based monitoring system for web services. In: The 16th International Command and Control Research and Technology Symposium: CCT2011, vol II, Orlando, FL, April 2011, pp 84–89
24. Simpson WR, Chandersekaran C (2011) An agent-based web-services monitoring system. Int J Comput Technol Appl (IJCTA), 2(9):675–685
25. Simpson WR, Chandersekaran C, Wagner R, (2011) High Assurance Challenges for Cloud Computing. In: Proceedings World Congress on Engineering and Computer Science 2011, Lecture Notes in Engineering and Computer Science, vol I, San Francisco, October 2011, pp 61–66
26. Chandersekaran C, Simpson WR (2012) Claims-Based Enterprise-Wide Access Control. In: Proceedings World Congress on Engineering 2012, The 2012 International Conference of Information Security and Internet Engineering. Lecture Notes in Engineering and Computer Science, vol I, London, July 2012, pp 524–529
27. Simpson WR, Chandersekaran C (2012) Assured content delivery in the enterprise. In: Proceedings World Congress on Engineering 2012, The 2012 International Conference of Information Security and Internet Engineering. Lecture Notes in Engineering and Computer Science, vol I, London, July 2012, pp. 555–560
28. Simpson WR, Chandersekaran C (2012) Enterprise high assurance scale-up. In: Proceedings World Congress on Engineering and Computer Science 2012. Lecture Notes in Engineering and Computer Science, vol 1, San Francisco, October 2012, pp. 54–59
29. Chandersekaran C, Simpson WR (2012) A uniform claims-based access control for the enterprise. Int J Sci Comput 6(2):1–23. ISSN:0973-578X

Chapter 16
Regional Differences in Iwate Prefecture Road Usage Recovery Following the 2011 Tohoku Earthquake

Noriaki Endo and Hayato Komori

16.1 Introduction

16.1.1 The 2011 Tohoku Earthquake

The 2011 Tohoku Earthquake Fig. 16.1 struck the northeastern coast of Japan on March 11, 2011. Subsequently, the region was severely affected by the tsunami. Following these natural disasters, the electricity, water, and gas supplies were shut down in both coastal and inland areas. Furthermore, the road travel was disrupted in many parts of the region.

16.1.2 Purpose of the Study

The purpose of our study was to evaluate regional differences in the road usage recovery in the Iwate Prefecture following the 2011 Tohoku Earthquake.

Therefore, we divided the Iwate Prefecture into four areas according to their geographic positions and features: the Northern Inland, Southern Inland, Northern Coastal, and Southern Coastal areas.

During the disaster, these four areas were affected differently. The strength of the earthquake in the southern areas was much stronger than that in the northern areas. And there are some geographical differences between Northern and Southern Costal areas.

N. Endo (✉) · H. Komori
Iwate University, Morioka City, Iwate, Japan
e-mail: n_endo@gw2k.hss.iwate-u.ac.jp

H. Komori
e-mail: h_komori@gw2k.hss.iwate-u.ac.jp

S.-I. Ao et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-981-10-2717-8_16

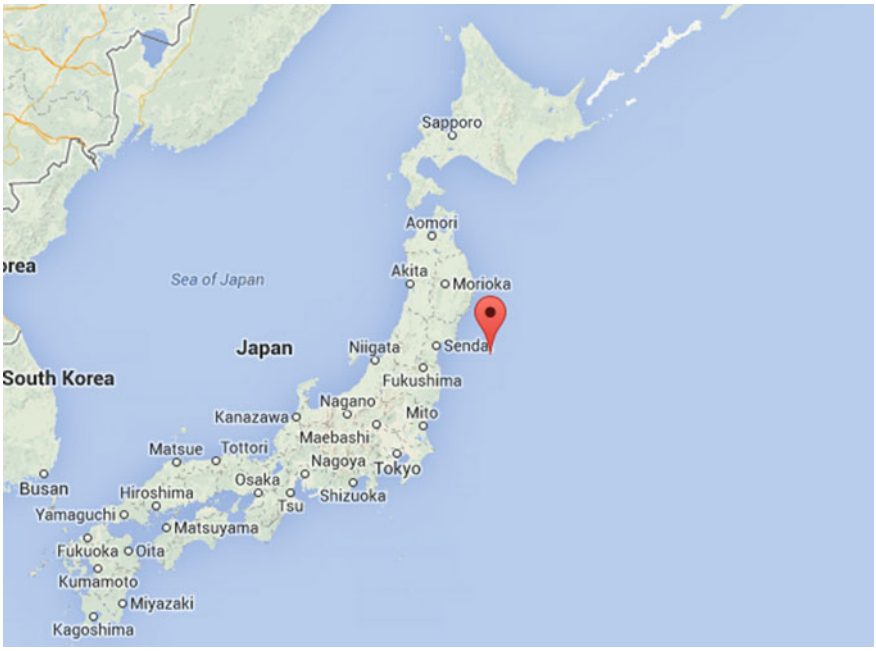


Fig. 16.1 Center of the 2011 Tohoku Earthquake occurred on March 11, 2011 (https://www.google.co.jp/maps/)

Therefore, we assume that there were specific differences between the four studied regions during the road recovery process following the disaster.

16.2 Our Previous Studies

16.2.1 Previous Study 1

Our previous study 1 [1] was focused on the use of the main roads in the Southern Coastal area of Iwate Prefecture Fig. 16.2. The usable distances of the main roads following the 2011 Tohoku Earthquake have been calculated from the G-BOOK telematics data [2].

The main findings of this study are as follows:

(1) The usable distance of the roads in a weekly period has continuously increased from March 18 to April 7, 2011, but it has fluctuated thereafter.

Fig. 16.2 Main roads in the Southern Coastal area of Iwate Prefecture (http://map.yahoo.co.jp/)

(2) Defining the cumulative usable distance up to September 30, 2011 as 100 %, it has been determined that 80 % of the road distance has been usable by April 7, 2011 and 90 % by April 29, 2011.

(3) The use of the main road in the coastal area of Iwate Prefecture has been completely recovered by April 29, 2011.

16.2.2 Previous Study 2

Our previous study 2 [3] was focused on the use of the main roads not only in the Southern but also in the Northern Coastal area of Iwate Prefecture.

The cumulative usable road distance ratio of the main roads has been precisely calculated from the G-BOOK telematics data [2] for each city using the free and open source geographical information system software QGIS. The main findings in this study [3] are listed below.

(1) The change in the cumulative usable road distance ratio during the research period differed from one city to the next.

(2) The ratio increases in the usable distances of Kuji, Iwaizumi, and Noda were extremely delayed. In our study, we were able to determine related roads by analyzing the maps generated by the QGIS software. For Kuji and Iwaizumi, the delays were mostly dependent on the Iwate Prefectural Road number 7 (the Kuji-Iwaizumi line). For Noda, the delay was mostly dependent on the Iwate Prefectural Road number 273 (the Akka-Tamagawa line).

(3) In our previous study 1 [1], we determined that the use of the main road in the Southern Coastal area of Iwate Prefecture was completely recovered by April 29, 2011.

However, in this study, when we have precisely observed the change in the usable road distance ratio during the research period for each city, the ratio increase in the usable road distance of Kamaishi has been delayed compared with other Southern Coastal cities.

For Kamaishi City, the delay was mostly dependent on the Iwate Prefectural Road number 249 (the Sakuratoge-Heita line).

16.2.3 Previous Study 3

This paper is the extended version of our third paper [4].

16.3 Telematics Data and Vehicle-Tracking Map

Telematics is a general term encompassing telecommunications and informatics. A telematics service provides various personalized information for users, especially for drivers of automobiles. G-BOOK is a telematics service provided by Toyota Motor Corporation.

To calculate the usable distance of the main roads, we applied the vehicle tracking map originally created by Hada et al. [5] after the 2007 Niigataken Chuetsu-oki earthquake.

That vehicle tracking map was based on telematics data provided by Honda Motor Company. Similarly, in our study, we used the vehicle tracking map based on telematics data provided by Toyota's G-BOOK system Figs. 16.3 and 16.4.

Registered members of G-BOOK can access telematics services to acquire GPS data for car navigation systems and interactive driving data, such as traffic jam points, road closures, and weather reports.

Such comprehensive data acquisition is possible because the telematics system server receives accurate location data (geographic coordinates) from its registered members.

Telematics services are extremely useful to drivers. Because the accurate driving routes of registered users remain in the system server, they are accessible to traffic researchers in various fields.

16.4 Research Methods

The current study used the research methods of our previous study 2 [3]; only the research area was different.

Fig. 16.3 Vehicle-tracking map of eastern Japan

16.4.1 Research Area

Our previous study 1 [1] was focused only on the Southern Coastal area of Iwate Prefecture Fig. 16.4, and our previous study 2 [3] was focused on the entire coastal area of Iwate Prefecture (i.e., both Southern and Northern Coastal areas) Fig. 16.4.

The current study was focused on the entire area of Iwate Prefecture (i.e., Southern and Northern Coastal areas as well as Southern and Northern Inland areas) Fig. 16.4.

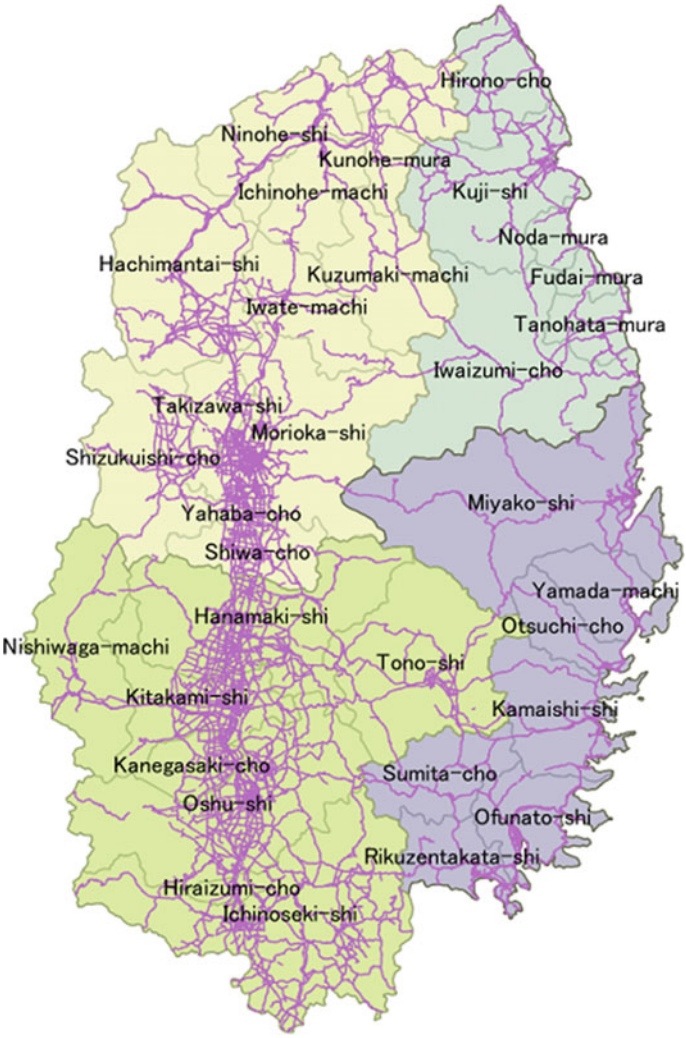


Fig. 16.4 Vehicle-tracking map of Iwate Prefecture is shown by the *violet lines*. The perimeter of a city is shown by a *gray polygon*. Iwate Prefecture has been divided into four areas: Northern Costal, Southern Coastal, Northern Inland, and Southern Inland areas

16.4.2 Research Materials

In our current study, we have used the vehicle-tracking maps built from the G-BOOK telematics data that is available on the Internet on March 18, 2011 following the 2011 Tohoku Earthquake [2].

The data used in this study have been collected between March 18 and September 30, 2011 (i.e., approximately six months following the 2011 Tohoku Earthquake).

16.4.3 System

Hardware:
The computations have been performed on a standard PC laptop with a Core i5-4200U CPU (1.6 GHz) and 4 GB memory (SONY VAIO PRO 11).

Software:
The software QGIS version 2.6.1 [6] and LibreOffice Calc 4.2.7 spreadsheet software [7] running on the Windows 7 Professional operating system have been used in this study. It is well-known that QGIS is one of the most popular geographic information systems used worldwide.

Prior to the above mentioned applications for geographical data processing, we have used the ogr2ogr software [8] on the Linux operating system along with Vine Linux 4.2 [9], which is a Linux distribution developed by a Japanese Linux community.

Note that QGIS, LibreOffice Calc, ogr2ogr, and Vine Linux are open source softwares freely available on the Internet.

16.4.4 Data Processing

(1) The vehicle-tracking maps constructed from the G-BOOK telematics data have been provided in the Google map KMZ format. For our analysis, we have first converted the KMZ files to SHP files (i.e., shape-files), which are compatible with ArcGIS using the ogr2ogr software.

(2) Next, the data coordinates have been converted from the terrestrial latitude and longitude to the x and y coordinates in a rectangular coordinate system.

(3) To reduce the computation time, the data file has been clipped to small files containing only the research area.

(4) After merging daily data into weekly data and removing duplicate data, we have been able to calculate the exact usable road distance available for a given week.

In this context, a usable road is one on which at least one vehicle has been probed during the observation period.

The purpose of converting the daily data to weekly data was to smooth the daily fluctuations in the traffic flows.

(5) Next, we have calculated the proportion of the cumulative distance up to the specified date. Note that the cumulative distance up to September 30, 2011 was considered 100 %.

(6) Finally, we evaluated the regional differences in the road usage recovery in the Iwate Prefecture following the 2011 Tohoku Earthquake. We divided the Iwate Prefecture into four areas according to their geographic positions and features, i.e., the Northern Inland, Southern Inland, Northern Coastal, and Southern Coastal areas.

Table 16.1 Regional difference in road recovery (cumulative usable road distances (meters) and ratios)

	Mar 31	Apr 08	Apr 15	Apr 22	Apr 29	May 27	Jun 24	Jul 29	Aug 26	Sep 30
Northern inland	1850873.51	2070475.04	2152283.40	2251706.55	2296780.33	2403134.24	2483876.38	2532471.47	2571798.73	2603629.91
	0.711	0.795	*0.827*	0.865	0.882	*0.923*	0.954	0.973	0.988	1.000
Southern inland	2386739.37	2722218.13	2870724.95	2991566.34	3074037.86	3201569.09	3310978.41	3396994.03	3429255.89	3465835.57
	0.689	0.785	*0.828*	0.863	0.887	*0.924*	0.955	0.980	0.989	1.000
Northern coastal	394082.91	449373.30	471389.15	524694.94	564884.21	589275.55	614317.97	618734.27	668603.07	675390.65
	0.583	0.665	0.698	0.777	*0.836*	0.872	*0.910*	0.916	0.990	1.000
Southern coastal	666655.53	779216.76	814157.35	821573.46	843427.24	855066.71	903310.83	910464.22	920120.72	934580.47
	0.713	*0.834*	0.871	0.879	*0.902*	0.915	0.967	0.974	0.985	1.000

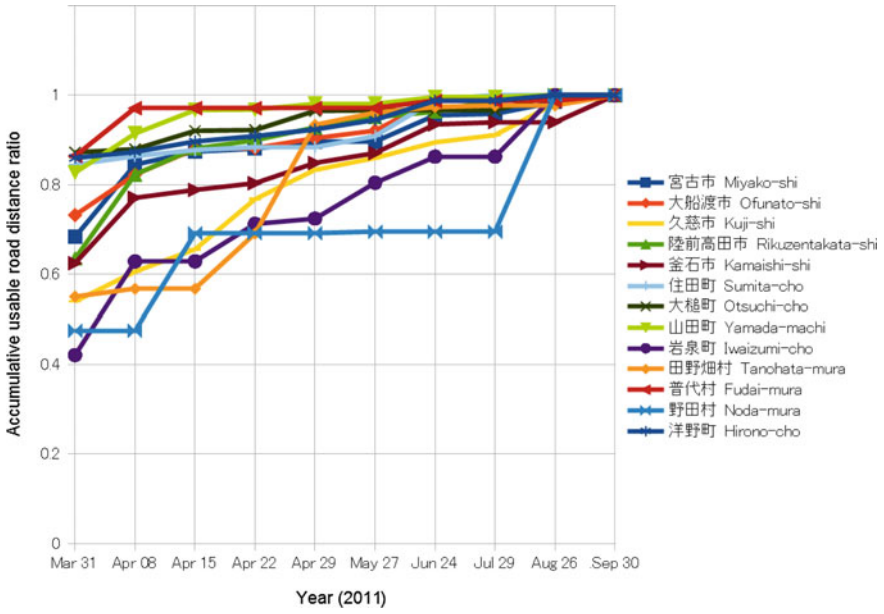


Fig. 16.5 Cumulative usable road distance ratio for the coastal area [3]. The *vertical* scale displays the cumulative distance proportion of the usable roads (relative to the cumulative distance on September 30, 2011) for each date

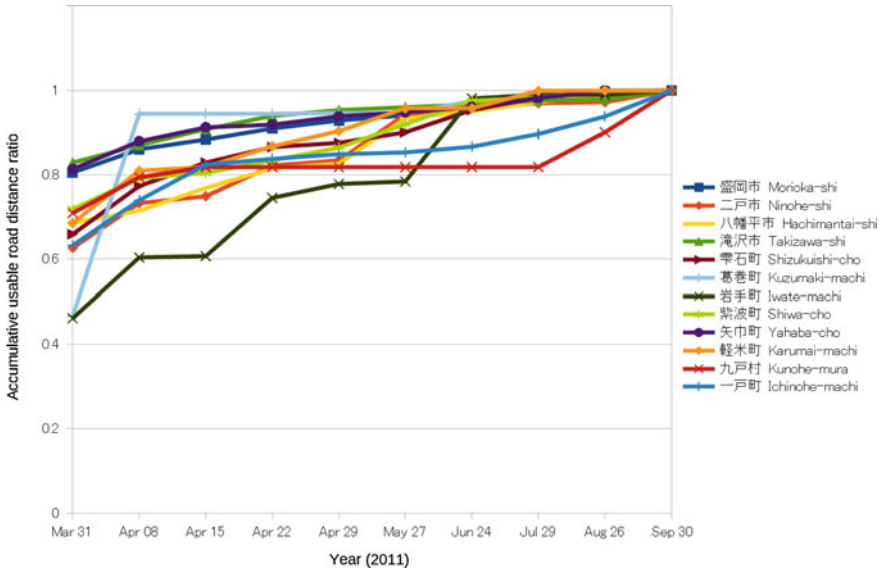


Fig. 16.6 Cumulative usable road distance ratio for the Northern Inland area. The *vertical* scale displays the cumulative distance proportion of the usable roads (relative to the cumulative distance on September 30, 2011) for each date

(7) Furthermore, we identified the municipalities whose road recovery was exceptionally delayed. Subsequently, we determined the related roads by analyzing maps generated using the QGIS software (Fig. 16.5).

16.5 Results

16.5.1 Regional Road Usage Recovery Differences

Defining the cumulative usable distance up to September 30, 2011 as 100 %, the percentages of the usable road distances have the values given in Table 16.1. In this Table 16.1, the upper lines indicate the cumulative usable road distances (in meters) and the lower lines represent the cumulative usable road distance ratios.

(1) The Northern Inland area

It was determined that 80 % of the road distance has been usable by April 15, 2011 and 90 % by May 27, 2011.

(2) The Southern Inland area

It was determined that 80 % of the road distance has been usable by April 15, 2011 and 90 % by May 27, 2011.

The recovery speed in both the Northern and Southern Inland areas have been slightly slower than that in the Southern Coastal area.

(3) The Northern Coastal area

It was determined that 80 % of the road distance has been usable by April 29, 2011 and 90 % by June 24, 2011.

The recovery speed in the Northern Coastal area has been extremely slower compared with that in the Southern Coastal area.

(4) The Southern Coastal area

As we have already described in our previous study 1 [1], 80 % of the road distance has been usable by April 7, 2011 and 90 % by April 29, 2011.

16.5.2 Cities Whose Road Recovery Was Extremely Delayed

(1) The Northern Inland area Fig. 16.6

The ratio increases in the usable distance of Iwate Town, Kunohe, and Ichinohe have been extremely delayed.

We were able to determine related roads by analyzing the maps generated by the QGIS software.

For Iwate Town, the road whose recovery was significantly delayed is the Iwate Prefectural Road number 158 (the Yabukawa-Kawaguchi line).

For Kunohe, the road whose recovery was significantly delayed is the Iwate Prefectural Road number 5 (the Ichinohe-Yamagata line).

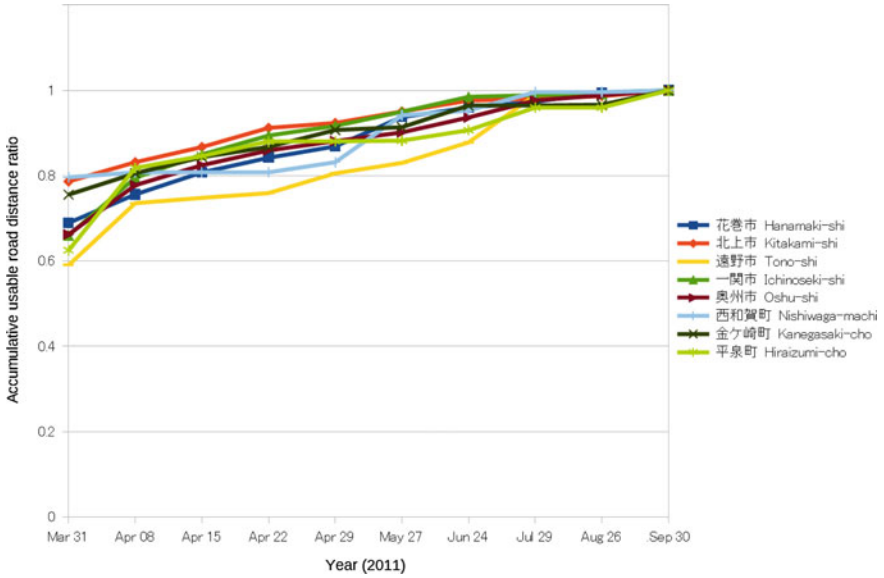


Fig. 16.7 Cumulative usable road distance ratio for the Southern Inland area. The *vertical* scale displays the cumulative distance proportion of the usable roads (relative to the cumulative distance on September 30, 2011) for each date

For Ichinohe, the road whose recovery was significantly delayed is the road forward to the Okunakayama Ski Field.

(2) The Southern Inland area Fig. 16.7

The ratio increases in the usable distance of Iwate Town, Kunohe, and Ichinohe have been extremely delayed.

We were able to determine related roads by analyzing the maps generated by the QGIS software.

For Tono, the road whose recovery was significantly delayed is Iwate Prefectural Road number 160 (Tsuchibuchi-Tatsusobe line).

For Hiraizumi (the town of National Heritage), the roads whose recovery was significantly delayed are the roads forward to the Motsuji Temple, Hiraizumi Kominkan, and Satoyama.

(3) The Northern Coastal area

According to our previous study [3], the ratio increases in the usable distance of Kuji, Iwaizumi, and Noda have been extremely delayed compared with other northern coastal cities.

(4) The Southern Coastal area

As it has been shown in [3] that the ratio increase in the usable distance of Kamaishi has been extremely delayed compared with other Southern Coastal cities.

16.6 Discussion

In this study, we evaluated the regional differences in the road usage recovery in the Iwate Prefecture following the 2011 Tohoku Earthquake. The Iwate Prefecture was divided into four areas, i.e., the Northern Inland, Southern Inland, Northern Coastal, and Southern Coastal areas.

According to the results of our study, we concluded that the recovery conditions of the regional roads following the 2011 Tohoku Earthquake differed based on region. For the Southern Coastal area, as we have shown in the first previous study [1], it was determined that 80 % of the road distance was usable by April 8, 2011 and 90 % was usable by April 29, 2011. According to our findings, for both the northern and Southern Inland areas, 80 % of the road distance was usable by April 15, 2011 and 90 % was usable by May 27, 2011, i.e., the recovery speed in both the northern and Southern Inland areas was slightly slower compared to that in the Southern Coastal area. For the Northern Coastal area, 80 % of the road distance was usable by April 29, 2011 and 90 % was usable by June 24, 2011, which means that the recovery speed in the Northern Coastal area was much slower than that in the Southern Coastal area.

We assume that the recovery difference between the Northern and Southern Coastal areas has been caused by differences in their geographical features. In the Iwate Prefecture, the coastal features observed in the Northern Coastal area are primarily characterized by coastal terraces. Conversely, the coastal features observed in the Southern Coastal area are primarily characterized by the Rias' coast. Nevertheless, further investigations should be performed concerning this point in future studies.

In addition, road recovery efforts have focused more on regions that were heavily affected by the earthquake and subsequently by the tsunami, in particular, the Southern Coastal area. In fact, the road recovery process was intensified in the Southern Coastal area to accelerate rescue operations, provide temporary housing for citizens, and rebuild the infrastructure of the region.

Acknowledgements We would like to thank all the people worldwide who assisted the people of Japan following the 2011 Tohoku Earthquake. In addition, we would like to thank the people who created and maintained the useful software and telematics archives used in this study.

References

1. Komori H, Endo N (2013) Road distance traveled by vehicles following the 2011 Tohoku earthquake, calculated by G-BOOK telematics data. In: Proceedings of the 9th International Conference on Signal-Image Technology & Internet-Based Systems, 870–874
2. Vehicle Tracking Map Constructed by G-BOOK Telematics Data (In Japanese): http://g-book.com/pc/spot/Tohoku-Jishin.asp
3. Endo N, Komori H (2015) Analysis of vehicle tracking maps in iwate prefecture following the 2011 Tohoku earthquake. In: Proceedings of The International MultiConference of Engineers and Computer Scientists 2015. Lecture Notes in Engineering and Computer Science, Hong Kong. 18–20 March 2015, pp 124–128
4. Endo N, Komori H (2015) Regional difference on road recovery in iwate prefecture following the 2011 Tohoku earthquake. In: Proceedings of The World Congress on Engineering and Computer Science 2015. Lecture Notes in Engineering and Computer Science, San Francisco, USA, 21–23 Oct 2015, pp 162–167
5. Hada Y, Suzuki T, Shimora H, Meguro K, Kodama N (2009) Issues and future prospect on practical use of probe vehicle data for disaster reduction–provision of the vehicle tracking map in the 2007 Niigataken Chuetsu-oki Earthquake–(In Japanese). J Jpn Assoc Earthq Eng, 9(2), 148–159. http://www.jaee.gr.jp/stack/submit-j/v09n02/090211_paper.pdf
6. Official website of QGIS, A Free and Open Source Geographic Information System. http://www.qgis.org/en/site/
7. Official Website of the LibreOffice Project. http://www.libreoffice.org/
8. Official website of GDAL: ogr2ogr. http://www.gdal.org/ogr2ogr.html
9. Official website of Vine Linux (In Japanese). http://www.vinelinux.org/
10. Vehicle Tracking Map Constructed by G-BOOK Telematics Data http://g-book.com/disasterMap/ALL/Tohoku-Jishin_YYYYMMDD.kmz (For example, Tohoku-Jishin_20110318.kmz. The data files of tracking map are still available on February 3, 2015.)

Chapter 17
Computational Complexity Study of RSA and Elgamal Algorithms

A.E. Okeyinka

17.1 Introduction

Cryptography is concerned with the study of how to keep secrets secret. Its classical task is to provide confidentiality. However, in recent times, the scope of cryptography has expanded beyond issues of confidentiality. Its domain now covers the study of techniques for message integrity, identity, authentication, digital signatures and so forth. The rapid growth of electronic communication means that issues in information security are of increasing practical importance [1]. Many cryptographic algorithms have been developed among which are the following:

A. *RSA:* This is a public key. It is a bijective function and computationally efficient. It was designed by Rivest, Shamir, and Adleman.
B. *ElGamal*: ElGamal is a discrete logarithm algorithm. It is a one-way function, and contains no trap door.
C. *DES:* This is Data Encryption Standard. It uses a 56-bit key and operates on 64-bit blocks of data.
D. *HASH*: This is also known as 'fingerprint' or 'message digest'. It is used for computing condensed representation of a fixed length message.
E. *MD5:* This is a 128-bit message digest function, developed by Ron Rivest.

A cryptographic algorithm is a set of mathematical rules used in encryption and decryption. In addition to securing data being communicated, there is also the need to ensure that the data which is communicated is authentic. A digital signature is a means of ensuring that an electronic document is authentic. "Authentic" in this

He was formerly with Ladoke Akintola University of Technology, Ogbomoso, Nigeria.

A.E. Okeyinka (✉)
Computer Science, Landmark University, Omu-Aran, Nigeria
e-mail: ae.okeyinka@gmail.com

S.-I. Ao et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-981-10-2717-8_17

context implies that the receiver knows the person who created the message, and he knows that the message has not been altered since it was created. A digital signature mechanism consists of an algorithm for generating the signature as well as an associated verification algorithm. Digital signatures are designed to provide authentication and also non-repudiation. In this study, the RSA and Elgamal algorithms including their digital signatures are implemented and compared.

17.1.1 Classification of Computational Problems

The categories of computational problems include the following:

a. Solvable problems
b. Unsolvable problems
c. Solvable but impracticable problems
d. Solvable but intractable problems

A problem is regarded as computationally solvable if there exists at least an algorithm for solving it. This implies that all the problems that we have been able to solve so far using the computer belong to this category. On the other hand, a problem for which there exists no known algorithm at the moment for solving it is an unsolvable problem. A well-known example of the problems in this category is the halting problem. The halting problem can be stated as follows: "Given a description of an arbitrary computer program, decide whether the program finishes running or continues to run forever". This implies: given a program and an input, decide whether the program will halt or will run forever. The halting problem is a decision problem and has been proved to be an unsolvable problem (so far).

Next, some problems are solvable but impractical. The impracticability of solving them arises from the amount of computational resources required to solve them. The enormity of such resources is beyond both the computer and human being. A problem that could take today's fastest computer billions of millions of years to solve is an example. Usually, problems in this category are called combinatorial problems. They are problems whose solution spaces explode exponentially; most of them could be modeled as the Traveling Salesman Problem (TSP). They are very simple to characterize but difficult to solve. They are classified as NP-complete problems. In real life, heuristics are used to solve problems that are solvable but impractical. Heuristics give solutions that are not optimal, but good enough, usually such solutions should be at least 90 % optimal. In a few cases, heuristics could yield optimal solutions, though they are not guaranteed to do so. However, there is hope that the problems in this category might become solvable someday in the future when faster computers or algorithms or both are invented.

The other category consists of problems that are solvable but intractable. Problems that are solvable but intractable will remain impracticable as it has been proved that no amount of faster algorithms or invention of faster computers could make them practical. Such problems are better left unattended to since no amount of

thinking can proffer solutions to them. Formally, the subject of computational complexity theory classifies these computational problems as follows:

a. P: These are problems that can be solved in polynomial time ("P" stands for polynomial)
b. NP: This stands for "non-deterministic polynomial time" where non-deterministic is just a way of talking about guessing a solution. A problem is in NP if we can quickly, in polynomial time; test whether a solution is correct without worrying about how hard it might be to find the solution. Problems in P and NP are relatively easy to solve. It should be noted that NP does not stand for "non-polynomial".
c. PSPACE: Problems that can be solved using a reasonable amount of memory without regard to how much time the solution takes.
d. EXPTIME: These are problems that can be solved in exponential time. This class contains everything in the classes P, NP and PSPACE.
e. NP-HARD: NP-hard is a class of problems that are informally, at least as hard as the hardest problems in NP. A polynomial-time algorithm for an NP-hard problem implies a polynomial-time algorithm for every problem in NP.
f. NP-Complete: A problem is NP-Complete if it is both NP-Hard and an element of NP (or NP-easy). NP-Complete problems are the hardest problems in NP. If anyone finds a polynomial-time algorithm for even one P-Complete problem, then that would imply a polynomial-time algorithm for every NP-complete problem.
g. CO-NP: This is the opposite of NP. If the answer to a problem in CO-NP is No, then there is a proof of this fact that can be checked in polynomial time. For example, the circuit satisfiability problem is in NP. If the answer is YES then any set of m input values that produces TRUE output is a proof of this fact; we can check the proof by evaluating the circuit in polynomial time. It is widely believed that circuit satisfiability is not in P or CO-NP, but nobody actually knows.

Every decision in P is also in NP. If a problem is in P, we can verify YES answer in polynomial time re-computing the answer from scratch. Similarly, any problem in P is also in CO-NP. One of the most important open questions in theoretical computer science is whether or not P = NP. Nobody knows.

The class of problems that can be efficiently solved by quantum computers is called BQP, for "bounded error, quantum, polynomial time". Quantum computer only runs probabilistic algorithms, so BQP on quantum computers is the counterpart of BPP ("bounded error, probabilistic, polynomial time") on classical computers. It is defined as the set of problems solvable with a polynomial-time algorithm, whose probability of error is bounded away from one half [2].

A quantum computer is said to solve a problem if, for every instance, its answer will be right with high probability. If that solution runs in polynomial time, then that problem is in BQP. BQP is suspected to be disjoint from NP-complete and a strict superset of P, but that is not known. Both integer factorization and discrete log

are in BQP. Both of these problems are NP problems suspected to be outside BPP, and hence outside P. Both are suspected not to be NP-complete. There is a common supposition that quantum computers can solve NP-complete problems in polynomial time. That is not known to be true, and is generally suspected to be false [3].

17.2 Research Motivation

In addition to creating new algorithms to solve problems that are so far regarded as unsolvable or impractical, the research gradient in computational complexity is also skewing towards algorithmalgorithm efficiency. It is not enough to invent an algorithm; indeed, considering the computational efficiency of such an algorithm vis-à-vis the existing ones professing to do same task is of great importance. Many cryptographic algorithms abound but they are not equally efficient. In that case there is need to measure and compare their level of computational efficiency. Doing so, would enable us to know which of the algorithms should be used in specific situations for overall maximum efficiency. Furthermore, a reflection on their performance may suggest the need for more study of the algorithms to establish whether or not a more efficient algorithm could be obtained by hybridization or concatenation. So far it has been established that RSA is more energy efficient than Elgamal [4]; however, other performance parameters need to be investigated and studied.

In this paper, RSA and ElGamal algorithms including digital signatures are studied. The choice of these two algorithms is not arbitrary. RSA is a classical technique and most security systems in use today were based on RSA. In short RSA appears to be the most acceptable technique for securing electronic data communication. RSA, was proposed in 1977, [5]. It was patented in the United States by Massachusetts Institute of Technology (MIT) September 20, 1983, [5]. Although the patent for RSA expired September 21, 2000, [6], RSA has become the most popularly implemented public-key cryptosystem [5]. Elgamal on the other hand was proposed in1985, [7]. It is an extension of the Diffie-Hellman key agreement protocol. It is a non-deterministic algorithm, [7]. So the goal of this study is to determine and compare the complexity of RSA and Elgamal algorithms; given that RSA is deterministic and Elgamal is non-deterministic.

17.3 Research Methodology

Both the RSA and Elgamal cryptographic algorithms with digital signature are implemented using C# programming language on the same programming environment. Each algorithm consists of three phases: Key generation, Encryption and decryption, Signing and verification.

The C# program takes as inputs ten different text data one by one; each character of the text document is converted into its ASCII form and used appropriately in the algorithms in computing cipher text information, which is sent to the receiver by the sender. The cipher text information received by the receiver is decrypted by the module meant for that in order to extract the original message. The length of the text used as input is automatically determined by the C# code. The signature generation, and signature verification module of the code determines the validity of the signature. The execution times of each input text as a whole are observed using the computer internal clock for both Elgamal and RSA algorithms. The execution times are compared to determine which of the two algorithms is more computationally efficient.

17.4 RSA and Elgamal Algorithms

The two algorithms (Elgamal and RSA) are presented below:

17.4.1 Elgamal Algorithm

Elgamal cryptosystem requires a modular exponentiation operation. The security strength of the cipher is a function of the sizes of the modulus; it is based on the discrete logarithm.

(i) Key Generation

This process generates required keys (private key and public key) for both encryption and decryption. The algorithm is stated as follows:

(a). Generate a large Prime number p
(b). Choose a Generator number a subject to the following conditions

$1 < a < p\text{-}1$

To ensure that the value of a picked is a generator number, additional conditions have to be considered as follows

(a). Find $Ø = p - 1$
(b). Find all the factors of Ø i.e. (f1, f2…fn).
(c). a is a generator number if and only if $wi = a^{Ø/fi} \bmod p <> 1$, for all qi
(d). Choose an integer x such that $1 < x < p - 2$, x is the private key
(e). Compute $d = a^x \bmod p$
(f). Public key information = (p, a, d)
Private key = x

(ii) Encryption and Decryption Algorithm

The encryption is done using the public key information while the Decryption is done using the private key information.

(1) Encryption:

(a). The sender receives the public key information only, which will enable her to encrypt
(b). The sender encodes the message m by converting its string representation to its corresponding numerical value
(c). The sender chooses an integer k such that $1 < k < p - 2$
(d). The sender computes $y = a^k \bmod p$
(e). The sender also computes $z = (d^k * m) \bmod p$
(f). The sender then sends the cipher text information $C = (y, z)$ to the receiver

(2) Decryption
The following steps are taken to decrypt a cipher text:

(a). The receiver needs the private key x to decrypt
(b). The receiver picks up the cipher text information $C = (y, z)$ He then computes $r = y^{p-1-x} \bmod p$
(c). The receiver finally computes $m = (r * z) \bmod p$ to extract the original message

(3) Signing and Verification Algorithm
This signature process aims at signing a message to ensure message authentication and integrity. There are two processes involved in this section, they are; Signature Generation Process Signature Verification Process.

(4) Signature Generation Process: The sender should do the following:

(a). Pass the numerical representation m of the message into an hash function to produce an hashed message M (i.e. M = hashfunction(m))
(b). Choose a secret key x such that $1 < x < p - 1$
(c). Choose random integer k with $1 <= k <= (p - 1)$ and $\gcd(k, p - 1) = 1$ (gcd is the grand common divisor)
(d). Compute $h = a^k \bmod p$
(e). Compute: $k\_Inverse = k^{-1} \bmod p$
(f). Compute the value: $s = (M - (x * h)) * - k\_Inverse \bmod (p - 1)$
(g). Compute $s = (p - 1) - s$

(5) The signature is the tuple (h, s)
For the signature verification process, the receiver should do the following

(a). Collect the signature (h, s)
(b). Compute $b = a^x \bmod p$

(c). Compute var1 = ($b^{h} * h^{s)}$ mod p
(d). Compute var2 = a^{M} mod p

If var1 == var2 then signature is valid otherwise invalid.

17.4.2 RSA Algorithm

The security of RSA is inherent in the difficulty of factoring large numbers. The RSA encryption and decryption algorithms require a single modular exponentiation operation. The size of the modulus determines the security strength of the cipher [8].

(i) Key generation
The algorithm is stated as follows;

(a). Generate two large random (and distinct) primes p and q, each roughly the same size.
(b). Compute n = p.q and Ø = (p-1) (q-1).
(c). Select a random integer e, where $1 < e < Ø$, such that the greatest common divisor, gcd(e, Ø) = 1.
(d). Use Extended Euclidean algorithm to compute the unique integer where $1 < d < Ø$, such that $e \equiv 1 (\text{mod } Ø)$.
(e). Sender's public key is (n, e) and private key is d.

(ii) Encryption and Decryption

(a). Encryption: This is done using the public key (n, e). $c = m^{e} \text{ mod } n$. m is the message.
(b). Decryption: This is done using the private key (d, n), $m = c^{d} \text{ mod } n$.
(c). Signing and Verification: This has two stages which are the signature generation and signature verification.
(d). Signature Generation: The Sender should do the following:
i. Compute m = R (m), an integer in the range [0, n-1].
ii. Compute $s = m^{d} \text{ mod } n$.
iii. Sender's signature form is s.
(e). Signature verification: To verify Sender's signature:

Compute var1 = m mod n
Compute var2 = s^{e} mod n

If var1 = var2, Signature valid else invalid

17.5 Results

The results obtained are shown below as figures and tables. (Figures 17.1, 17.2, 17.3 and Tables 17.1, 17.2, 17.3)

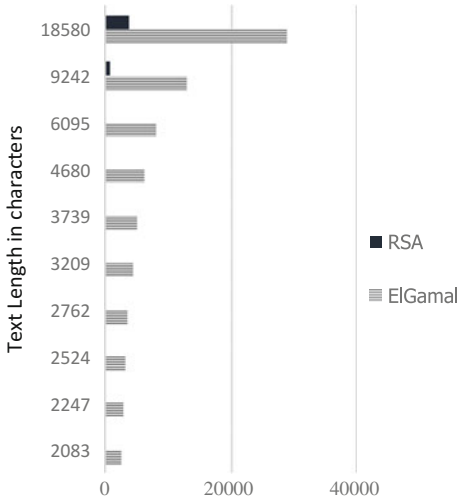


Fig. 17.1 Execution times for encryption and signing [9]

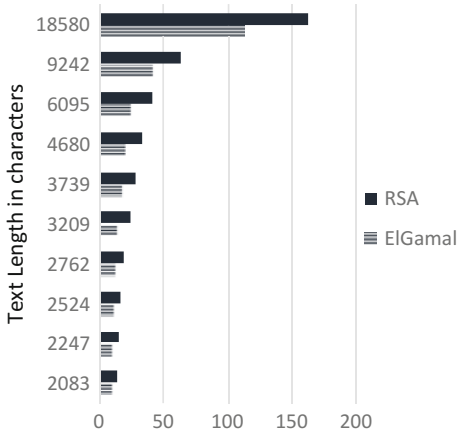


Fig. 17.2 Execution times for decryption [9]

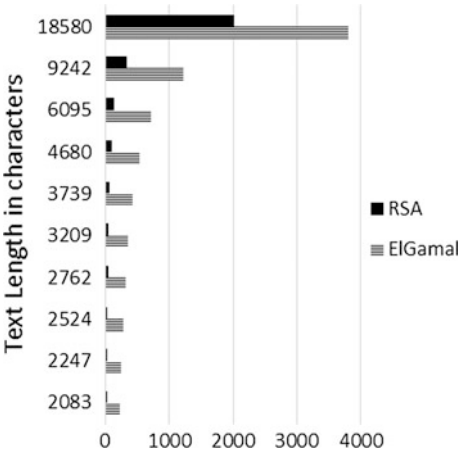


Fig. 17.3 Execution times for signature verification [9]

Table 17.1 Execution times for encryption and signing

Text Length in characters	ElGamal (milliseconds) (seconds)	RSA (milliseconds) (seconds)
18580	29083.6349 (29)	3818.8579 (3)
9242	13232.2388 (13)	641.4635 (0)
6095	8380.3855 (8)	224.8633 (0)
4680	6502.8839 (6)	142.941 (0)
3739	5199.6898 (5)	108.6065 (0)
3209	4462.4788 (4)	83.7358 (0)
2762	3731.9332 (3)	67.6062 (0)
2524	3490.1043 (3)	62.7788 (0)
2247	3115.9792 (3)	56.3892 (0)
2083	2892.6267 (2)	49.9453 (0)

Table 17.2 Execution times for decryption

Text Length in character	ElGamal (milliseconds)(seconds)	RSA (milliseconds) (seconds)
18580	111.9454 (0)	162.4227 (0)
9242	40.795 (0)	63.6866 (0)
6095	23.8803(0)	40.5696 (0)
4680	19.4903(0)	32.777 (0)
3739	16.9437(0)	28.0812 (0)
3209	13.2764 (0)	23.5664 (0)
2762	12.1137(0)	19.2534(0)
2524	10.3053 (0)	16.4246 (0)
2247	9.4394 (0)	14.9514 (0)
2083	8.5182 (0)	14.3429 (0)

Table 17.3 Execution times for signature verification

Text Length in characters	ElGamal (milliseconds) (seconds)	RSA (milliseconds) (seconds)
18580	3803.3193 (3)	2013.0053 (2)
9242	1216.9236 (1)	332.1506 (0)
6095	705.5812 (0)	133.1629 (0)
4680	526.0838 (0)	94.382 (0)
3739	413.8463 (0)	61.7007 (0)
3209	348.0134 (0)	50.902 (0)
2762	302.8908 (0)	39.5771 (0)
2524	277.0682 (0)	32.6042 (0)
2247	242.8458 (0)	27.1777 (0)
2083	224.3202 (0)	27.0507 (0)

17.6 Analysis of Results

The Execution times for both the Elgamal and RSA algorithms are shown on the Tables and Figures. The times are measured in milliseconds, but converted to seconds as displayed on the result templates. We observe and deduce as follows from the results obtained.

In the encryption and signing process, the RSA performs better than Elgamal in all cases. In the decryption process, the Elgamal outperforms RSA; meaning that text messages are decrypted faster by Elgamal than does the RSA technique. In the signature verification process, the RSA again performs better than the Elgamal approach.

When viewed as a single tool, the RSA is superior to the Elgamal algorithm in terms of computational speeds. This, in part, explains why the RSA algorithm has been and is still being used in designing many security protocols for data communication.

17.7 Conclusion and Future Research

From this study, we have observed that even though the RSA is superior to the Elgamal on the overall assessment, it is not as efficient as Elgamal when the rate of data decryption is considered. It is therefore fathomable that a platform that will hybridize both approaches may yield a more efficient technique than either the Elgamal or RSA algorithm. Hence efforts at designing a hybrid algorithm of these two techniques are strongly recommended as candidates for further research work. Furthermore, other performance evaluation parameters apart from energy and speeds may be investigated. Measures such Halstead, Cyclomatic, Lines-of-code and related ones could be computed, to enable us conclude with greater probability which of RSA and Elgamal algorithms is more efficient for pragmatic purposes.

Acknowledgement The efforts of Oyewole Samson Opeoluwapo in computer programming are greatly recognized and appreciated.

References

1. Delfs Hans, Knebl Helmut (2007) Introduction to Cryptography. Springer-Verlag, Berlin Heidelberg
2. Quantum Computer: www.en.wikipedia.org/wiki/Quantum_computer, Accessed on 7th June 2016
3. Bernstein E, Vazirani U (1997) Quantum complexity theory. SIAM J Comput, 26(5)
4. Luo X, Zheng K (2004) Encryption algorithms comparisons for wireless networked sensors. In: IEEE Int Conf Syst Man Cybern Coll Comput Sci. Zhejiang University, China
5. International Association of Engineers [Online]. Available: http://www.iaeng.org (Elbirt AJ (2008) Understanding and applying cryptography and data security. Auerbach Publications, Taylor and Francis Group)
6. Rivest R (1996) The MD5 Message-Digest Algorithm Internet Request for Comments, presented by Rump session of Advances in Cryptography—CRYPTO'91
7. El Gamal T (1985) A public key cryptosystem and a signature scheme based on discrete logarithms. IEEE Trans Inf Theory 31(4):469–472
8. Padmavathi D, Shanmuga Priya M (2009) A survey of attacks, mechanisms and challenges in wireless, sensor networks. Int J Comput Sci Inf Secur 4 (1, 2)
9. Okeyinka AE (2015) Computational speeds analysis of RSA and ElGamal algorithms on text data. In: Proceedings of The World Congress on Engineering and Computer Science 2015, WCECS 2015. Lecture Notes in Engineering and Computer Science, 21–23 October, 2015, San Francisco, USA, pp 115–118

Chapter 18
Ontology-Based Context Modeling and Reasoning for a Smart Space

Moeiz Miraoui

18.1 Introduction

Smart spaces are a research field under the banner of pervasive and ubiquitous computing where devices dynamically and proactively (without explicit intervention of users) adapt their behavior to current context changes. So far, there is no explicit definition of smart spaces which are also called "smart environments" or "ambient intelligence". The most well-known definition of smart spaces is the one gave by D. Cook and S. Das of the MAVHome project in their book [1]: "Smart space is able to acquire and apply knowledge about its environment and to adapt to its inhabitants in order to improve their experience in that environment". The number of smart appliances and devices in the home and office has grown dramatically in recent years thanks to the tremendous evolution of embedded systems. Unfortunately, each component usually performs a single function and there is no synchronization with other components or the environment. The principal objective of researches in this field is to move from environments filled with smart devices to smart environments. Previous solutions proposed suffer from three main drawbacks: (a) interaction and cooperation between devices is missing, (b) context definition and modeling are not clear enough and specific to some particular aspects of smart spaces, (c) the services adaptation task is not context-aware or do not deal in depth with context-awareness which makes the devices able of acting autonomously on behalf of users. In our previous work [2], our aim was to propose an ontology-based modeling of both smart living room environment and contextual information which enables common understanding of context and enhances its sharing. As enhancement, we propose in this paper a more complete process consisting of both ontology-based context modeling and reasoning on explicit

M. Miraoui (✉)
University of Gafsa, Gafsa, Tunisia
e-mail: moeizmiraoui@gmail.com

S.-I. Ao et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-981-10-2717-8_18

context in order to deduce implicit context to improve the quality of services by well understanding context.

The remainder of this paper is structured as follows. In Sect. 18.2 we discuss some related works. Section 18.3 describes the overall environment of an exemplary smart living room as example of a smart space. Section 18.4 presents the ontology-based modeling of the smart living room and contextual information. Section 18.5 presents the context reasoning technique. Finally, Sect. 18.6, draws the conclusions and future work of this paper.

18.2 Related Work

The development of an intelligent, independent and adaptable to changing conditions environment is a goal that has existed for decades. Research on this topic is more intense in recent years. Several researchers have worked on smart spaces such as homes, offices, universities, hospitals, hotels, cars and other private or public places. Context-awareness is an important characteristic of smart spaces and context modeling is a basic step for the development of such spaces. In this section we will focus only on ontology-based context modeling approaches in smart spaces. Among the earliest works on ontology-based context modeling, the SOUPA (standard ontology for ubiquitous and pervasive applications) [3]. SOUPA is composed of two sets of ontologies: SOUPA core ontologies and SOUPA extension ontologies. Core ontologies try to define a generic vocabulary which is universal for different pervasive computing applications. Extension ontologies (extended from SOUPA core ontologies) define additional ontologies to support specific kind of applications and provide examples for extending future ontologies. Pantsar-Syvaniemi et al. [4] proposed a novel context ontology for smart spaces that exploits some parts from SOUPA and enhances context-awareness with three dimensions: physical, digital and human. CoBrA (context broker agent) [5] is an architecture to support the development of context-aware systems in an intelligent space (smart house, smart car, etc.…). In CoBrA where defined ontologies collections called COBRA-ONT for modeling contextual information in such intelligent spaces. COBRA-ONT are expressed with OWL which defines typical concepts associated with places, agents and events. Ontology plays a crucial role in CoBrA, it helps the context broker (server) to share contextual information with other agents and permit him to make context reasoning. The ontology is categorized in four classes: (1) physical place ontology, (2) agent (human or software) ontology, (3) agent localization agent ontology, (4) agent activity context ontology. SOCAM (Service-Oriented Context-Aware Middleware) [6] is an architecture for the building and rapid prototyping of context-aware services in intelligent environment. The context ontologies are divided into upper ontology which captures general context knowledge about the physical world and composed of: computational entity, location, user and activity. The domain-specific ontologies which are a collection of low level ontologies that defines the details of general concepts and

their properties in each sub-domain. Abdulrazak et al. [7] presented an ontology for smart environments that takes into account the new notions namely the referentiality in terms of localization and temporality, and the problem of the environmental change. The design of their ontology was based on the general idea which states that, a being lives and interacts in an environment with a certain dynamic. From there, they extracted the major concepts of their top level ontology, which are: Being, Environment and Dynamic. Saleemi et al. [8] modeled and processed context information using their development tool and Nokia's Smart-M3 architecture. They referred to all information that characterize the situation of a user as his context. They divided the user's context in two broad categories, namely atomic context and inferred context. They modeled user's context using six context dimensions: Time, Locality, Devices, Activity, Occupancy and Associations. They have defined an inference rule using a 3-clauses pattern. Ming Li [9] proposed a context model based on ontology for smart space composed of the top layer ontology model and the specific field onto- logy model. The top layer ontology model is the abstract of context information in smart space composed of seven classes: Person, Location, Time, Activity, Physical Entity, Virtual Entity, and Service, the specific field ontology model is the embodiment of the top layer ontology model in the specific field. He proposed a rule-based reasoning system for the adaptation task. Haesung Lee and Joonhee Kwo [10] modeled contexts which are generally distributed in smart home environment using the ontology technology. The set of modeled contexts includes home domain-based contexts and social relationship-based contexts. They tried to provide smart home-based health care services to a specific user in pervasive and seamless way using context awareness computing technique. They defined three general categories for context items: Biomedical sensor data, Environmental sensor data and Social relationship data. The context model is made of four sub-ontologies: (1) the user domain ontology (2) the home domain ontology (3) the function management ontology (4) the social context ontology. They defined a set of first-order rules to determine if a function has to be triggered. Zachary Wemlinger and Lawrence Holder [11] presented the Casas Ontology for Smart Environments (COSE). They presented a smart environment domain ontology, where the main concepts in the COSE ontology are: buildings, occupants, sensors and human activities. Moji Wei et al. [12] presented an ontology-based Home Service Model to retrieve and invoke services according to user's needs automatically. In the ontology, they differentiated function from need which are usually confused with each other. They proposed an upper ontology as a fixed viewpoint, two domain ontologies which are Function Concept Ontology and Context Concept Ontology. They used Maslow's hierarchy of needs as guidance and consult existed services to construct their function concept ontology for service annotation. Wan-rong Jih et al. [13] proposed a Context-aware Service Platform, implemented in JADE agent platform, and utilize Semantic Web technologies to analyze the ambient contexts and deliver services. They integrated ontology and rule-based reasoning to automatically infer high-level contexts and deduce a goal of context-aware services. They defined a context ontology as a representation of common concepts about the smart space environment and an

if-then rules for the reasoning. Xinhua Zhu et al. [14] proposed a system modeling for a Smart-home Healthy Lifestyle Assistant System (SHLAS). They introduced an Ontology-based domain knowledge and context model to capture and represent the agents, and agent behavior which provides agents with reasoning ability. Through behavioral analysis, habits can be acknowledged and personalized services can be recommended to help the member achieve a healthy lifestyle. Chao Li et al. [15] Developed a context-aware lightning control system for smart meeting rooms. They used an ontology-based context modeling approach and a rule based system for context reasoning. Chahuara et al. [16] Presented an audio-controlled smart home based on a framework composed of knowledge representation module using a two level ontology, a situation recognition module based on the SWRL logic reasoner and a decision making module based on the Markov logic network (using weighted logic rules) to deal with uncertainty and imprecision of context information. Jih and Yung-jen Hsu [17] proposed an agent-based architecture for building context-aware systems in a smart space composed of an ontology-base context modeling, an OWL DL reasoner and an AI planner for the services plan generation.

Most of proposed approaches of ontology-based context modeling do not offer a complete description of contextual information and do not cover all aspects of context in a smart space. Most of them were not based on a clear and concise definition of context. In order to propose more extensible and reusable ontologies, most of them suggest the use of two categories of ontologies. One composed of basic ontology concepts and the other is adapted according to the application domain. The core ontology (basic) differs from one method to another and its basic classes depend on the definition of context adopted by authors. The proposed models fail in presenting a generic ontology for context which cover all aspect of context in a smart space which could limit their usability and extensibility. Most of previous works have neglected the context reasoning phase which could enormously improve the context-awareness aspect.

18.3 The Smart Living Room

18.3.1 Description

An exemplary smart living room is composed of a set of equipment of three main types: (a) appliance which can be smart TV, satellite receiver, heater, air conditioner and home cinema, (b) Furniture which can be window blinds, a set of light bulbs and sofa, (c) communication device which can be a land-line phone or the user's mobile phone. All these devices should provide a set of services through different forms (or modes) to the user (s) occupying the living room. These services should take into account the user's preferences and triggered according to the current context collected from different sensors installed in the living room (Fig. 18.1).

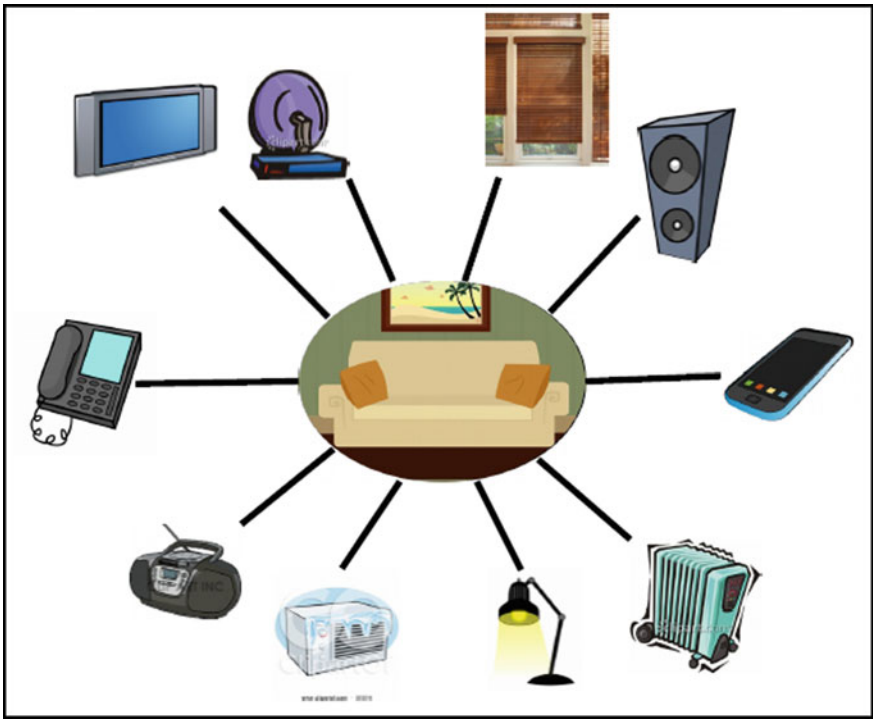


Fig. 18.1 Basic components of an exemplary living room

18.3.2 Sample Operation Scenario

In the initial state of the living room (empty), everything is off (TV, radio/music player, light, air conditioning, heater, window blinds closed, etc.). At the entrance of the user and detection of his presence, the lighting system composed of the window blinds and light bulbs starts first to adjust the light to the preferences of the user and the current context. Then, the climate system composed of an air conditioner and heater starts to adjust the room temperature and air condition to the user's preference and according to the current context. The former system will trigger only if the system perceives at least one user seated on the sofa to avoid wasting energy if the user makes a simple entry and exit from the living room without the intention of staying. After a small period of time (δT), the multimedia system then starts to turn on the TV or radio/music player and home cinema depending on the preferences of the user if he is alone or surrounded. Finally, after another small period of time (ΦT), the communication system starts to provide services tailored to the user's context, such as forwarding the ring of an incoming call to the home cinema if the noise is too high in the living room or display a picture of a phone on the TV screen in order to attract the user's attention.

The dynamic aspect of the smart living room described in the above scenario can be modeled using a simple timed automaton (Fig. 18.2) to show the state change over time. There are two main states of the smart living room: empty and occupied. In the first state (i.e., initial and final state), the network sensors perceive that there is no one in the living room, so all the equipment should be off. As soon as the network sensors perceives the presence of one or more user, the smart living room moves to the second state, which in turn contains four sub-states that describe the preferred order of system booting. The light system will start first (occupied 0), and after perceiving at least one user seated on the sofa, the climate system (occupied 1) will start. Then, after a small period of time (δT), the multimedia system will start (occupied 2), and finally after another small period of time (ϕT) the communication system will start (occupied 3). At any time during the occupied state of the smart living room, if the network sensors perceive that there is no body in the room (presence = 0), the smart living room will transit directly to the empty state.

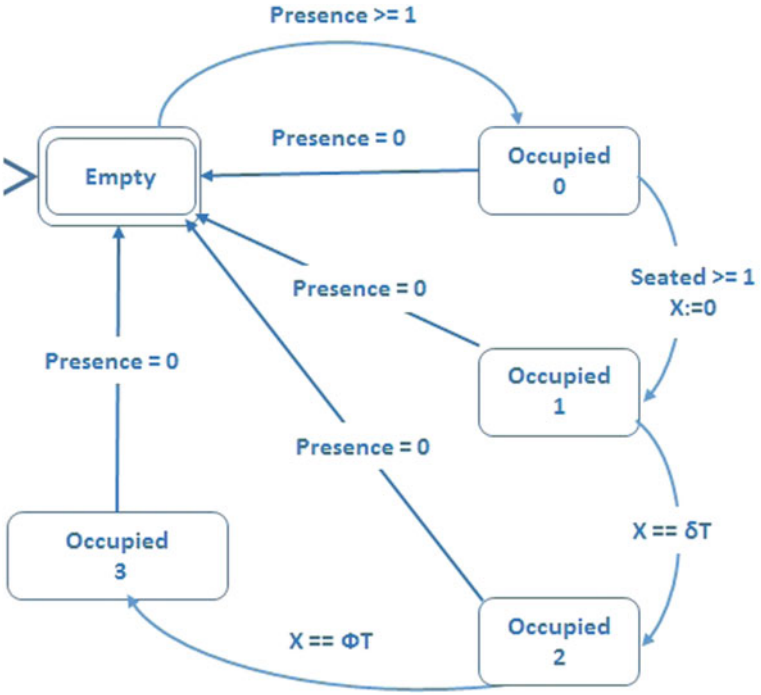


Fig. 18.2 Modeling of the dynamic aspect of the smart living room

18.4 Ontology-Based Modeling

Ontology is a set of structured concepts that are organized in a graph where relations can be either semantic relations or composition and heritage relations. The main objective of ontology is to model the set of knowledge of a given domain. This means to choose a manner for describing domain information in a form comprehensible by computers. It provides a representative vocabulary for a given domain and permits a consistent interpretation of data [18]. Ontology is a powerful tool for knowledge sharing, reuse and expression of complex situations compared to other data modeling techniques such as key-value, object oriented, logic, mark-up schema and graphical models [19, 20]. In addition, ontology supports formal logical reasoning over the situation ontology for consistency checking, subsumption reasoning and implicit knowledge inference.

18.4.1 Smart Living Room Modeling

A smart living room is composed of a set of equipment which can be appliances, furniture or communication devices. Each equipment provides a set of services to the users (occupants) through several forms taking into account their preferences. The basic concepts of a smart living room that we can extract from the description and the scenario above are: sensors, light system, cooling and heating system, multimedia system, communication system, user, user's preferences, services, service's forms and context. These concepts are related to each other in some ways. The context uses the user's preferences and perceived by the sensors network to trigger the light system, cooling and heating system, multimedia system and communication system. Each of these systems has a set of subclasses and can provide several services through different forms. Figure 18.3 shows a simplified ontology-based model of the smart living room and Fig. 18.4 shows a detailed one.

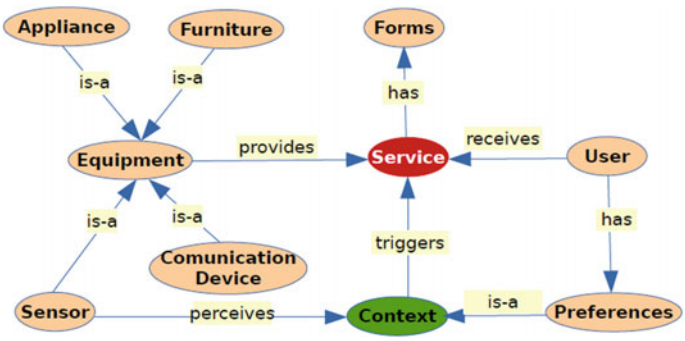


Fig. 18.3 Simplified ontology-based model of the living room

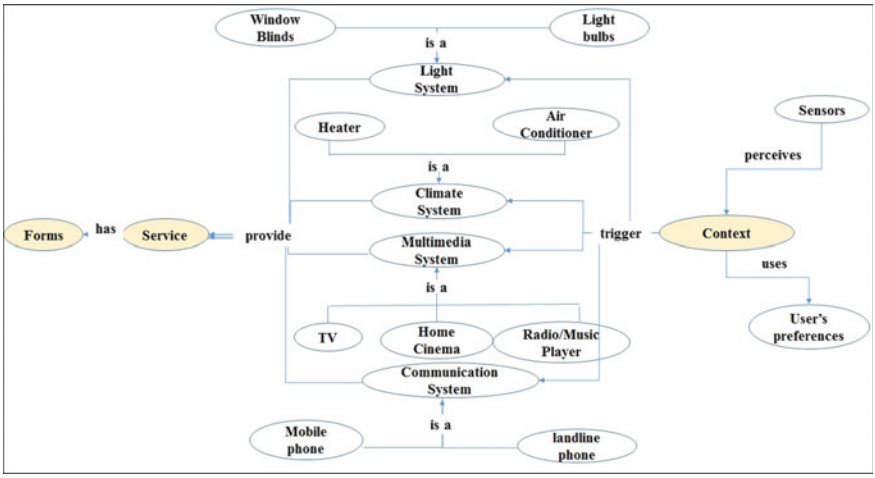


Fig. 18.4 Detailed ontology-based model of the living room

18.4.2 Context Modeling

Context-aware applications should provide proactively (without explicit user's intervention) adapted services to both users and applications according to the current context. Context-awareness is a basic concept of pervasive and ubiquitous computing as well as smart environment or ambient intelligence. From this reveals the importance of the concept of context which considered as a specific kind of knowledge that requires a good and clear understanding and specially a precise establishment of its components. One basic step before context modeling consists of clearly define context.

Several definitions of context were proposed. Some of them were based on enumerating contextual information (localization, nearby people, time, date, etc.) like those proposed in [21–23]. Others were based on providing more formal definitions in order to abstract the term like the one proposed by Dey [24]. Most of these definitions were specific to a particular domain such as human-computer interaction and localization systems and need an adaptation effort to be used in a particular application. In our previous work [25] we have made a survey of existing definitions of context and proposed a service-oriented definition of context as follows: "Any information that triggers a service or changes the quality (form or mode) of a service if its value changes." This definition is sufficiently abstract and helps to limit the set of contextual information. The definition was based on the concept of service because such one plays a crucial role in the operation of a pervasive computing system and can be easily adopted in the context of smart spaces.

Based on this definition, the modeling process starts (first step) by specifying for each equipment of the smart living room the set of services that can be provided.

Table 18.1 Service triggering information

Equipment	Provided service	Triggering information
Window blinds	Lighting	User's presence
TV and satellite receiver	Entertainment	Seated user
Home cinema	Entertainment	Seated user
Heater	Heating	Seated user
Air conditioner	Air conditioning	Seated user
Radio and music player	Entertainment	Seated user
Light bulbs	Lighting	User's presence
Phone	Communication	Incoming call

For each service we should specify also the set of information which their change of values will trigger the service (Table 18.1).

The second step of the modeling process consists of specifying for each service the set of forms through which the services can be provided. We should also specify for each form of service the set of information which their change of values will change the form of a service. The set of these forms are fixed extensible according to how the living room should operate (Table 18.2).

The third step of the modeling process consists of making the union of the two previous sets to get the final list of contextual information. This information will

Table 18.2 Service form changing information

Equipment	Service's forms	Changing form information
Window blinds	Closed, mostly closed, half-opened, mostly opened, totally opened	User's presence, indoor light, outdoor light, User's preference
Light	Off, low light, average light, high light	Indoor light, outdoor light, day type, time, user's preference
TV and satellite receiver	Off, on preferred channel, on other channel	Day type, time, surrounding, user's preference
Home cinema	Off, on low, on average, on high	Date, time, surrounding, user's preference
Heater	Off, on preferred, on other	Temperature, surrounding
Air conditioner	Off, on preferred, on other	Temperature, surrounding
Radio and music player	Off, on preferred radio station, on preferred song, on other	Day type, time, surrounding, user's preference
Phone	Default mode, divert to answer machine, divert to home cinema, divert to home cinema with icon on TV	Noise, surrounding, time, light, day type

Table 18.3 context elements and their possible values

Context element	Possible values
Seated user	0, 1, several
User's presence	Yes, no
Indoor light	Dark, low, average, high
Outdoor light	Dark, low, average, high
Noise	Silent, low, average, high
Time	Morning, afternoon, evening, night, late night
Day type	Work day, week-end
Surrounding	Alone, surrounded
Temperature	Very low, low, average, high, very high
Incoming call	Yes, No
User's preference	To be fixed according to the user

compose the global context and in our case will be composed of the following elements with their possible values (Table 18.3).

Figure 18.5 shows the ontology-based model of context in our smart living room composed of the main concepts and relations between them.

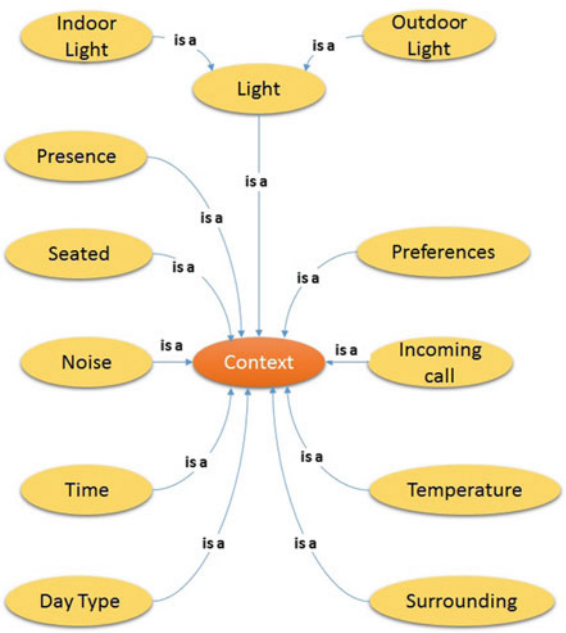


Fig. 18.5 Ontology-based context model of the living room

18.5 Context Reasoning

One major advantage of ontology-based context modeling resides in the possibility to carry out logical reasoning. Ontology-based context models support reasoning tasks in a better way than many other modeling approaches. According to The World Wide Web Consortium (W3C) [26], "Inference" means that automatic procedures can generate new relationships based on the data and based on some additional information in the form of a vocabulary, e.g., a set of rules. Context reasoning means to automatically find facts that are implicit in the ontology given explicitly stated facts. It consists of ‰finding what you know, but you don't know you know it - yet. Mainly it permits to deduce relevant new contextual information that cannot be explicitly provided by sensors of the smart space. Ontology reasoning permit to do three main tasks: (a) check context inconsistency and conflicts caused by imperfect sensing of contextual information, (b) check whether concepts are non-contradictory and (c) find subsumption relationships between classes and instances (for instance a user is sitting on the sofa of the living room, that means his inside the living room too, because the sofa is located in the living room). Many inference tasks can be reduced to subsumption reasoning which in turn can be reduced to concepts satisfiability. There are many automated tools for ontology reasoning. Pellet is one of the most common reasoning engines used for reasoning with Protégé OWL models. in our work we have used the pellet [27, 28] ontology reasoner to check the consistency of our ontology and infer new implicit context information.

Another goal of context reasoning consists of deducing and recognition of the current activity inside the living room based on the value of current context which could help a lot to provide accurate and adapted services to user (s). There are several known contexts where it is very easy to deduce the activity inside the living room. Among those contexts we can cite some examples as shown in Table 18.4 (the symbol? used in Table 18.4 means whatever value).

Table 18.4 Deducing activity from current context

Context elements	Context 1 values	Context 2 values	Context 3 values
Seated user	1	Several	0
User's presence	Yes	Yes	1
Indoor light	Very low	High	Average
Outdoor light	?	Dark	?
Noise	Silent	high	?
Time	Afternoon	night	?
Day type	Week-end	Week-end	?
Surrounding	Alone	Surrounded	Alone
Temperature	?	?	?
Incoming call	no	?	?
User's preference	?	?	?
Activity	The user is making a nap and equipment should be adjusted to a context of a nap	The user is with his family or friends talking and discussing equipment should be adjusted to a context of a home meeting	The user is inside the living room without intention to stay there (not sitting on sofa). equipment should be adjusted to a context of a simple entry/exit

18.6 Conclusion and Future Work

The first step of developing context-aware applications consists of context modeling based on a clear and concise definition of context. In this paper, we have introduced an ontology-based model for a smart living room. The modeling process is composed of two stages: (1) the living room modeling and (2) contextual information modeling based on our previous definition of context. The main advantages of our approach are: (1) context modeling is based on a simple and clear definition of context, (2) the proposed ontology is more generic and cover almost all concepts of context in a smart living room and (3) the modeling process is very simple and leads in an easy way to the elements of context. The use of ontology-based context modeling permit to make context reasoning for implicit knowledge inference in an efficient manner. Our future work will focus on achieving a context-aware services adaptation based on first order logic.

References

1. Cook D, Das S (2004) Smart environments: Technology, protocols and applications. John Wiley & Sons, p 424. ISBN: 0-471-54448-5
2. Miraoui M, El-etriby S, Tadj C, Abid AZ (2015) Ontology-based context modeling for a smart living room. In: Proceedings of The World Congress on Engineering and Computer

Science 2015, WCECS 2015. Lecture Notes in Engineering and Computer Science, 21–23 Oct 2015, San Francisco, USA, pp 127–132
3. Chen H, Perich F, Finin TW, Joshi A (2004) Soupa: standard ontology for ubiquitous and pervasive applications. In: MobiQuitous, IEEE Computer Society, pp 258–267
4. Pantsar-Syvaniemi S, Simula K, Ovaska E (2010) Context-awareness in smart spaces. In: IEEE Symposium on Computers and Communications (ISCC) Riccione
5. Chen H, Finin T, Joshi A (2003) An ontology for context-aware pervasive computing environments. Knowl Eng Rev, 18, 197–207. Springer Verlag
6. Gu T, Wang XH, Pung HK, Zhang DQ (2004) An ontology-based context model in intelligent environments. In: Proceedings of Communication Networks and Distributed Systems Modeling and Simulation Conference, San Diego, California, January 2004
7. Abdulrazak B, Chikhaoui B, Gouin-Vallerand C, Fraikin B (2010) A standard ontology for smart spaces. Int J Web Grid Serv 6(3):244–268
8. Mohsin Saleemi M, Rodríguez ND, Lilius J, Porres I (2011) A framework for context-aware applications for smart spaces. In: Balandin S, Koucheryavi Y, Hu H (eds) Proceedings of The 4th conference on Smart Spaces ruSMART 2011, Lecture Notes in Computer Science. Springer, pp 14–25
9. Li M (2011) Ontology-based context information modeling for smart space. In: 10th IEEE International Conference on Cognitive Informatics & Cognitive Computing (ICCI*CC), 2011, Banff, AB, pp 278–283
10. Lee H, Kwo J (2013) Ontology model-based situation and socially-aware health care service in a smart home environment. Int J Smart Home, 7(5):239–250
11. Wemlinger Z, Holder L (2011) The cose ontology: bringing the semantic web to smart environments. In: ICOST'11 Proceedings of the 9th international conference on Toward useful services for elderly and people with disabilities: smart homes and health telematics, pp 205–209
12. Wei M, Xu J, Yun J, Xu L (2012) Ontology-based home service model. Comput Sci Inf Syst 9 (2):813–838
13. Jih W, Hsu JY (2010) Agent-based context-aware service in a smart space. Agent-Based Ubiquitous Computing, Atlantis Ambient and Pervasive Intelligence, vol 1, pp 131–146
14. Zhu X, Yu Y, Ou Y, Luo D, Zhang C, Chen J (2013) System modeling of a smart-home healthy lifestyle assistant agents and data mining interaction. Lect Notes Comput Sci 7607:65–78
15. Li C, Sun L, Hu X (2012) A context-aware lighting control system for smart meeting rooms. Systems Engineering Procedia, vol 4, Information Engineering and Complexity Science–Part II, pp 314–323
16. Chahuara P, Portet F, Vacher M (2013) Making context aware decision from uncertain information in a smart home: a markov logic network approach. In: Fourth International Joint Conference on Ambient Intelligence, Dublin
17. Jih WR, Hsu JY (2009) Agent-based Ubiquitous Computing, vol 1 of Atlantis Ambient and Pervasive Intelligence, chapter Agent-Based Context-Aware Service in a Smart Space, Atlantis Press, June 2009, pp 131–146
18. Miraoui M, Tadj C, Amar CB (2008) Context modeling and context-aware service adaptation for pervasive computing systems. Int J Comput Inf Sci Eng (llCISE) 2(3):148–157
19. Strang T, Linnhoff-Popien C (2004) A context modeling survey In: The first International Workshop on Advanced context modeling, Reasoning and management, UbiComp 2004
20. Bettini C, Brdiczka O, Henricksen K, Indulska J, Nicklas D, Ranganathan A, Riboni D (2010) A survey of context modelling and reasoning techniques. Pervasive Mob Comput 6(2):161–180
21. Schilit S, Theimer M (1994) Disseminating active map information to mobile hosts. IEEE Netw 8(5):22–32
22. Brown PJ, Bovey JD, Chen X (1997) Context-aware applications: from the laboratory to the marketplace. IEEE Pers Commun 4(5):58–64

23. Ryan N, Pascoe J, Morse D (1997) Enhanced reality fieldwork: the context-aware archaeological assistant. Comput Appl Archeol
24. Dey AK (2001) Understanding and using context. J Pers Ubiquitous Comput 5:4–7
25. Miraoui M, Tadj C (2007) A service oriented definition of context for pervasive computing. In: Proceedings of the 16th International Conference on Computing, Mexico city, Mexico, Nov 2007
26. The world wide web Consortium (W3C), https://www.w3.org/standards/semanticweb/inference (visited June 2016)
27. Pellet reasoner, https://www.w3.org/2001/sw/wiki/Pellet (visited june 2016)
28. Parsia B, Sirin E (2004) Pellet: an OWL DL reasoner. Poster. In: Third International Semantic Web Conference (ISWC2004), Hiroshima, Japan, Nov 2004

Chapter 19
Performance Analysis of Heed Over Leach and Pegasis in Wireless Sensor Networks

Gaurav Kumar Nigam and Chetna Dabas

19.1 Introduction

Wireless sensor Networks (WSN) consists of a large number of low powered sensor nodes typically in the range of hundreds to thousands in number that are multi-functional and they are deployed in a hostile environment [1]. Sensor nodes are tiny in size that have finite battery power that is non replaceable. There are various applications of WSN like Military application, health application, home application etc. [2]. There are various challenges and design issues in WSN like node deployment, routing, energy consumption, fault tolerance, coverage, connectivity and QoS i.e. quality of service. We may find Routing as a major challenge in WSN since there are huge amount of sensor nodes involved, therefore it is needed to develop a global addressing proposal for the distribution of sensor nodes. Energy consumption is the main focus in WSN as the battery power is limited and it is not suited to recharge the battery as the node are randomly deployed in an area as needed. So, various energy efficient routing techniques are used for energy conservation in WSN.

Energy conservation is a important phenomenon in Wireless sensor Networks. Since they are randomly distributed in a hostile environment it is difficult to monitor the battery power at regular intervals. If a node dies earlier then it may cause instability in network [2, 3]. So, to achieve the goal all nodes should be in working condition. One barrier we found in this is the unbalanced energy rate. There are already various techniques to improve the consumption of energy rate like clustering, data aggregation and routing. For a large scale WSNs, we may find multiple

G.K. Nigam (✉)
Jaypee Institute of Information Technology, Sector-128, Near Jaypee Wish Town, Noida, India
e-mail: gauravnigam786@yahoo.com

C. Dabas
Jaypee Institute of Information Technology, Sector-62, Noida, India
e-mail: chetna.dabas@jiit.ac.in

S.-I. Ao et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-981-10-2717-8_19

base stations as one of the possible solution. This is performed by shortening the distance or path between the base station and the sensor node and thereby saving the consumption of energy for operation of transmissions. To accomplish the efficiency, group of multiple base stations is to be placed in the sensing area. However, energy saving is not a effective measure to increase the network lifetime because there may be depletion of uneven energy in partitioning of network and the coverage ratio, if it is low then it lowers the network performance. Designing efficient energy consumption protocol is the main focus in WSN as the we are using battery power, that is limited, and it is not suited to recharge the battery as the node are randomly deployed in an area as needed [4]. So, various energy efficient routing techniques are used for energy conservation in WSN [5, 6].

The author in the paper [7] explains the various routing protocols that is based on layers, architecture and functionality of WSN. Further a study of 15 routing protocols is given along with their comparison.

WSN have tiny micro sensor nodes that are deployed in an environment that have a fixed energy constraint. There are various energy efficient protocols that are needed to increase the network lifetime. In [8] the author has designed, implemented and evaluated the performance of a reliable energy aware routing protocol (REAR) and thereby compared it with existing routing protocols. REAR uses DATA-ACK service to achieve successful transmission of data. The REAR architecture consists of total three threads; they are: Transmission, Receiver and Processing thread. The data sensed by the sensor node in the real field is transmitted by the transmission thread. The routing message from the source sensor node that is delivered from MAC layer is received by the receiver thread. Processing thread takes out from the queue the data information and process it.

REAR set up the route by looking at the current energy level information of nodes.

In [9] the author has discussed the mobile wireless sensor networks that are deployed in the human body to monitor temperature of body, heart beat or any other activity [10]. There are few mobile nodes in WBAN (Wireless Body Area Sensor Networks) that reinforces scalability and fault tolerance. The author proposed an efficient routing protocol by using GPS (global positioning system) and energy level information of the sensor nodes. The advantage of this protocol is that it uses limited routing space by selecting the best node in the path using unicast transmission and thereby reducing the overall network overhead and energy consumption. There are two category of routing that is used by the author. They are:

(1) Location-aided routing (LAR)
(2) Geographic and energy aware routing (GEAR)

LAR is also called as on-demand routing and it discovers the route by using limited flooding.

GEAR is a proactive routing and uses a heuristics approach to route a packet to the destination. Each node gathers the cost information from reaching to its

neighbor and each node makes a routing table and broadcasts its location information and residual energy available for further processing.

Routing protocols can be categorized into multiple sections [1, 11]. One of the category among multiple one is hierarchical routing that is based on the assumption that energy efficiency is maintained by all sensor nodes. The nodes with higher energy observe and transmit the information in the network whereas nodes with lesser energy simply sense in their area of interest [10]. The formation of group of sensors, called a cluster, and the assigning of tasks makes a major contribution to scalability, lifetime and energy efficiency [12, 13]. Low-energy adaptive clustering hierarchy (LEACH), Power efficient gathering in sensor information system (PEGASIS), Hybrid energy efficient distributed (HEED) are hierarchal routing protocols.

LEACH called as low energy adaptive clustering hierarchy is a routing protocol for wireless sensor networks in which there is a immovable base station and all sensor nodes are equal in terms of computation power and resources. LEACH conserves energy via adaptive clustering and aggregation. There are various rounds in LEACH and in each round there are two phases: set up phase (Advertisement and cluster setup) and steady phase (schedule creation and data transmission). PEGASIS is a version of LEACH where a single node is chosen as a leader and whose task in each round is to send fused data to the base station. PEGASIS is a based on chain communication. In PEGASIS, each node communicates to its close neighbour and collecting data from all the neighbours and thereby transmitting the final collated data to the base station. Collected data moves from node to node, gets combined and eventually a pre selected node transmits it to the sink. It involves chain construction, gathering data and then transmitting this data.

HEED is designed to select different cluster heads in a network, considering remaining amount of energy of the node. It consists of a 3-tier process viz, Choosing a cluster head, selection of cluster head and finally decision on node's status is taken.

19.2 Problem Formulation

One of the most critical aspects of WSN is the energy management and the key to energy management and deploying an appropriate sensor network is to know the requirements of the system. We try to implement a routing protocol best suited to the system using our knowledge of the routing protocol's performance under various parameters.

Factors like speed of the node, pause-time, size and area of the network, number of traffic sources, and type of routing affect the routing performance of networks [14].

Thus, we select most critical parameters for WSN to be packet delivery ratio, Throughput, energy level and delay.

19.3 Metric Considered

MATLAB offers support for different parameters for the measurement of the performance of the WSN under different routing protocols. The parameters used here for performance evaluation are Packet delivery ratio, End to end, Energy consumed and Throughput. The metrics used to evaluate the performance can be defined as:

(i) Average end to end delay (expressed in seconds): It is defined as the time taken by a packet to traverse across a network from source to destination.
(ii) Energy consumed (measured in kJ): It is defined as the total rate at which energy is consumed by sensor nodes in a WSN within a certain specific time period.
(iii) Packet delivery ratio (expressed in number of packets): It is defined as the total proportion of successfully delivered packets from source to the sink that is sent by all sensor nodes in the network.
(iv) Average throughput (measured in bits per second): It is defined as the average number of packets successfully received per unit time by the sink node.

19.4 Methodology

For executing LEACH, PEGASIS and HEED protocol, MATLAB simulator is used. Plotting of performance is done for all the routing protocols to show that HEED protocol is the improved version of LEACH and PEGASIS as it provides smooth and less degradation in performance. It uses the residual energy for finding the probability of node to become CH. In the simulation, up to 100 nodes are deployed in a 100 m × 100 m region that is divided into 10 × 10 grids.

A scenario is used to simulate our three routing protocols. The performance evaluation is done by simulating in the MATLAB simulator and results in the form of graphs are obtained (Table 19.1).

The values for this simulation scenario are given below:

Table 19.1 Parameters of simulation

Parameters	Values
Network size	100 m × 100 m
Energy consumed	10 pJ/bit/m^2
Message size	3000 bits
Number of nodes	100
Initial energy, E_0	0.4 J
Aggregation energy, E_{DA}	5nJ/bit

19.5 Simulation Results

The simulation is performed in MATLAB simulation tool. The area of simulation is 100 m × 100 m.

In the simulation we have used various parameters like throughput, (PDR) Packet delivery ratio, energy level and delay. The three protocols (HEED, LEACH and PEGASIS) are compared using these parameters. The outputs of simulation of various routing protocols are discussed below.

(A) Throughput: The figure below display the x-graph for throughput investigation. In the graph below, it is seen that the throughput in HEED protocol is productive when it is correlated with LEACH and PEGASIS (Fig. 19.1).

(B) Packet delivery ratio: The figure below display the x-graph for packet delivery investigation. In the below graph, HEED has the higher packet delivery ratio compared with LEACH and PEGASIS (Fig. 19.2).

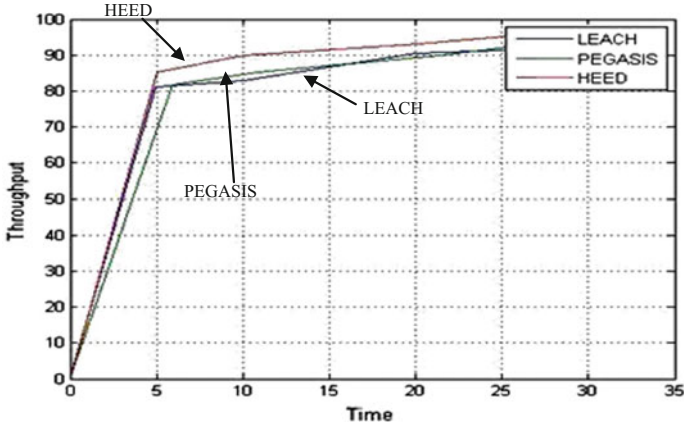


Fig. 19.1 Throughput versus time

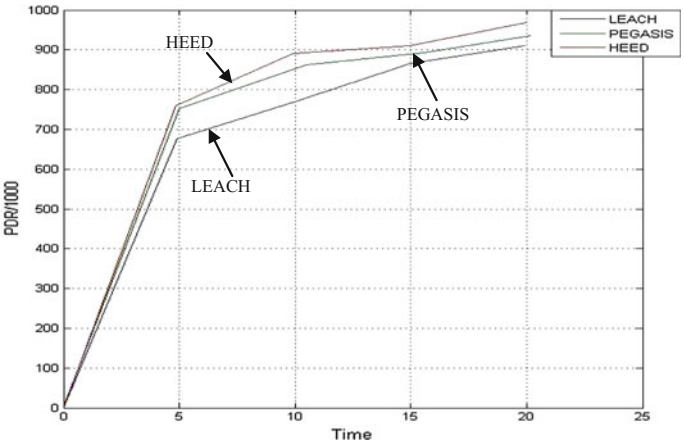


Fig. 19.2 PDR versus time

(C) Energy level: The figure below display the x-graph for energy investigation. In the below graph, HEED has the higher energy level than the other two protocols (Fig. 19.3).
(D) Delay: The figure below display the x-graph for delay investigation. In the below graph, we can find that there is minimum delay in HEED protocol (Fig. 19.4).

The results are represented in the form of the Table 19.2. The routing protocols are numbered as:

1. LEACH
2. PEGASIS
3. HEED

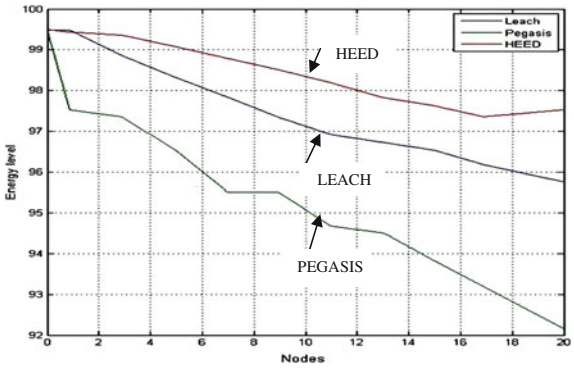


Fig. 19.3 Energy level versus number of nodes

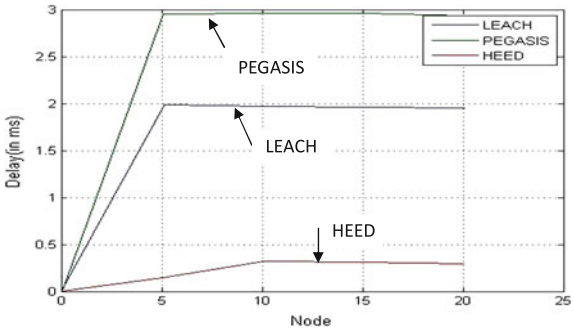


Fig. 19.4 Delay versus nodes

Table 19.2 Results after simulation of protocols

Protocol	Delay (sec)	Energy level (%)	Throughput (%)	PDR (%)
1	1.9	93	82	88
2	3.0	90	93	90
3	0.4	96	96	94

19.6 Conclusion

This chapter concludes the result in the scenario to observe the variation in the working of the three routing protocols LEACH, PEGASIS and HEED to analyze their behaviour w.r.t. the metrics mentioned above. The simulation results of the three routing protocols shows that HEED protocol is the best protocol and it shows almost the best performance for all the metrics. The chapter gives comparison of the three hierarchical routing protocols on the basis of some metrices. The comparision is done on the basis of simulation performed.

References

1. Liu X (2012) A survey on clustering routing protocols in wireless sensor networks. Sensors,12(8): 11113–11153
2. Akyildiz IF, Su W, Cayirci E, Sankarasubramanium Y (2002)Wireless sensor networks: a survey. Comput Netw J, 38(4):393–422
3. Rajagopalan R, Varshney PK (2006) Data-aggregation techniques in sensor networks: a survey. IEEE Commun Surv Tutor 8(4):48–63
4. Nigam,GK (2010) Technical challenges and constraints for spectrum planning in GSM networks. Int J Comput Sci Eng
5. Wei Chunjuan, Yang Junjie, Gao Yanjie, Zhang Z (2011) Cluster-based routing protocols in wireless sensor networks: A survey. Int Conf Comput Sci Netw Technol 3:1659–1663
6. Younis O, Fahmy S (2004) HEED: a hybrid, energy-efficient, distributed clustering approach for ad hoc sensor networks. IEEE Trans Mob Comput 3(4):366–379
7. Nigam GK, Dabas C (2015) A survey on protocols and routing algorithms for wireless sensor networks. In: Proceedings of The World Congress on Engineering and Computer Science 2015, WCECS 2015, Lecture Notes in Engineering and Computer Science, San Francisco, USA, 21–23 Oct 2015, pp 719–722
8. Shin K-Y, Song J, Kim JW, Yu M, Mah PS (2007)The 9th International Conference on Advanced Communication Technology, vol 1, pp 525–530
9. Kim K, Kim J, Lee I-S, Lee H, Yoon M, Han K (2009) First International Conference on Networks and Communications, 2009, NETCOM'09, pp 396–399
10. Shen X, Chen J, Wang Z, Sun Y (2006) Grid Scan: a simple and efficient approach for coverage issue in Wireless Sensor Networks. IEEE Int Commun Conf, 8:3480–3484
11. Sohrabi K, Gao J, Ailawadhi V, Pottie GJ (2002) Protocols for self-organization of a wireless sensor network. IEEE Pers Commun 7:16–27
12. Al- Karaki JN, Kamal AE (2004) Routing techniques in wireless sensor networks: a survey. IEEE Wirel Commun 11(6): 6–28

13. Seetharam A,Dasgupta S, Das D (2007) Lifetime enhancement in wireless sensor networks through selective data handover. In: Third International conference on wireless communication and sensor networks, December 2007, pp 55–60
14. Perkins DD, Hughes HD, Owen CB (2002) Factors affecting the performance of ad hoc networks. IEEE Int Conf Commun, 4:2048–2052

Chapter 20
Spectral Properties and Error Rate Performance of Digital Chirp Communication System

Mohammad Alsharef, Abdulbaset M. Hamed and Raveendra K. Rao

20.1 Introduction

Chirp modulation, also known as a linear frequency modulation, can be viewed as a type of spread spectrum signaling technique in which a carrier is swept over a wide-band during a given data pulse interval. This spreading technique produces a signal whose bandwidth is much wider than the information bandwidth. Because chirp signals possess properties that are useful in wireless communication, they are expected to be utilized in a variety of future communication systems. In particular, the growing interest in chirp signals is due to the advances in Surface Wave Acoustic (SAW) technology, which offers a rapid close-to-optimum method for both generation and correlation of wide-band chirp pulses [1]. In 2007, IEEE introduced Chirp Spread Spectrum (CSS) physical layer in its standard 802.15.4a [2]. Furthermore, single-chip transceivers for wireless communication in the industrial, scientific, and medical (ISM) band have been developed and are commercially available [3]. In [4], Alsharef and Rao have evaluated the performance of M-ary chirp signals over several fading channels. In [5], they have proposed M-ary signaling technique for chirp modulation and have evaluated its performance over Additive White Gaussian Noise (AWGN) channel for coherent and non-coherent detection. Gupta, Mumtaz, Zaman and Papandreou-Suppappola [6] have evaluated the performance of chirp modulation in a frequency-hopped CDMA over several channel models. The error performance of chirp modulation over frequency selective and non-selective fading channels are

M. Alsharef (✉)
Taif University, Taif, Saudi Arabia
e-mail: m.alsharef@tu.edu.sa

A.M. Hamed · R.K. Rao
University of Western Ontario, London, Canada
e-mail: ahamed6@uwo.ca

R.K. Rao
e-mail: rrao@uwo.ca

S.-I. Ao et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-981-10-2717-8_20

investigated in [7]. The performance analysis with closed-form bit error probability expressions for chirp modulation in the maximum ratio combining (MRC) diversity system has been investigated in [8]. Hengstler, Kasilingam and Costa [9] have proposed a novel chirp modulation spread spectrum technique that utilizes antipodal signaling and have derived an expression for its bit error rate and bandwidth efficiency. In [10], Kadri and Rao have evaluated the performance of weak binary chirp signals in ε-mixture noise for coherent and non-coherent detection. Phichet, Tran and Tawil have proposed a new system called multi linear chirp frequency hopping code division multiple access (MLC-FH-CDMA) and studied its performance over Nakagami-Rice fading channel [11]. In [12], the authors applied the concept of frequency hopping (FH) spread spectrum technique to the multi-user chirp modulation system and evaluated the performance over multi-path fading channels. In the literature, there are several studies on chirp modulation; however, the impact of the shadowing on chirp signals has not received much attention. M-ary chirp signalling provides increased immunity to channel noise. In this paper, chirp modulation for transmission of M-ary data is considered and then expressions for average error probability in closed form are derived for several fading channels. In particular, Rayleigh, Nakagami-m and Generalized-K fading channels are considered in this work. The effect of fading and shadowing on the performance of the receiver are examined for different values of fading parameters c and m. Numerical results are presented to illustrate the effect of channel on the performance of these chirp signals.

The paper is organized as follows. In Sect. 20.2, M-ary chirp modulation is described and illustrated. Also, optimum coherent receiver and its performance in AWGN channel for M-ary chirp signals are treated. In Sect. 20.3, symbol error probabilities over Rayleigh, Nakagami-m and Generalized-K fading channels are derived. Section 20.4 is devoted for the spectra of M-ary chirp signals. The paper is concluded in Sect. 20.5.

20.2 M-ary Chirp System Model

20.2.1 M-ary Chirp Signalling Technique

The ith modulated M-ary chirp signal is given by [13]:

$$S_i(t) = \sqrt{\frac{2E_s}{T_s}}\cos(w_c t + \phi(t, d_i) + \theta), \quad 0 \le t \le T_s \tag{20.1}$$

where E_s is the symbol energy in the symbol duration T_s seconds, w_c is the angular carrier frequency, $\phi(t, d_i)$ is the information carrying phase, θ is the starting phase ($t = 0$) and is assumed to be zero with no loss of generality. The information carrying phase $\phi(t, d_i)$ for chirp modulation is given by:

$$\phi(t, d_i) = \begin{cases} 0, & t \le 0, \quad t > T_s \\ d_i\pi \left\{ h\left(\frac{t}{T_s}\right) - w\left(\frac{t}{T_s}\right)^2 \right\}, & 0 \le t \le T_s \\ d_i\pi q = d_i\pi(h - w), & t = T_s \end{cases} \tag{20.2}$$

where $d_i, i = 1, 2, ..., M$, takes values from the set $\{\pm 1, \pm 3, \ldots, \pm(M-1)\}$ and:

$$d_i = \begin{cases} +i, & if\ i\ odd \\ -(i-1), & if\ i\ even \end{cases} \tag{20.3}$$

for example, for 4-ary chirp modulation, there are 4 possible waveforms $S_1(t), S_2(t)$, $S_3(t), and\ S_4(t)$ corresponding to data $+1, -1, +3, -3$, respectively. In (20.2), h and w are dimensionless parameters: h represents the initial peak-to-peak frequency deviation divided by the symbol rate $\frac{1}{T_s}$, and w represents the frequency sweep width divided by the symbol rate $\frac{1}{T_s}$. Since $h = (q + w)$, we choose (w, q) to be independent signal modulation parameters. Figure 20.1 shows an example of 4-ary chirp modulation (quadrature chirp signals) where the data d_i takes values $\pm 1, \pm 3$.

20.2.2 Optimum Correlator-Receiver

Figure 20.2 shows the optimum coherent receiver for detection of M-ary chirp signals in AWGN. This receiver will compute M functions $\Lambda_1, \Lambda_2, \ldots, \Lambda_M$ given by:

$$\Lambda_i = \int_0^{T_s} r(t)\, S_i(t)\, dt, \quad i = 1, 2, ..., M \tag{20.4}$$

and arrives at an optimum decision based on the largest of these M values. Thus, if:

$$\Lambda_k = \max\left\{\Lambda_1, \Lambda_2, \ldots, \Lambda_M\right\} \tag{20.5}$$

then the receiver decides the transmitted data as:

$$d = \begin{cases} +k, & k\ odd \\ -(k-1), & k\ even \end{cases} \tag{20.6}$$

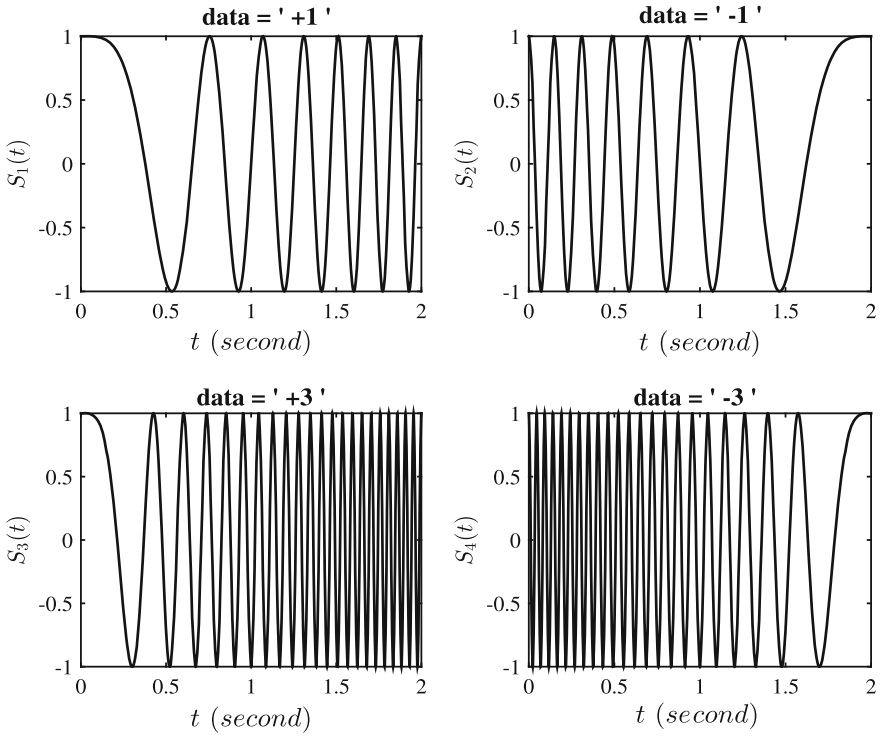


Fig. 20.1 4-ary chirp modulated signals as a function of time

20.2.3 Symbol Error Rate Performance

Suppose that the transmitted signal is $S_i(t)$. The received signal is given by:

$$r(t) = S_i(t) + n(t), \quad 0 \leq t \leq T_s \tag{20.7}$$

where $n(t)$ is the additive white Gaussian Noise (AWGN) with zero-mean and one-sided power spectral density (PSD) of N_0 *watts*/*Hz*. The conditional probability of an error given $S_i(t)$ is transmitted is given by:

$$P(\varepsilon|S_i) = Prob[\Lambda_1 > \Lambda_i \cup \Lambda_2 > \Lambda_i \cup ... \cup \Lambda_M > \Lambda_i|S_i] \tag{20.8}$$

The conditional probability in (20.8) can be union bounded using the identity:

$$P(\Lambda_1 \cup \Lambda_2 \cup .... \cup \Lambda_n) \leqslant \sum_{j=1}^{n} P(\Lambda_j) \tag{20.9}$$

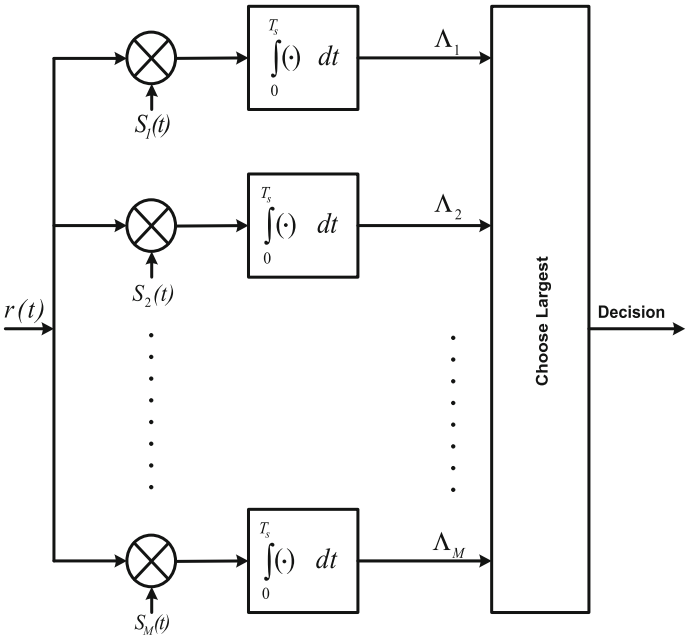


Fig. 20.2 Optimum coherent receiver for M-ary chirp modulation

Thus, (20.8) can be written as:

$$P(\varepsilon|S_i) = \sum_{\substack{j=1 \\ j\neq i}}^{M} Prob[\Lambda_j > \Lambda_i|S_i] \tag{20.10}$$

Averaging over all possible S_i, the expression for average symbol error probability is given by:

$$P(\varepsilon) \leqslant \sum_{j=1}^{M} \sum_{\substack{i=1 \\ i\neq j}}^{M} P(S_j) P_r \left[\Lambda_j \geqslant \Lambda_i \mid S_i\right] \tag{20.11}$$

It is noted that in (20.11), Λ_i's are Gaussian random variables. Thus, (20.11) can be written as:

$$P(\varepsilon) \leqslant \frac{1}{M} \sum_{j=1}^{M} \sum_{\substack{i=1 \\ i\neq j}}^{M} Q\left[\sqrt{\frac{E_s}{N_0}(1-\rho(i,j))}\right] \tag{20.12}$$

$$Q(x) \triangleq \frac{1}{\sqrt{2\pi}} \int_{x}^{\infty} e^{-\frac{t^2}{2}} dt$$

which also can be written as:

$$Q(x) = \frac{1}{\pi} \int_{0}^{\frac{\pi}{2}} e^{\left(-\frac{x^2}{2sin^2(\theta)}\right)} \; d\theta \tag{20.13}$$

and the quantity $\rho(i,j)$ is the normalized correlation between $S_i(t)$ and $S_j(t)$ and is defined as [5]:

$$\rho(i,j) = \frac{1}{E_s} \int_{0}^{T_s} S_i(t)\, S_j(t)\, dt, \quad i,j = 1, \ldots, M \tag{20.14}$$

using (20.1) in (20.14), the normalized correlation can be written as:

$$\rho(i,j) = \left[\frac{\cos(\Omega)}{\sqrt{2\zeta w}}\, \mathbb{C} + \frac{\sin(\Omega)}{\sqrt{2\zeta w}}\, \mathbb{S} \right] \tag{20.15}$$

where

$$\mathbb{C} = \mathbf{C}(u_h) - \mathbf{C}(u_l), \;\; \mathbb{S} = \mathbf{S}(u_h) - \mathbf{S}(u_l)$$

$$\Omega = \frac{\pi \zeta h^2}{4w}, \; \zeta = |d_i - d_j|$$

$$u_h = \sqrt{\frac{\zeta}{2}} \frac{(w-q)}{\sqrt{w}}, \;\; u_l = \sqrt{\frac{\zeta}{2}} \frac{(w+q)}{\sqrt{w}}$$

the value d_i is the data associated with the signal $S_i(t)$ and d_j is the data associated with the signal $S_j(t)$. The function $\mathbf{C}(.)$ and $\mathbf{S}(.)$ are the Fresnel cosine and sine integral which are given by:

$$\mathbf{C}(u) = \int_{0}^{u} \cos(\frac{\pi x^2}{2}) \; dx, \quad \mathbf{S}(u) = \int_{0}^{u} \sin(\frac{\pi x^2}{2}) \; dx$$

Table 20.1 Optimum values of q and w for 2-, 4-, and 8- chirp signals

Modulation size (M)	(q, w)
2	(0.36, 1.52)
4	(0.4, 2.4)
8	(0.95, 0.25)

The performance of the optimum coherent receiver can be evaluated using Eqs. (20.12)–(20.15). The optimum modulation parameters (q, w) are those that minimize the symbol error probability given by (20.12) [5]. These are given in Table 20.1. For all illustrations in the paper these optimum modulations parameters are used.

20.3 Performance of Optimum Receiver over Fading Channels

In this section, we consider evaluating the performance of optimum receiver over short-term and long-term fading (shadowing). In particular, closed form expressions for average symbol probability of error for Rayleigh, Nakagami-m and Generalized-K channel models are derived. The received chirp signal over the fading channel can be written as:

$$r(t) = h(t) * S_i(t) + n(t)$$

where $h(t) = \alpha\delta(t)$ is the impulse response of the channel and $n(t)$ is AWGN. Hence, the instantaneous SNR per symbol and the average SNR are described as $\gamma = \alpha^2 E_s/N_0$ and $\bar{\gamma} = \Omega E_s/N_0$, where $\Omega = E\{\alpha^2\}$. The average symbol error probability of M-ary chirp system (P_s) over fading channel is determined by averaging the conditional error probability over the Probability Density Function (PDF) of the fading model and is given by [14]:

$$P_{av} = \int_0^\infty p_\gamma(\gamma) P_e(\gamma)\ d\gamma \tag{20.16}$$

20.3.1 Rayleigh Fading Channel

Using (20.16), the average error probability P_{av} is evaluated for Rayleigh fading channel using:

$$p_\gamma(\gamma) = \frac{1}{\bar{\gamma}} \exp\left(-\frac{\gamma}{\bar{\gamma}}\right), \quad \gamma \geqslant 0 \tag{20.17}$$

Using (20.12), (20.13), (20.16) and changing the order of integration, the P_{av} can be expressed as:

$$P_{av} = \frac{1}{\pi \overline{\gamma} M} \sum_{j=1}^{M} \sum_{\substack{i=1 \\ i \neq j}}^{M} \int_0^{\frac{\pi}{2}} \int_0^{\infty} \exp\left(-\frac{(1-\rho(i,j))\gamma}{2\sin^2(\theta)} - \frac{\gamma}{\overline{\gamma}} \right) d\gamma \, d\theta \tag{20.18}$$

integrating (20.18) by using [15], the symbol error rate can be written as:

$$P_s = \frac{1}{M} \sum_{j=1}^{M} \sum_{\substack{i=1 \\ i \neq j}}^{M} \left[1 - \sqrt{\frac{1}{1 + \frac{1}{\frac{\Omega E_s}{2N_0}(1-\rho(i,j))}}} \right] \tag{20.19}$$

Figure 20.3 shows the bit error probability for 2-ary chirp over AWGN channel, BPSK, 2-FSK, and DPSK versus the normalized SNR over Rayleigh fading channel. It is noted that the performance of binary chirp is very close to the performance of BPSK modulation over Rayleigh fading channel and better than the performance of 2-FSK and DPSK systems. In Figs. 20.4 and 20.5, the bit error probability performances of the 2-ary chirp system are illustrated as a function of q and w, respectively. It is observed that as the values of the modulation parameters deviate from the optimum values, the performance degrades.

Figures 20.6 and 20.7 shows the symbol error probability performance of 4-ary and 8-ary chirp systems, respectively. For both systems, we have found that the

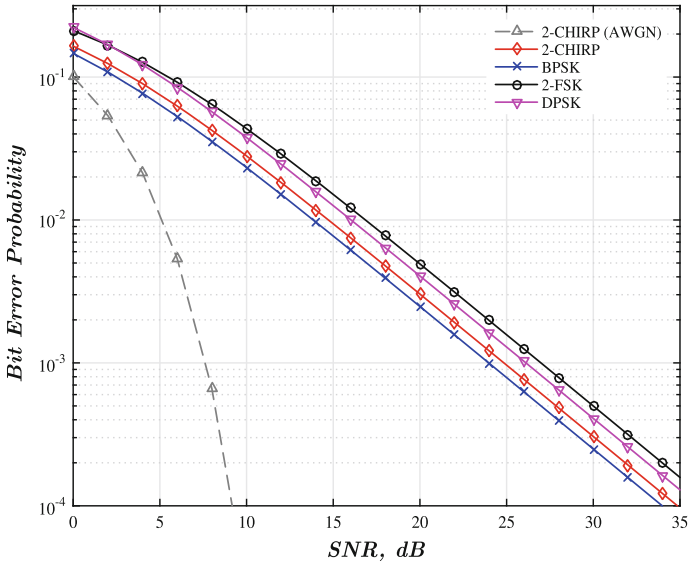


Fig. 20.3 Performance of 2-ary chirp system over Rayleigh fading channel

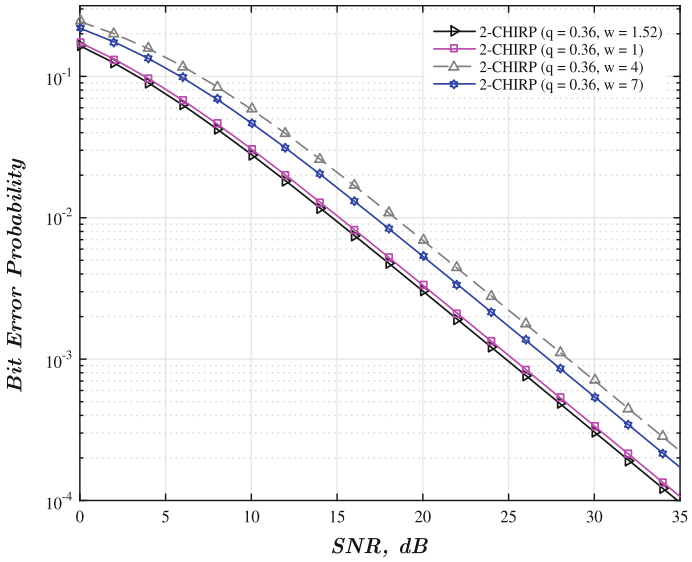


Fig. 20.4 Performance of 2-ary chirp system over Rayleigh fading channel as a function of $w = 1.52, 1, 4, 7$, for a fixed value of q

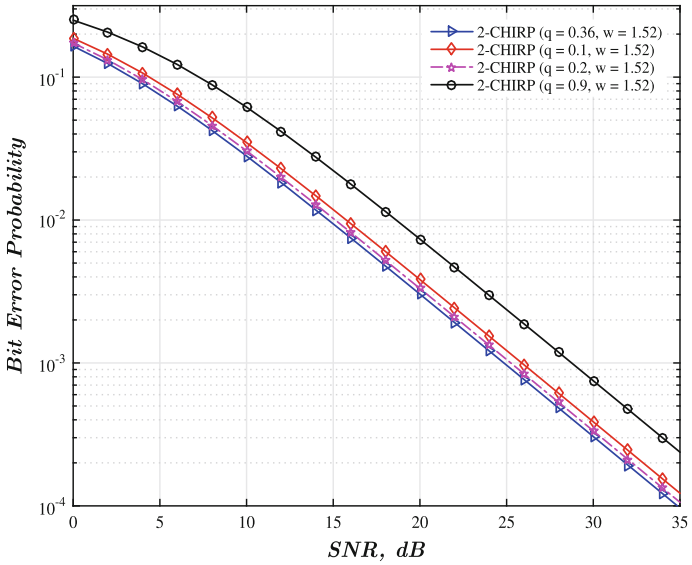


Fig. 20.5 Performance of 2-ary chirp system over Rayleigh fading channel as a function of $q = 0.36, 0.1, 0.2, 0.9$, for a fixed value of w

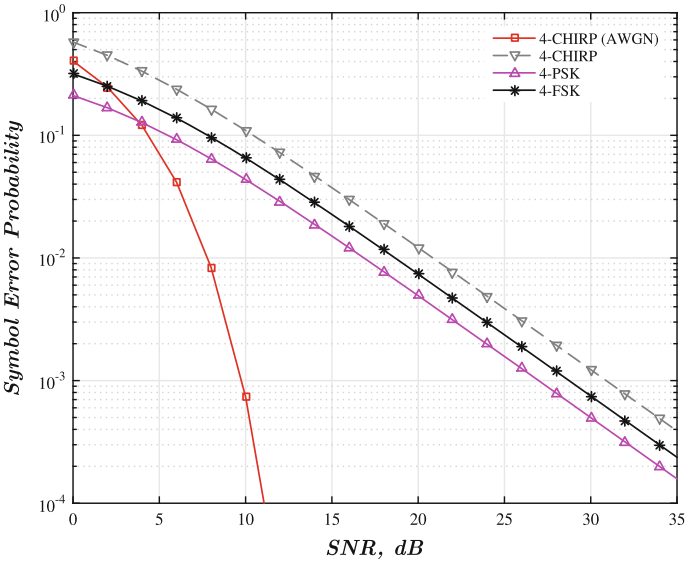


Fig. 20.6 Performance of 4-ary chirp system over Rayleigh fading channel

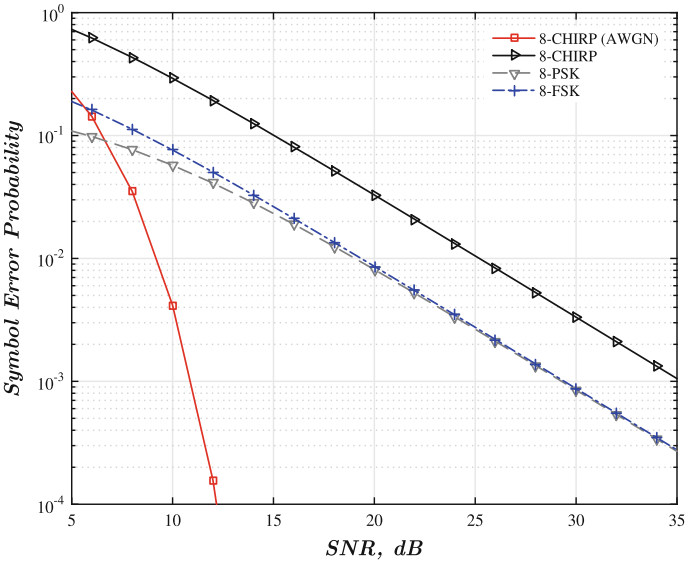


Fig. 20.7 Performance of 8-ary chirp system over Rayleigh fading channel

performance of the chirp system is worse than than the performance of 4-PSK and 4-FSK over Rayleigh fading channel.

20.3.2 Nakagami-m Fading Channel

The PDF of Nakagami-m model is given by [14]:

$$p_\gamma(\gamma) = \left(\frac{m}{\overline{\gamma}}\right)^m \frac{\gamma^{m-1}}{\Gamma(m)} e^{-m\frac{\gamma}{\overline{\gamma}}}, \quad \gamma \geqslant 0 \tag{20.20}$$

m is the fading parameter which can be used to generate several fading channel model ($m = 1$ represents Rayleigh model when there is no line of sight (LOS) component). By expressing the exponential term in (20.20) as $e^{-m\frac{\gamma}{\overline{\gamma}}} = G_{0,1}^{1,0}\left(\frac{m\gamma}{\overline{\gamma}} \mid \begin{matrix} \\ 0 \end{matrix}\right)$ and $Q(\sqrt{(1-\rho(i,j))\gamma}) = \frac{1}{\sqrt{\pi}} G_{1,2}^{2,0}\left((1-\rho(i,j))\gamma \mid \begin{matrix} 1 \\ 0,\ 1/2 \end{matrix}\right)$ in (20.12) using [(8.4.3/1), (8.4.14/2)] [16] respectively, where $G_{a,b}^{q,p}\left(x \mid \begin{matrix} \cdots \\ \cdots \end{matrix}\right)$ is the Meijer G-function [17]. (20.16) yields to the integral:

$$P_{av} = \frac{1}{M\sqrt{\pi}} \sum_{j=1}^{M} \sum_{\substack{i=1 \\ i\neq j}}^{M} \int_0^\infty \gamma^{m-1} G_{1,2}^{2,0}\left((1-\rho(i,j))\gamma \mid \begin{matrix} 1 \\ 0,\ 1/2 \end{matrix}\right) G_{0,1}^{1,0}\left(\frac{m\gamma}{\overline{\gamma}} \mid \begin{matrix} \\ 0 \end{matrix}\right) d\gamma \tag{20.21}$$

By solving the integral in (20.21), a closed form expression for the symbol error rate over Nakagami-m channel can be written as:

$$P_s = \frac{1}{M\sqrt{\pi}\Gamma(m)} \sum_{j=1}^{M} \sum_{\substack{i=1 \\ i\neq j}}^{M} G_{2,2}^{2,1}\left(\frac{(1-\rho(i,j))}{m} \frac{\Omega E_s}{2N_0} \mid \begin{matrix} 1,\ 1-m \\ 0,\ 1/2 \end{matrix}\right) \tag{20.22}$$

Figures 20.8, 20.9 and 20.10 show the performance of 2, 4, and 8-ary chirp modulation over Nakagami-m fading channel. For the special case when $m = 1$, Nakagami-m reduces to the well known Rayleigh model. The performance approaches AWGN performance as $m \to \infty$ because the line of sight (LOS) path dominates the received signal. Also, it is noted that as M increases the symbol error rate increases for a fixed value of SNR. This observation can be utilized to design an adaptive modulation technique by changing the modulation order to increase the bandwidth efficiency or to improve the error rate performance to meet a certain quality of service [18].

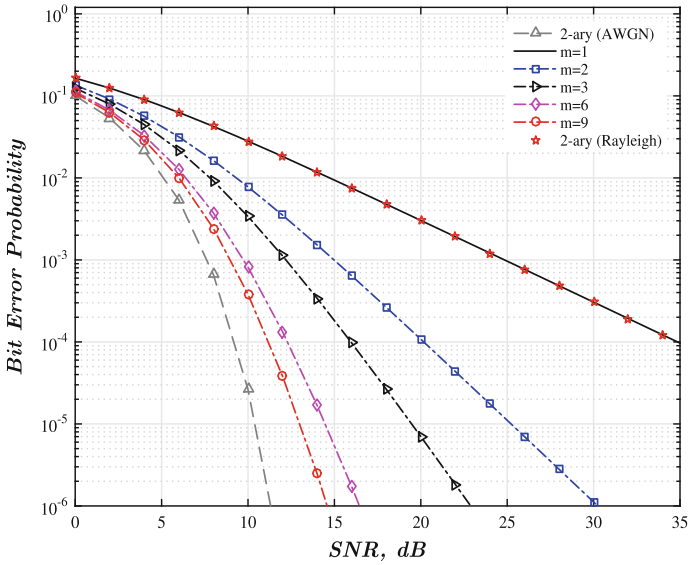


Fig. 20.8 Performance of 2-ary chirp system over Nakagami-*m* fading channel as a function of *m*

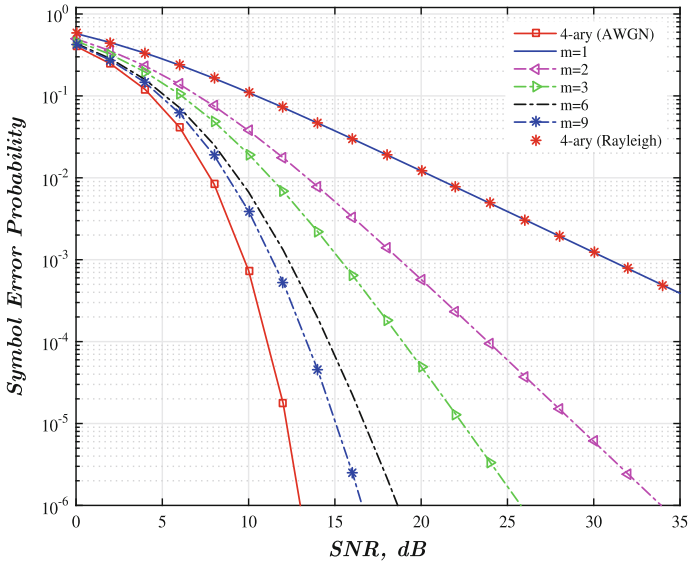


Fig. 20.9 Performance of 4-ary chirp system over Nakagami-*m* fading channel as a function of *m*

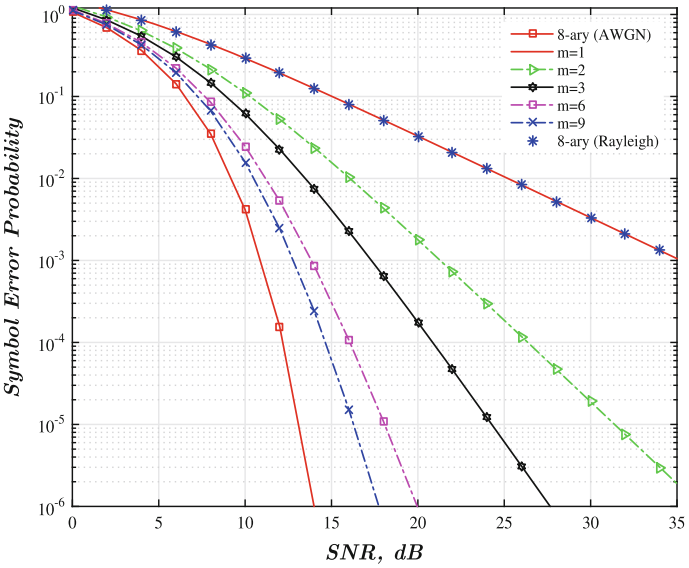


Fig. 20.10 Performance of 8-ary chirp system over Nakagami-m fading channel as a function of m

20.3.3 Generalized-K Fading and Shadowing Channel

The Generalized-K is a composite model that is used to describe the fading and shadowing channel characteristics. This model represents a wireless channel that subjects to short and long-term fading. The PDF of the Generalized-K model is given by:

$$p_\gamma(\gamma) = 2\left(\frac{cm}{\overline{\gamma}}\right)^{\frac{c+m}{2}} \frac{\gamma^{\frac{c+m-2}{2}}}{\Gamma(c)\Gamma(m)} K_{c-m}\left(2\sqrt{\frac{cm}{\overline{\gamma}}\gamma}\right), \gamma \geqslant 0 \tag{20.23}$$

where $K_{c-m}(.)$ is the modified Bessel function of order $c-m$, $\Gamma(.)$ is the Gamma function [17]. The coefficients c and m are the shadowing and fading's parameters; respectively. As m and c increase, the fading and shadowing become less sever. For m and $c \to \infty$, the channel approaches that of AWGN. using (20.12), (20.13), (20.16), (20.23) and expressing the modified Bessel function in (20.23) as $K_{c-m}\left(2\sqrt{\frac{cm}{\overline{\gamma}}\gamma}\right) = \frac{1}{2}G_{0,2}^{2,0}\left(\frac{cm}{\overline{\gamma}}\gamma \,\middle|\, \frac{(c-m)}{2}, \frac{-(c-m)}{2}\right)$ and $Q(\sqrt{(1-\rho(i,j))\gamma}) = \frac{1}{\sqrt{\pi}}G_{1,2}^{2,0}\left((1-\rho(i,j))\gamma \,\middle|\, \begin{matrix}1\\0,\,1/2\end{matrix}\right)$ using [(8.4.23/1), (8.4.14/2)] [16]. Thus, the integral of (20.31) can be expressed in terms of Miejer G-function as:

$$I = \frac{1}{2\sqrt{\pi}} \int_0^{\infty} \gamma^{\frac{c+m}{2}-1} G_{1,2}^{2,0}\left((1-\rho(i,j))\gamma \,\middle|\, \begin{matrix} 1 \\ 0,\ 1/2 \end{matrix} \right) G_{0,2}^{2,0}\left(\frac{c\,m}{\bar{\gamma}}\gamma \,\middle|\, \frac{(c-m)}{2}, \frac{-(c-m)}{2} \right) d\gamma \quad (20.24)$$

average symbol error probability ((P_{av})) of chirp modulation over K_G-channel is obtained in a closed form as:

$$P_s \geqslant \frac{1}{M\sqrt{\pi}\Gamma(c)\Gamma(m)} \cdot \sum_{j=1}^{M} \sum_{\substack{i=1 \\ i\neq j}}^{M} G_{3,2}^{2,2}\left(\frac{(1-\rho(i,j))}{m} \frac{\Omega E_s}{2N_0} \,\middle|\, \begin{matrix} 1,\ 1-c,\ 1-m \\ 0,\ 1/2 \end{matrix} \right) \quad (20.25)$$

Figures 20.11, 20.12 and 20.13 show the performance of 2, 4 and 8-ary chirp system over K_G fading channel as a function of c and m. For comparison purposes, symbol error rate for Rayleigh fading and AWGN channel are also plotted. Because of the shadowing effect, the performance over K_G channel is poorer than the performance over AWGN and Rayleigh channel. Several techniques may be used to mitigate the effect of shadowing and fading on the performance. For example, at a symbol error rate of 10^{-4} for 2-ary chirp system, performance deteriorates by 5, 14, and 27 *dB* for generalized-K with $c = 6, m = 8$, generalized-K with $c = 6, m = 2$,

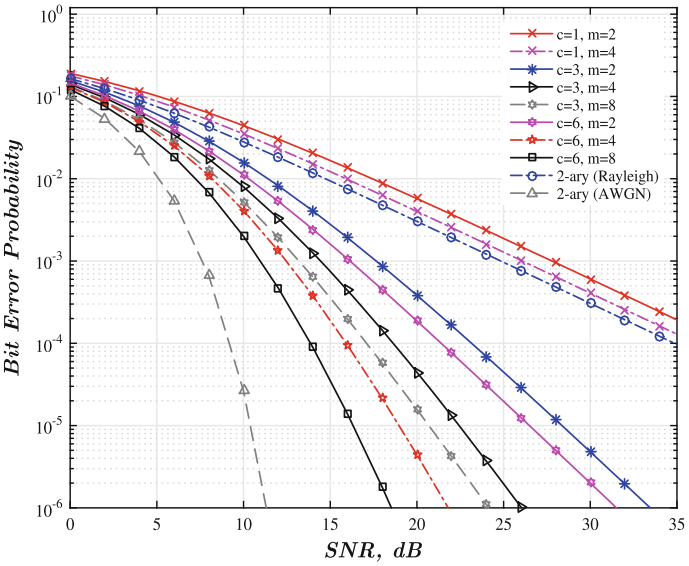


Fig. 20.11 Performance of 2-ary chirp system over K_G fading channel as a function of c and m

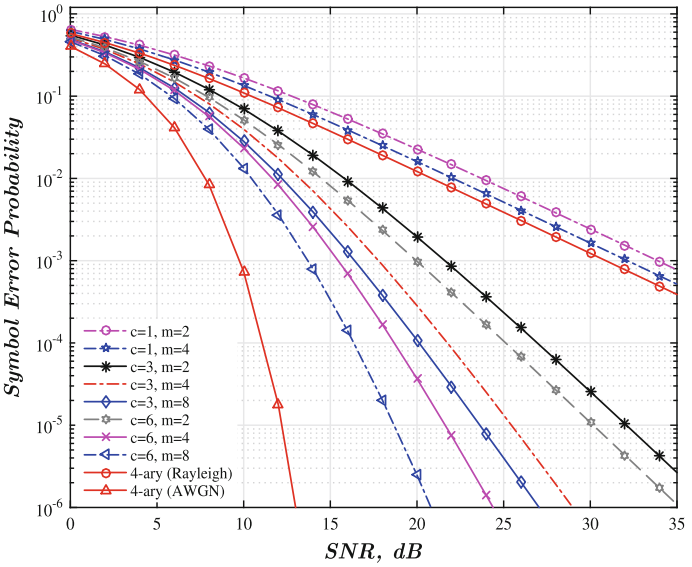


Fig. 20.12 Performance of 4-ary chirp system over K_G fading channel as a function of c and m

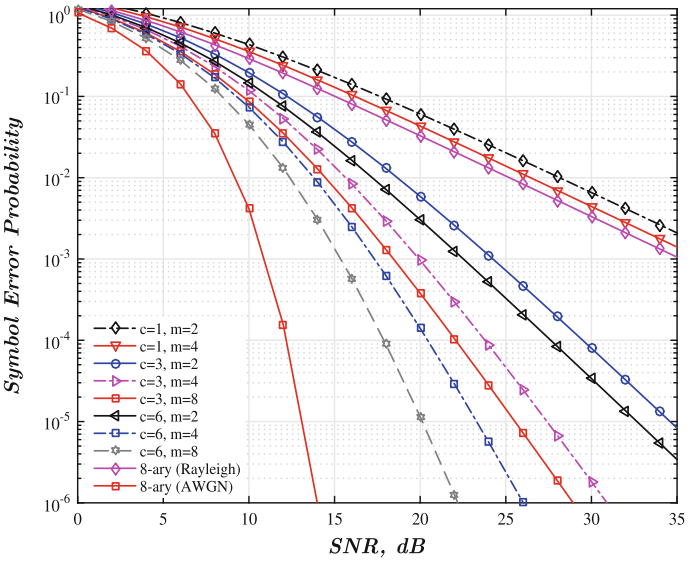


Fig. 20.13 Performance of 8-ary chirp system over K_G fading channel as a function of c and m

and Rayleigh channel model, respectively, with relative to the performance in AWGN channel. To compensate for this degradation in performance, channel coding and/or diversity techniques could be used. However, these techniques increase the channel bandwidth and system complexity.

20.4 Spectra of the *M*-ary Chirp Signals

The power spectrum of a stochastic signal $x(t)$ is defined as [13]:

$$G(f) = \lim_{\lambda \to \infty} \left[\frac{2}{\lambda} G_{\lambda}(f) \right], \quad f > 0 \tag{20.26}$$

where

$$G_{\lambda}(f) = \mathbb{E} \left\{ | \int_0^{\lambda} x(t) \exp \left[-j2\pi ft \right] \, dt|^2 \right\} \tag{20.27}$$

and $\mathbb{E}$ denotes the expectation operator. Using this definition of the power spectrum, many authors have derived closed-form expressions for the spectra of digital signals. Applying the direct method developed in [19] to *M*-ary chirp signals, we have been able to calculate the power spectra for different modulation order. The method can handle arbitrary *M*-ary data and works for arbitrary sets of modulation parameters q and w. The one-sided low pass spectrum is given by:

$$G(f) = \frac{2}{\hat{T}} [p(f) + 2Re\{F(f) *$$

$$F_b^*(f) \exp(-j2\pi f \hat{T}) + F(f) F_b^*(f) \Lambda(f)\}] \tag{20.28}$$

where

$$\Lambda(f) = \exp(-j4\pi f \hat{T}) \frac{C(\hat{T})}{1 - C(\hat{T}) exp(-j2\pi f \hat{T})} \tag{20.29}$$

and

$$b_n(t) = \pi \int_0^t [d_{1n} f(\tau) + d_{2n} f(\tau - T_s) + \ldots$$

$$+ d_{kn} f(\tau - kT_s)] \, d\tau, n = 1, 2, \ldots M^k \tag{20.30}$$

with

$$F_n(f) = \int_0^{\hat{T}} \exp(-j2\pi ft + jb_n(t)) \, dt \tag{20.31}$$

$$C(T) = E\{\exp(iB_n)\} \tag{20.32}$$

$$B_n = b_n(\hat{T}) \tag{20.33}$$

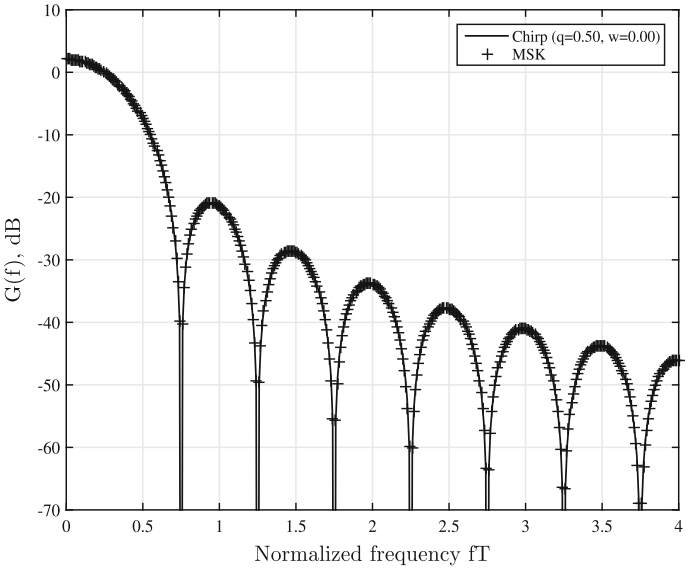


Fig. 20.14 Power spectra of the 2-ary chirp system ($q = 0.5, w = 0.0$) and MSK

$$F(f) = E\{F_n(f)\} = M^{-k} \sum_n F_n(f) \tag{20.34}$$

$$F_b^*(f) = E\{F_n^*(f) \exp(jB_n)\} = M^{-k} \sum_n F_n^*(f) \exp(jB_n) \tag{20.35}$$

$$P(f) = E\{|F_n(f)|^2\} = M^{-k} \sum_n |F_n(f)|^2 \tag{20.36}$$

E denotes the expectation operator with respect to the index $n = 1, 2, \cdots M^k$, $\hat{T} = kT_s$ is the signal interval in seconds. In (20.30), the function $f(\tau), f(\tau - T_s), \ldots, f(\tau - kT_s)$ denotes the frequency function of the M-ary during each time interval. The expressions (20.28)–(20.36) have been evaluated using Matlab. Equation 20.31 involves numerical integration for which the adaptive Simpson quadrature method was used. As a cross check, we have first considered the results of an established power and bandwidth efficient modulation scheme, the MSK. In the chirp pulse, if the modulation parameter w is set to zero, and q is 0.5, the pulse shape becomes identical to the pulse shape of the MSK. Therefore, the spectra of the binary chirp ($q = 0.5$ and $w = 0$) is sketched by using the direct method and by using the closed-form expression of the squared magnitude of the Fourier transform, given in [20]. In Fig. 20.14, the power spectrum of the 2-ary chirp system along with spectrum of the MSK are depicted. It is noted that both the spectra are identical.

Figure 20.15 shows the power spectra for 2-ary chirp systems using sets of modulation parameters that minimize the probability of error [13]. It is observed that the

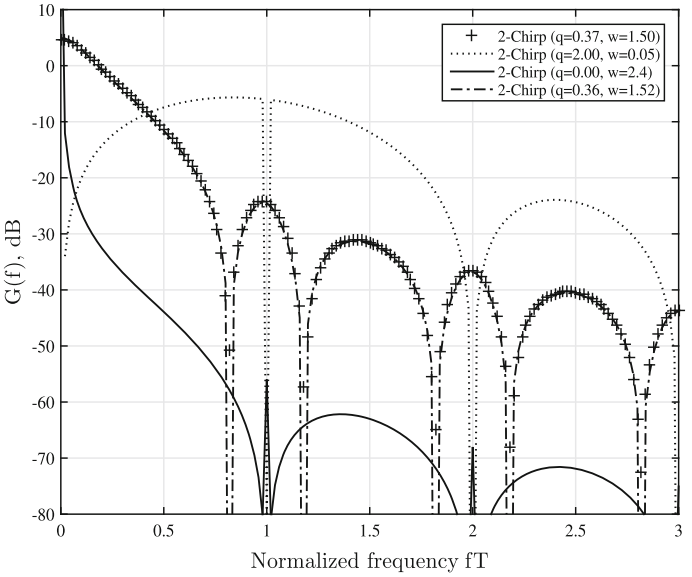


Fig. 20.15 Power spectra of the 2-ary chirp system

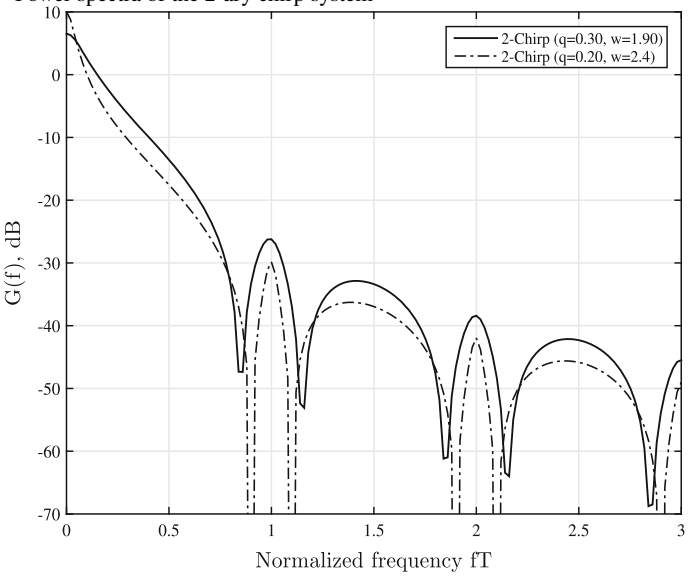


Fig. 20.16 Power spectra of the 2-ary chirp system

2-ary systems can offer different bandwidth occupancies by varying q and w of the set of modulation parameters (q, w).

Optimum 2-ary chirp system can offer 99 % bandwidth at a normalized frequency of 0.56 Hz/bit/s, that is slightly better than the bandwidth offered by the

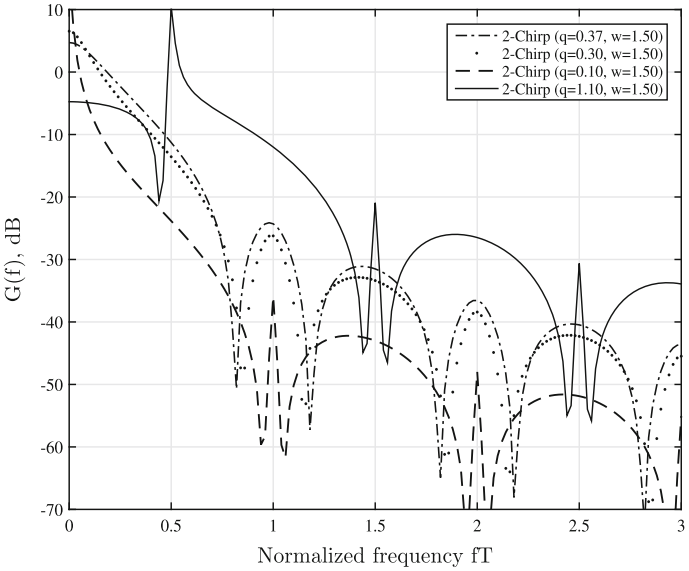


Fig. 20.17 Power spectra of 2-ary chirp system for a fixed value of w

MSK (0.60 Hz/bits/s), but has a wider main lobe. However, this set of modulation parameters does not provide a gain in dB relative to the MSK. Also, it is noted that other sets of modulation parameters can have different energy distributions; thus, power/bandwidth trades-offs are possible with these signals. Figure 20.16 shows the power spectra for modulation parameter sets that maximize the Euclidean distance over different observation lengths [13]. It is noted that the difference between the maximum value of the power in the main lobe and the maximum value in the first side lobe is nearly 20 dB for 2-ary with $(q = 0.30, w = 1.90)$ and 22 dB with $(q = 0.20, w = 2.40)$. This shows significant signal energy distributions in the side lobes. To examine the sensitivity of a 2-ary chirp system with respect to the modulation parameters, Fig. 20.17 shows the power spectra for different sets of modulation parameters, with a fixed value of $w = 1.5$. It is noted that the spectra is highly sensitive to small changes in the modulation parameters q. To compare power spectra as a function of the modulation level M, Figs. 20.18 and 20.19 show the spectra of 4-ary and 8-ary chirp systems. It is found that as M increases, the spectra become increasingly smooth; however, the signal energy distribution is almost the same throughout the spectrum.

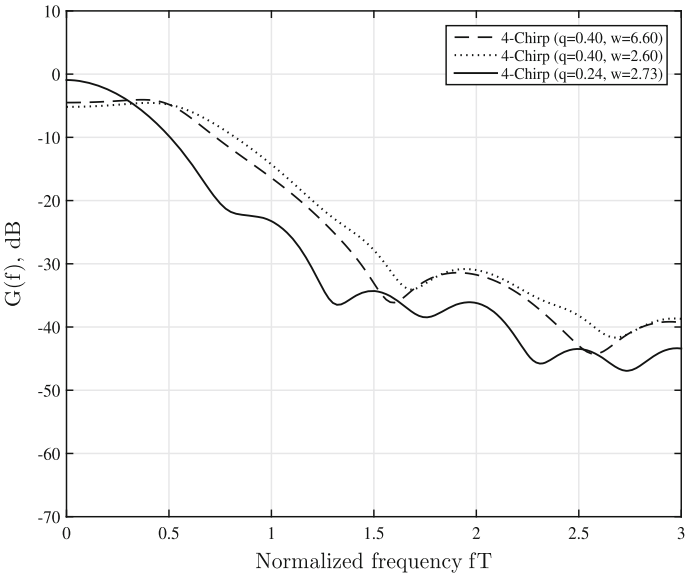


Fig. 20.18 Power spectra of a 4-ary chirp system

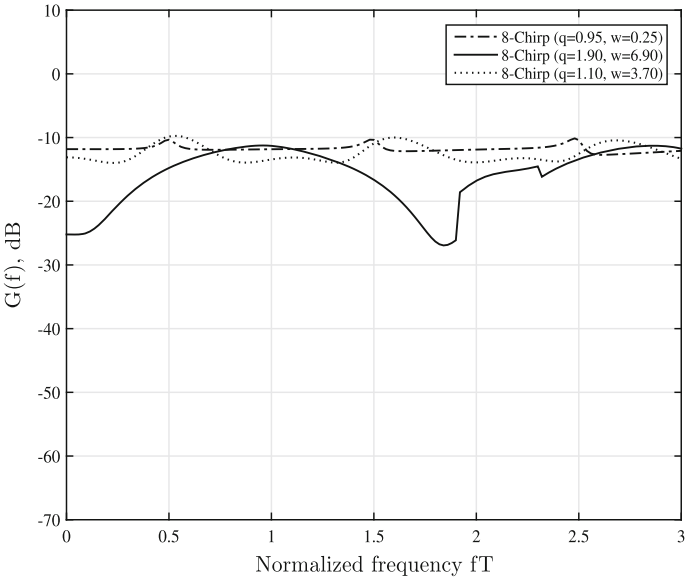


Fig. 20.19 Power spectra of an 8-ary chirp system

20.5 Conclusion

In this paper, closed-form SER expressions for M-ary chirp systems over Rayleigh, Nakagami-m and Generalized-K channels have been derived. The SER performances have been plotted as a function of normalized SNR for different modulation orders M and as a function of fading and shadowing parameters. One of the major contributions of this paper is the derivation of closed-form expressions for the performance over Rayleigh, Nakagami-m and Generalized-K channel models for M-ary chirp communication system. Also, a method called the direct method to calculate the power spectra of arbitrary M-ary signals is given. The method is very general and applies to arbitrary level of modulation, M, and can handle arbitrary frequency pulse shapes. It is observed that spectrum of 2-ary system is quite sensitive to even small variations in the sweep width parameter, w, of the modulation. Also, in 2-ary system there is significant energy distribution in the side lobes. As w increases, the ripple associated with the spectra of 2-ary system decreases. In general, as M increases, the spectra become smoother; however, the signal energy distribution becomes spread for a wide range of spectrum. M-ary system is thus an attractive modulation that can provide wide-band signals with good error rate behavior.

Acknowledgements The first author would like to gratefully thank Taif University for their support.

References

1. Huemer M, Koppler A, Ruppel CCW, Reindl L, Springer A, Weigel R (1999) SAW based chirp Fourier transform for OFDM systems. In: Ultrasonics Symposium, 1999. Proceedings. 1999 IEEE, vol.1, pp 373–376, vol.1, 17–20 Oct. 1999
2. The IEEE 802.15 Low Rate Alternative for Wireless Personal Area Networks (WPANs). http://www.ieee802.org/15/pub/TG4a.html
3. Nanotron technologies for embedded location platform. http://nanotron.com/EN/
4. Alsharef M, Hamed AM, Rao RK (2015) Error rate performance of digital chirp communication system over fading channels, In: Proceedings of The World Congress on Engineering and Computer Science 2015. Lecture Notes in Engineering and Computer Science, 21–23 October, 2015, San Francisco, USA, pp 727–732
5. Alsharef M, Rao RK (2015) M-ary chirp modulation for coherent and non-coherent data transmission. In: 2015 IEEE 28th Canadian Conference on Electrical and Computer Engineering (CCECE), pp 213–219, 3–6 May 2015
6. Gupta C, Mumtaz T, Zaman M, Papandreou-Suppappola A (2003) Wideband chirp modulation for FH-CDMA wireless systems: coherent and non-coherent receiver structures. In: IEEE International Conference on Communications, 2003. ICC'03, vol 4, pp 2455–2459, 11–15 May 2003
7. Khan MA, Rao RK, Wang X (2013) Closed-form error probability for M-ary chirp modulation in frequency-selective and -nonselective fading channels. In: 2013 26th Annual IEEE Canadian Conference on Electrical and Computer Engineering (CCECE), pp 1–4, 5–8 May 2013
8. Khan MA, Rao RK, Wang X (2013) Performance analysis of MRC-chirp system over independent and correlated fading channels. In: 2013 26th Annual IEEE Canadian Conference on Electrical and Computer Engineering (CCECE), pp 1–4, 5–8 May 2013

9. Hengstler S, Kasilingam DP, Costa AH (2002) A novel chirp modulation spread spectrum technique for multiple access. In: 2002 IEEE Seventh International Symposium on Spread Spectrum Techniques and Applications, vol 1, pp 73–77
10. Kadri A, Rao RK (2007) Binary chirp signals in—mixture noise: coherent and noncoherent detection. In: International Symposium on Signals, Systems and Electronics, 2007. ISSSE'07, pp 579–582, 30 July 2007–2 Aug 2007
11. Moungnoul P, Huang TT, Paungma T (2005) Investigation of multi-linear chirp FH-CDMA over fading channel model. In: 2005 Fifth International Conference on Information, Communications and Signal Processing, pp 1480–1484
12. El-Khamy SE, Shaaban SE, Thabet EA (1999) Frequency-hopped multi-user chirp modulation (FH/M-CM) for multipath fading channels. In: Proceedings of the Sixteenth National on Radio Science Conference, 1999. NRSC'99, pp C6/1–C6/8, 23–25 Feb 1999
13. Alsharef MA (2016) Constant-envelope multi-level chirp modulation: properties, receivers, and performance. Ph.D. dissertation, Department ECE, The University of Western Ontario, London
14. Simon MK, Alouini MS (2005) Digital communication over fading channels, 2nd edn. Wiley, New York
15. Hamed AM, Alsharef M, Rao RK (2015) Bit error probability performance bounds of CPFSK over fading channels. In: 2015 IEEE 28th Canadian Conference on Electrical and Computer Engineering (CCECE), pp 1329–1334, 3–6 May 2015
16. Prudnikov AP, Brychkov YA, Marichev OI (1990) Integrals and series. Vol. 3, Gordon and Breach Science Publishers, New York. More special functions; Translated from the Russian by GG Gould. MR 1054647 (91c:33001)
17. The Wolfram functions site for special mathematical functions (2015) Wolfram Research, Inc. http://functions.wolfram.com
18. Hamed AM, Rao RK, Primak SL (2014) Adaptive multidimensional modulation over faded shadowing channels. In: 2014 IEEE 27th Canadian Conference on Electrical and Computer Engineering (CCECE), pp 1–4, 4–7 May 2014
19. Wilson SG, Gaus R (1981) Power spectra of multi-h phase codes. IEEE Trans Commun 29(3):250–256
20. Proakis JG (2001) Digital communication, 4th edn. McGraw-Hill

Chapter 21
Power Factor Control Mechanism for Optimum Efficiency in Wind Generators and Industrial Applications

Jude K. Obichere, Milutin Jovanovic and Sul Ademi

21.1 Introduction

The brushless doubly-fed generator (BDFG) has been considered as a viable replacement to the traditional DFIG for wind generators [1–7]. In these applications, where only a limited variable speed capability is required (e.g. typically, in a 2:1 range or so [1, 4, 8]), the BDFG should retain the DFIG economic benefits of using a relatively smaller inverter (e.g. around 25 % of the machine rating), but with higher reliability and maintenance-free operation afforded by the absence of brush gear [9, 10].

The BDFG has two standard stator windings of different applied frequencies and pole numbers, unlike the DFIG. The primary (power) winding is grid-connected, and the secondary (control) winding is normally supplied from a bi-directional power converter. A BDFG reluctance type (Fig. 21.1), the brushless doubly-fed reluctance generator (BDFRG) [1–4], appears to be more attractive than its 'nested' cage rotor form, the brushless doubly-fed induction generator (BDFIG) [5–7, 11, 12]. This preference has been mainly attributed to the prospect for higher efficiency [2] and simpler control associated with the cage-less reluctance rotor [13]. However, the BDFG rotor must have half the total number of stator poles to provide the

J.K. Obichere (✉) · M. Jovanovic
Faculty of Engineering and Environment, Northumbria University Newcastle, Newcastle upon Tyne NE1 8ST, UK
e-mail: judekennedyobichere@yahoo.com

M. Jovanovic
e-mail: milutin.jovanovic@northumbria.ac.uk

S. Ademi
Institute for Energy and Environment, Department of Electronic and Electrical Engineering, University of Strathclyde, Glasgow G1 1XW, UK
e-mail: sul.ademi@strath.ac.uk

S.-I. Ao et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-981-10-2717-8_21

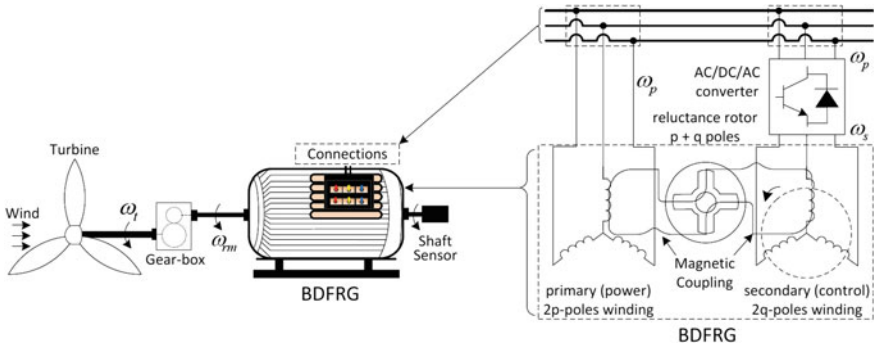


Fig. 21.1 Typical BDFRG drive set-up of variable speed WECS

rotor position dependent magnetic coupling between the stator windings required for the machine torque production [3].

With the recent introduction of the grid codes [14], another important BDFG merit is the superior low-voltage-fault-ride-through (LVFRT) capability to the DFIG [15–17]. It has been shown that owing to the larger leakage inductances and lower fault current levels, the LVFRT of the BDFIG may be accomplished safely without a crowbar circuitry [14, 18]. These potential LVFRT performance advantages over the DFIG can be carried over to the BDFRG featuring the leakage reactance values of the same order as the BDFIG.

Various control algorithms have been developed for the BDFRG including scalar control [1, 19], vector control (VC) [1, 20, 19, 21], direct torque control [19, 22], torque and reactive power control [23], direct power control [24], sliding mode power control [25], field-oriented control [21], and even non-linear Lyapunov control theory [8]. Although a comparative analysis of some of these control methods has been partly made in [19, 21] (and more detailed for the DFIG in [15, 26]), it is interesting that there is very little reported specifically on Power Factor Control (PFC) of the BDFRG being of utmost importance for generator applications [18]. In the BDFRG, the 'VC' term is commonly referred to as the primary winding voltage control, by analogy to the stator voltage control of DFIG.

With a proper selection of the reference frames and careful tuning of the dedicated PI controllers, a very good potential and dynamic response of the VC scheme has been demonstrated without knowledge of any machine parameters. The VC performance is examined using the optimum power factor control (OPFC) strategy [23, 24] on a large, custom-built 2 MW BDFRG [4]. This control objective has been considered because of the achievable efficiency gain at unity line power factor. Extensive realistic simulation results taking into account the usual practical effects (e.g. transducers' DC offset, noise in measurements, and a PWM power converter model) are presented to support the discussions in this paper.

21.2 Modeling Techniques of the BDFRG

The dynamic model of the BDFRG in arbitrary rotating reference frames, assuming motoring convention, can be represented using standard complex notation as [3]:

$$\left.\begin{aligned}
\mathbf{v}_p &= R_p\mathbf{i}_p + \frac{d\lambda_p}{dt} = R_p\mathbf{i}_p + \frac{d\lambda_p}{dt}\big|_{\theta_{p\,const}} + j\omega_p\boldsymbol{\lambda}_p \\
\mathbf{v}_s &= R_s\mathbf{i}_s + \frac{d\lambda_s}{dt} = R_s\mathbf{i}_s + \frac{d\lambda_s}{dt}\big|_{\theta_{s\,const}} + j\omega_s\boldsymbol{\lambda}_s \\
\boldsymbol{\lambda}_p &= \underbrace{L_p i_{pd} + L_{ps} i_{sd}}_{\lambda_{pd}} + j.\underbrace{(L_p i_{pq} - L_{ps} i_{sq})}_{\lambda_{pq}} \\
\boldsymbol{\lambda}_s &= \lambda_{sd} + j.\lambda_{sq} = \sigma L_s\mathbf{i}_s + \underbrace{\frac{L_{ps}}{L_p}\boldsymbol{\lambda}_p^*}_{\boldsymbol{\lambda}_{ps}}
\end{aligned}\right\} \tag{21.1}$$

where the primary and secondary winding are denoted by the subscripts 'p' and 's' respectively, $\sigma = 1 - \frac{L_{ps}^2}{L_p L_s}$ is the leakage factor, and λ_{ps} is the primary flux linking the secondary winding (i.e. the mutual flux linkage).

The fundamental angular velocity and torque relationships for the machine with p_r rotor poles and $\omega_{p,s} = 2\pi f_{p,s}$ applied angular frequencies to the respective 2p-pole primary and 2q-pole secondary windings are [3]:

$$\omega_{rm} = \frac{\omega_p + \omega_s}{p_r} \leftrightarrow n_{rm} = 60.\frac{f_p + f_s}{p_r} \tag{21.2}$$

$$T_e = \frac{3p_r}{2}\left(\lambda_{p_{sd}} i_{sq} - \lambda_{p_{sq}} i_{sd}\right) \tag{21.3}$$

$$T_a = J.\frac{d\omega_{rm}}{dt} = T_e - T_L(\omega_{rm}) - F.\omega_{rm} \tag{21.4}$$

Notice that $\omega_s > 0$ for 'super-synchronous' operation, and $\omega_s < 0$ at sub-synchronous speeds (i.e. an opposite phase sequence of the secondary to the primary winding) in (21.2) where $\omega_{syn} = \frac{\omega_p}{p_r}$ is the synchronous speed (for $\omega_s = 0$ i.e. a DC secondary) as with a $2p_r$-pole wound rotor synchronous turbo-machine. It is also worth mentioning that all the ω_p rotating vectors in the primary voltage/flux equations in (21.1) are in ω_p frame, whilst the corresponding secondary counterparts, including the λ_{ps} components in (21.3), are stationary in $p_r\omega_{rm} - \omega_p = \omega_s$ frame [3]. Given that λ_p and λ_{ps} in (21.3) are approximately constant by the primary winding grid connection, torque control can obviously be achieved through the secondary dq currents in the ω_s rotating frame.

Using (21.2), one can derive the mechanical power equation indicating individual contributions of each BDFRG winding:

$$P_m = T_e.\omega_{rm} = \underbrace{\frac{T_e.\omega_p}{p_r}}_{P_p} + \underbrace{\frac{T_e.\omega_s}{P_r}}_{P_s} = P_p.\left(1 + \frac{\omega_s}{\omega_p}\right) \tag{21.5}$$

The machine operating mode is determined by the power flow in the primary winding i.e. to the grid for the generating ($T_e < 0$) regime under consideration, while the secondary winding can either take or deliver real power (P_s) subject to its phase sequence i.e. the ω_s sign; the BDFRG would absorb (produce) $P_s > 0$ at sub- or super-synchronous speeds.

21.3 Controller System Mechanism

The detailed BDFRG system layout with a generic controller design is presented in Fig. 21.2. A standard phase-locked-loop (PLL) algorithm, readily available in the *Simulink* library, has been used to retrieve the stationary α-β frame angular positions (θ/θ_p) of the primary voltage/flux vectors from the measured voltages and/or currents. Furthermore, a conventional vector controller with space-vector PWM of the active rectifier has been implemented for control of DC link voltage and unity line power factor [21, 24, 27–30]. The primary real (P) and reactive (Q) power calculations are reference frame invariant and have been done using the stationary

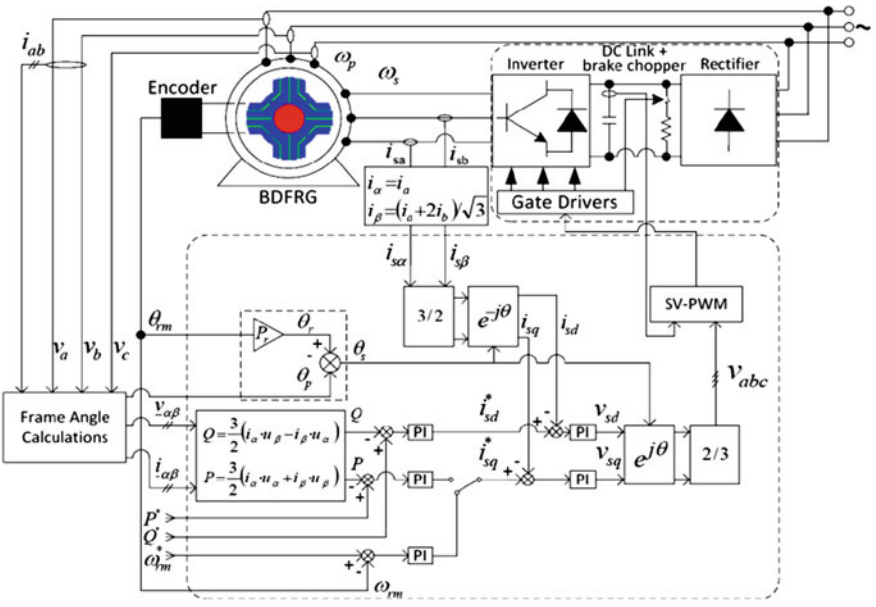


Fig. 21.2 Structural diagram of the BDFRG drive set-up

frame voltages ($v_{\alpha\beta}$) and currents ($i_{\alpha\beta}$) to avoid unnecessary conversions into their rotating d_s-q_s equivalents, and the use of time-consuming trigonometric functions, allowing so the higher control rates and superior performance in practice. The Q reference is often set to zero ($Q^* = 0$) for the unity primary power factor but can be any other value of interest for a given real power setting (P*) in power control mode, or the desired angular velocity ω^*_{rm} in variable speed systems. For example, either P* or ω^*_{rm} may correspond to the Optimum Power Point Tracking (OPPT) of a wind turbine [1, 8] while Q* was chosen to optimize efficiency of the WECS in this paper.

21.4 Voltage Control (Vc)

The control form expressions can be derived from the BDFRG space-vector model (21.1) in the natural reference frames, ω_p (i.e. d_p-d_q for the primary winding) and ω_s (i.e. d_s-q_s for secondary winding) rotating frames (Fig. 21.3), where the respective vector components appear as DC quantities. Substituting for $\mathbf{i}_p$ from the λ_p equation of (21.1) into $\mathbf{S}_p = \frac{3}{2}\mathbf{v}_p\mathbf{i}_p^*$ would lead to the following relationships for the primary mechanical and reactive power:

$$P_p = \frac{3}{2}\omega_p\left(\lambda_{p_{sd}} i_{sq} - \lambda_{p_{sq}} i_{sd}\right) \tag{21.6}$$

$$Q_p = \frac{3}{2}\omega_p\left(\frac{\lambda_p^2}{L_p} - \lambda_{p_{sd}} i_{sd} - \lambda_{p_{sq}} i_{sq}\right) \tag{21.7}$$

The control form expressions can be derived from the BDFRG space-vector model (21.1) in the natural reference frames, ω_p (i.e. d_p-d_q for the primary winding)

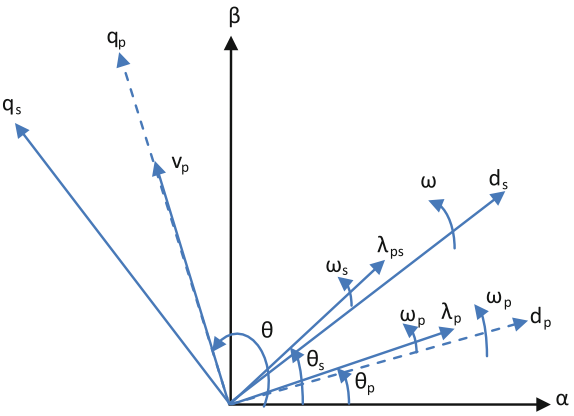


Fig. 21.3 Identification of primary voltage and flux vectors in a stationary α-β reference frame

and ω_s (i.e. d_s-q_s for secondary winding) rotating frames (Fig. 21.3), where the respective vector components appear as DC quantities.

Substituting for $\mathbf{i}_p$ from the λ_p equation of (21.1) into $\mathbf{S}_p = \frac{3}{2}\mathbf{v}_p\mathbf{i}_p^*$ would lead to the following relationships for the primary mechanical and reactive power:

$$P_p = \frac{3}{2}\omega_p\left(\lambda_{p_{sd}} i_{sq} - \lambda_{p_{sq}} i_{sd}\right) \tag{21.6}$$

$$Q_p = \frac{3}{2}\omega_p\left(\frac{\lambda_p^2}{L_p} - \lambda_{p_{sd}} i_{sd} - \lambda_{p_{sq}} i_{sq}\right) \tag{21.7}$$

A trade-off of the machine parameter independence is that VC of P_p and Q_p is coupled as both the i_{sd} and i_{sq} secondary current components appear in (21.6) and (21.7). The level of this cross-coupling can be reduced by aligning the q_p-axis of the reference frame to the primary voltage vector as shown in Fig. 21.3. In this case the primary flux vector (λ_p), would be phase-shifted ahead of the corresponding d_p-axis depending on the winding resistance values which are getting smaller with larger machines. Therefore, for the frame alignment choice as in Fig. 21.3, $\lambda_{p_{sd}} \gg \lambda_{p_{sq}}$ i.e. $\lambda_{p_{sd}} \approx \lambda_{ps}$ so that the approximations of (21.6) and (21.7) become:

$$P_p \approx \frac{3}{2}\omega_p\lambda_{ps} i_{sq} = \frac{3}{2}\frac{L_{ps}}{L_p}\omega_p\lambda_p i_{sq} \tag{21.8}$$

$$Q_p \approx \frac{3}{2}\frac{\omega_p\lambda_p^2}{L_p} - \frac{3}{2}\omega_p\lambda_{ps} i_{sd} \tag{21.9}$$

$$= \frac{3}{2}\frac{\omega_p\lambda_p}{L_p}\left(\lambda_p - L_{ps} i_{sd}\right) = \frac{3}{2}\omega_p\lambda_p i_{pd} \tag{21.10}$$

The P_p versus i_{sq} and Q_p versus i_{sd} functions above are nearly linear, which justifies the use of PI controllers in Fig. 21.2. Therefore, all inductance variations in (21.8)–(21.10) can be effectively taken care of by optimal tuning of the PI gains so their prior knowledge is not required for control purposes.

21.5 OPFC Mechanism of BDFRG Wind Generator

The preliminary performance of the VC scheme in Fig. 21.2 has been assessed using the parameters of a large-scale BDFRG [4, 28] in Table 21.1 for simulation studies. In order to make the simulations as realistic as possible, the following actions have been taken and/or assumptions made: (i) The power electronic models from the *SimPowerSystems* toolbox have been implemented; (ii) High-frequency uncorrelated white noise and unknown slowly varying DC offset have been

Table 21.1 The BDFRG design parameters, symbols and ratings

Rotor inertia [J]	3.8 kgm^2	Primary power [P]	2 MW
Primary resistance [R_p]	0.0375 Ω	Rated speed [n_r]	1000 rev/min
Secondary resistance [R_s]	0.0575 Ω	Stator currents [$I_{p,s}$]	1.5 kA rms
Primary inductance [L_p]	1.17 mH	Stator voltages [$V_{p,s}$]	690 V rms
Secondary inductance [L_s]	2.89 mH	Stator frequencies [$f_{p,s}$]	50 Hz
Mutual inductance [L_{ps}]	0.98 mH	Winding connections	Y/Y
Rotor poles [p_r]	4	Stator poles [p/q]	6/2

superimposed to the ideal signals to account for practical effects of the measurement noise and current/voltage transducers errors; (iii) Finally, the rotor position and speed information has been provided by a shaft sensor.

In a typical WECS, the turbine torque on the generator side of the gear-box for the maximum wind energy extraction in the base speed region (i.e. between the minimum 'cut-in', u_{min}, and the rated wind speed, u_r), can be represented as [1, 8]:

$$T_{opt} = \frac{A.\rho.C_p(\lambda_{opt}, \gamma).R^3}{2.g^3.\lambda_{opt}^3} . \omega_{rm}^2 = K_{opt} . \omega_{rm}^2 \tag{21.11}$$

where ρ is the air density, $C_p(\lambda, \gamma)$ is the power coefficient (i.e. the maximum turbine efficiency as $\lambda = \lambda_{opt}$ in this case), $\lambda_{opt} = \frac{R_t}{u}$ is the optimum tip speed ratio for a given wind speed u, ω_t is the turbine rotor angular velocity, γ is the pitch angle (normally fixed to zero to maximise C_p), R the radius of the circular swept area ($A = \pi R^2$), and $g = \frac{\omega_{rm}}{\omega_t}$ is the gear ratio. The shaft torque-speed profile in (21.4) is of the same form as (21.11):

$$T_L = -\frac{P_r}{\omega_r} . \left(\frac{n_{rm}}{n_{max}}\right)^2 \approx -19 . \left(\frac{n_{rm}}{1000}\right)^2 \text{kNm} \tag{21.12}$$

21.6 Simulation Results

The simulation results in Figs. 21.4, 21.6 and 21.8 have been produced by running the control algorithms in Fig. 21.2 at 5 kHz switching rate for the IGBT converter. The DC link voltage has been maintained at ≈1200 V by the PWM rectifier (i.e. the line-side bridge) supplied at 690 V, 50 Hz. The reference speed trajectory is set as a steep ramp signal suited for dynamically not very demanding wind power applications even under extreme turbulent wind conditions (Figs. 21.4 and 21.5).

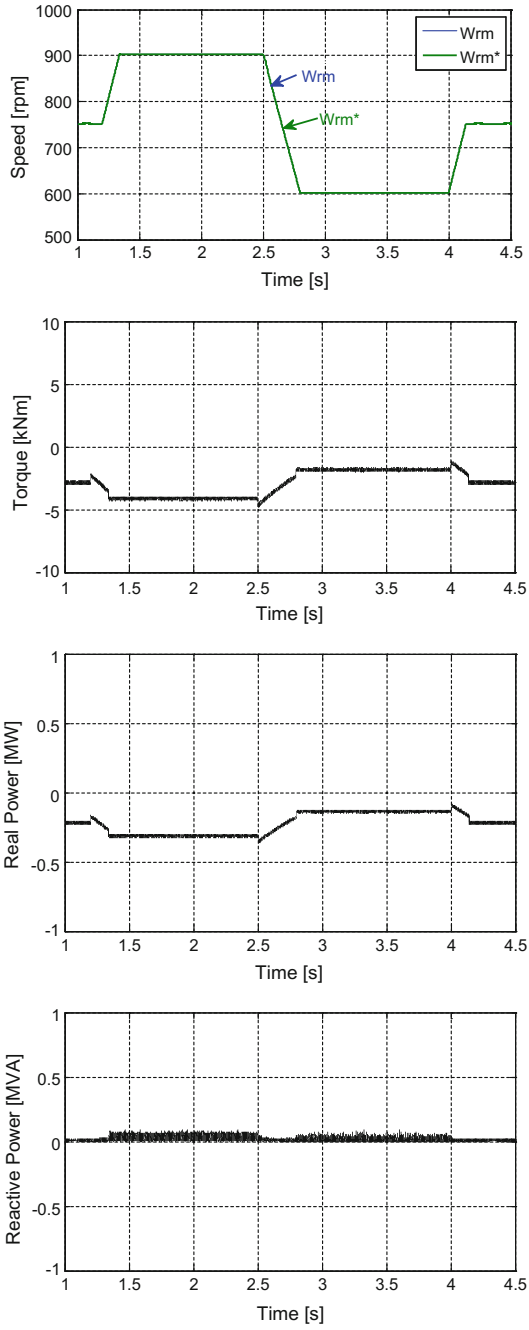


Fig. 21.4 Simulation of the BDFRG operation under OPFC in a limited range around synchronous speed (750 rpm)

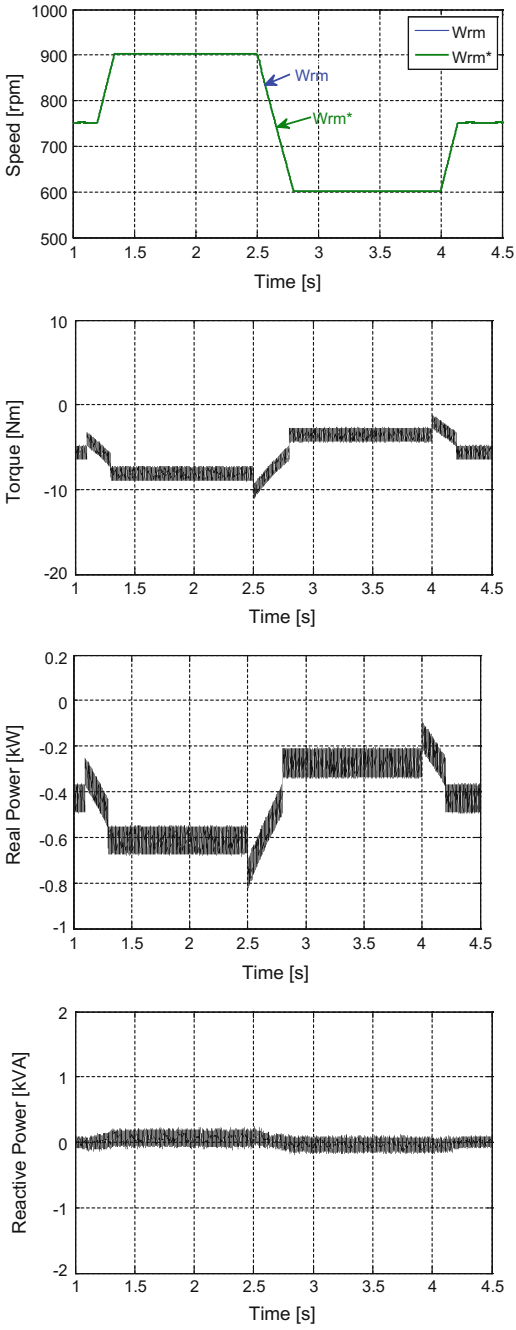


Fig. 21.5 Experimental validation of the simulations in Fig. 21.4 for the small-scale BDFRG

The top plots in Fig. 21.4 show the excellent speed tracking with no overshoot following the start-up period in both super- and sub-synchronous modes of the BDFRG over the limited speed range of 600–900 rpm. The primary electrical power (P) and electro-magnetic torque (T_e) curves reflect (21.12) for the specific speed settings. Except for a difference in losses, and considering that $\omega_p \approx$ const, P and T_e are directly related as follows from (21.5) and (21.6) which explains a close resemblance in their shape.

The T_e deviations from the desired profile during the speed transients refer to the acceleration or deceleration torque term in (21.4) depending on whether the machine is to speed-up ($T_a > 0$) or slow-down ($T_a < 0$). The reasonably accurate and smooth $Q_p \approx 0$ control can be observed to be little affected by the variable speed dependent loading disturbances, which means that the optimum power factor control (OPFC) conditions have been largely met. Note that the outer PI speed control loop is required to ensure effective variable speed operation at optimum tip-speed ratio.

The secondary ($i_{sd,q}$) and primary ($i_{pd,q}$) current waveforms in Fig. 21.6 show no transient over-currents as the PI regulators do not need to be saturated to allow accurate tracking of the desired trajectories for the moderate ramp speed variations. A close link between the active q currents and the real power (torque), as well as the magnetizing d currents and Q, is immediately visible from the respective waveforms. The coupling effects of the i_{sq} clearly manifest themselves as speed (torque) dependent disturbance (e.g. offsets) in the respective non-controllable i_{pd} profiles by analogy to the P and Q scenario in Fig. 21.4.

Figure 21.8 illustrates the step-wise PWM angular positions of the modulated secondary voltage vector (v_s), and the respective space sector variations, as well as

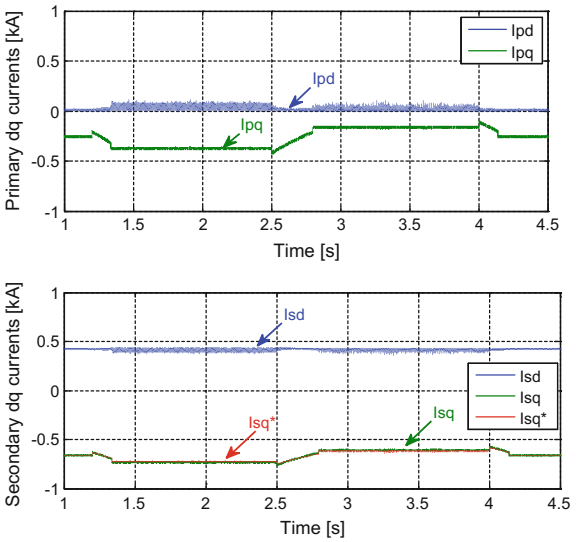


Fig. 21.6 Simulated OPF responses of the BDFRG current components in the corresponding rotating reference frames

the respective secondary currents, during a speed reduction, from 900 rev/min. to 600 rev/min. That is, the speed mode change, from super- to sub-synchronous. In the super-synchronous mode, v_s rotates anti-clockwise as indicated by the ascending angular steps for the same phase sequence of the windings and $\omega_s > 0$ in (21.2). The situation is reversed at sub-synchronous speeds when v_s rotates clockwise with the angular steps descending, which comes from the opposite phase sequence of the secondary to the primary winding since $\omega_s < 0$ in (21.2). Notice that v_s becomes stationary at synchronous speed as the secondary currents are then DC i.e. $\omega_s = 0$ in (21.2).

The respective phase voltage and current waveforms which are π-rad out of phase as expected for unity power factor control (−1 in generating mode) in steady-state period are discussed and demonstrated in [18].

21.7 Experiment and Results

The experimental validation of the simulation model was carried out using the small-scale BDFRG test rig described in [22–24]. The 6/2-pole machine prototype of Y/Y winding connection is rated at 1.5 kW, 1000 rev/min, 2.5 A, 400 V, 50 Hz with the following parameters: J = 0.1 kg^2, R_p = 11.1 Ω, R_s = 13.5 Ω, L_p = 0.41 H, L_s = 0.57 H, and L_{ps} = 0.34 H. The experimental plots in Figs. 21.5, 21.7 and 21.9 are respectively produced by executing the VC scheme in Fig. 21.2 on a Simulink® compatible dSPACE® platform at 5 kHz inverter switching rate under

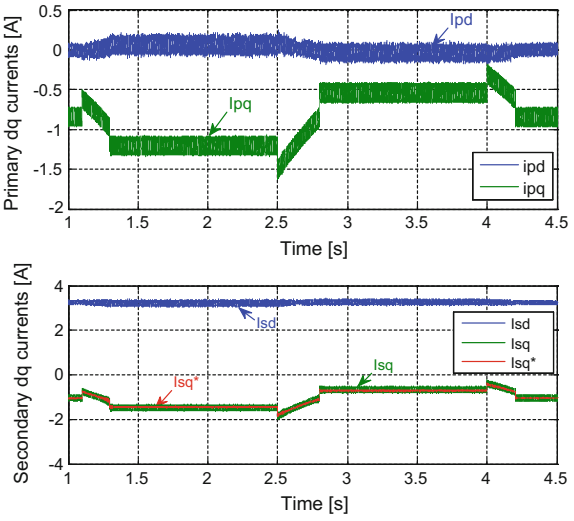


Fig. 21.7 Experimental validation of the simulations in Fig. 21.6, for the small-scale BDFRG

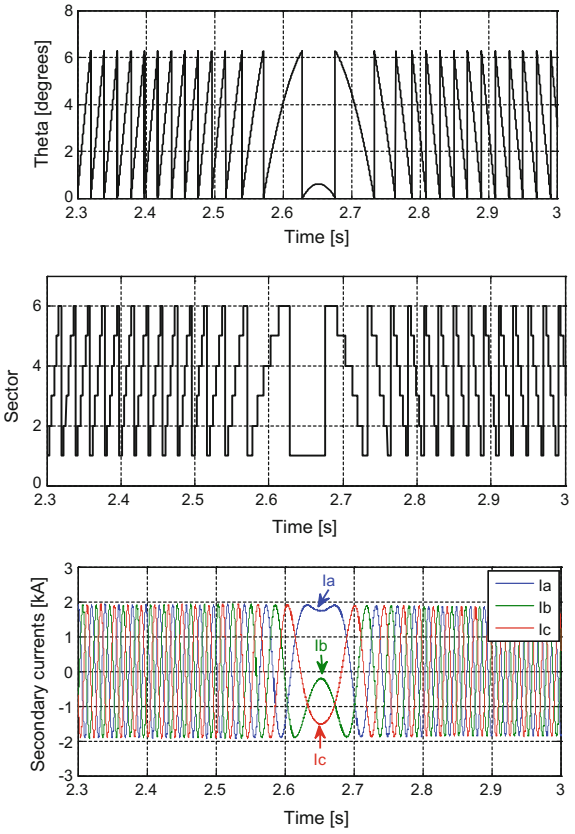


Fig. 21.8 PWM space sector, angular (theta) positions, secondary currents and switching times of active voltage vectors of the BDFRG showing a phase sequence speed reversal at unity power factor

similar operating conditions as simulated in Figs. 21.4, 21.6 and 21.8. A commercial DC machine drive has emulated the required variable prime mover (i.e. wind turbine) torque, measured by a Magtrol® torque transducer, as specified by (21.12) but with 14 Nm loading at 1000 rev/min. in real-time.

Although a generally good agreement between the simulation and the experimental results is more than obvious from the generated waveforms this can further be enhanced by the PI speed regulator. It should also be noted that unlike the large-scale case in Fig. 21.4, the reactive power Q, response in Fig. 21.5 are severely affected by the speed dependent P variations as a consequence of the much more pronounced cross-coupling effects of the i_{sq} term in (21.7) due to the unusually high winding resistances of the test BDFRG.

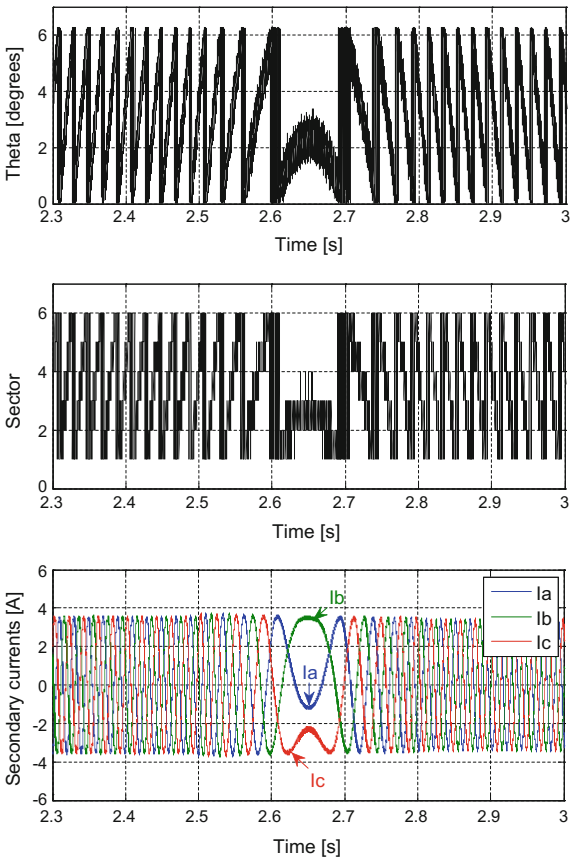


Fig. 21.9 Experimental validation of the simulations in Fig. 21.8 for the small-scale BDFRG

21.8 Conclusion and Future Work

A novel power factor control (PFC) algorithm for optimum efficiency of the BDFRG has been carried out and evaluated by comprehensive simulation studies and experimental validation, using small-scale BDFRG set-up, under typical operating conditions. Such vector control implementation can serve as a platform for further research on this emerging and prominent brushless machine topology for applications in large wind turbines and industries, where the cost advantages of partially-rated power electronics and high reliability of brushless structure can be fully utilized.

The simulation and experimental works in Matlab®/Simulink® have clearly demonstrated the immense potential and effectiveness of the controller(s) using the power factor control mechanism. The presented results are more than encouraging and can serve as basis for further research work on this machine.

References

1. Jovanovic MG, Betz RE, Yu J (2002) The use of doubly fed reluctance machines for large pumps and wind turbines. IEEE Trans Ind Appl 38:1508–1516
2. Wang F, Zhang F, Xu L (2002) Parameter and performance comparison of doubly-fed brushless machine with cage and reluctance rotors. IEEE Trans Ind Appl 38(5):1237–1243
3. Betz RE, Jovanovic MG (2003) Introduction to the space vector modelling of the brushless doubly-fed reluctance machine. Electr Power Compon Syst 31(8):729–755
4. Dorrell DG, Jovanovi´c M (2008) On the possibilities of using a brushless doubly-fed reluctance generator in a 2 MW wind turbine. In: IEEE industry applications society annual meeting, pp 1–8, Oct 2008
5. McMahon RA, Roberts PC, Wang X, Tavner PJ (2006) Performance of BDFM as generator and motor. IEE Proc-Electr Power Appl 153(2):289–299, Mar 2006
6. Poza J, Oyarbide E, Sarasola I, Rodriguez M (2009) Vector control design and experimental evaluation for the brushless doubly fed machine. IET Electr Power Appl 3(4):247–256
7. Protsenko K, Xu D (2008) Modeling and control of brushless doubly-fed induction generators in wind energy applications. IEEE Trans Power Electron 23(3):1191–1197
8. Valenciaga F, Puleston PF (2007) Variable structure control of a wind energy conversion system based on a brushless doubly fed reluctance generator. IEEE Trans Energy Convers 22 (2):499–506
9. Polinder H, van der Pijl F, de Vilder G, Tavner P (2006) Comparison of direct-drive and geared generator concepts for wind turbines. IEEE Trans Energy Convers 21(3):725–733
10. Spinato F, Tavner PJ, van Bussel G, Koutoulakos E (2009) Reliability of wind turbine subassemblies. IET Renew Power Gener 3(4):387–401
11. Tohidi S, Zolghadri M, Oraee H, Tavner P, Abdi E, Logan T (2012) Performance of the brushless doubly-fed machine under normal and fault conditions. IET Electr Power Appl 6 (9):621–627
12. Barati F, McMahon R, Shao S, Abdi E, Oraee H (2013) Generalized vector control for brushless doubly fed machines with nested-loop rotor. IEEE Trans Ind Electron 60(6):2477–2485
13. Knight A, Betz R, Dorrell D (2013) Design and analysis of brushless doubly fed reluctance machines. IEEE Trans Ind Appl 49(1):50–58
14. Long T, Shao S, Malliband P, Abdi E, McMahon R (2013) Crowbar-less fault ride-through of the brushless doubly fed induction generator in a wind turbine under symmetrical voltage dips. IEEE Trans Ind Electron 60(7):2833–2841
15. Cardenas R, Pena R, Alepuz S, Asher G (2013) Overview of control systems for the operation of DFIGs in wind energy applications. IEEE Trans Ind Electron 60(7):2776–2798
16. Marques G, Sousa D (2012) Understanding doubly fed induction generator during voltage dips. IEEE Trans Energy Convers 27(2):421–431
17. Tohidi S, Tavner P, McMahon R, Oraee H, Zolghadri M, Shao S, Abdi E (2014) Low voltage ride-through of DFIG and brushless DFIG: Similarities and differences. Electr Power Syst Res 110:64–72
18. Obichere JK, Jovanovic M, Ademi S (2015) Power factor control of large doubly-fed reluctance wind generators. In: Lecture notes in engineering and computer science: Proceedings of The World Congress on Engineering and Computer Science 2015, WCECS 2015. San Francisco, USA, pp. 210–215, 21–23 Oct 2015
19. Jovanovic M (2009) Sensored and sensorless speed control methods for brushless doubly fed reluctance motors. IET Electr Power Appl 3(6):503–513
20. Xu L, Zhen L, Kim E (1998) Field-orientation control of a doubly excited brushless reluctance machine. IEEE Trans Ind Appl 34(1):148–155
21. Ademi S, Jovanovic M, Obichere JK (2014) Comparative analysis of control strategies for large doubly-fed reluctance wind generators. In: Lecture notes in engineering and computer

science: proceedings of the world congress on engineering and computer science 2014, WCECS 2014. San Francisco, USA, pp 245–250, 22–24 Oct 2014
22. Jovanovi´c MG, Yu J, Levi E (2006) Encoderless direct torque controller for limited speed range applications of brushless doubly fed reluctance motors. IEEE Trans Ind Appl 42 (3):712–722
23. Chaal H, Jovanovi´c M (2012) Practical implementation of sensorless torque and reactive power control of doubly fed machines. IEEE Trans Ind Electron 59(6):2645–2653
24. Chaal H, Jovanovic M (2012) Power control of brushless doubly-fed reluctance drive and generator systems. Renew Energy 37(1):419–425
25. Valenciaga F (2010) Second order sliding power control for a variable speed-constant frequency energy conversion system. Energy Convers Manag 52(12):3000–3008
26. Tremblay E, Atayde S, Chandra A (2011) Comparative study of control strategies for the doubly fed induction generator in wind energy conversion systems: a DSP-based implementation approach. IEEE Trans Sustain Energy 2(3):288–299
27. Malinowski M, Kazmierkowski M, Trzynadlowski A (2003) A comparative study of control techniques for PWM rectifiers in ac adjustable speed drives. IEEE Trans Power Electron 18 (6):1390–1396
28. Ademi S, Jovanovic M (2014) Control of emerging brushless doubly fed reluctance wind turbine generators. In: Hossain J, Mahmud A (eds) Large scale renewable power generation, ser. green energy and technology. Springer, Singapore, pp 395–411
29. Ademi S, Jovanovic M (2015) Vector control methods for brushless doubly fed reluctance machines. IEEE Trans Ind Electron 62(1):96–104
30. Ademi S, Jovanovic M (2014) High-efficiency control of brushless doubly-fed machines for wind turbines and pump drives. Energy Convers Manag 81:120–132

Chapter 22
TBSC Compensator

Swapnil Patil, Khushal Shende, Dadgonda Patil and Anwar Mulla

22.1 Introduction

It is well documented in literature and through public discussions at various levels that a substantial power loss is taking place in our low voltage distribution systems on account of poor power factor, due to inadequate reactive power compensation facilities and their improper control. Switched LT capacitors can directly supply the reactive power of loads and improve the operating condition. Government of India has been insisting on shunt capacitor installations in massive way and encouraging the state electricity boards through Rural Electrification Corporation and various other financing bodies. The expansion of rural power distribution systems with new connections and catering to agricultural sector in wide spread remote areas, giving rise to more inductive loads resulting in very low power factors [1].

The voltages at the remote ends are low and the farmer's use high HP motors operating at low load levels with low efficiencies. This is giving rise to large losses in the distribution network. Thus there exists a great necessity to closely match reactive power with the load so as to improve power factor, boost the voltage and reduce the losses. The conventional methods of reactive power supply are through

S. Patil (✉) · D. Patil · A. Mulla
Department of Electrical Engineering, Annasaheb Dange College of Engineering and Technology, Ashta, Maharashtra, India
e-mail: swapnil6006@rediffmail.com

D. Patil
e-mail: dadasorpatil@gmail.com

A. Mulla
e-mail: ammaitp@rediffmail.com

K. Shende
Department of Electrical Engineering, Walchand College of Engineering, Sangli, Maharashtra, India
e-mail: khushalshende@gmail.com

S.-I. Ao et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-981-10-2717-8_22

switched LT capacitors, mostly in equal steps in various automatic power factor controllers developed by number of companies. In this paper, a more reliable, technically sound, fast acting and low cost scheme is presented by arranging the thyristor switched capacitor units in five binary sequential steps. This enables the reactive power variation with the least possible resolution. As there is reduction in loss with shunt compensation in the feeders, the efficiency increases and conservation of energy takes place. Besides the enhancement of transformer loading capability the shunt capacitor also improves the feeder performance, reduces voltage drop in the feeder and transformer, better voltage at load end, improves power factor and improves system security, increases over all efficiency, saves energy due to reduced system losses, avoids low power factor penalty, and reduces maximum demand charges [2, 3].

Induction motors (I.M.) load constitute a large portion of power system. Three-phase induction motors represent the most significant load in the industrial plants, over the half of the delivered electrical energy. Starting of induction motor may cause a problem of voltage sag in the power system. The IEEE defines voltage sag as: A decrease to between 0.1 and 0.9 p.u. in rms voltage or current at the power frequency for durations of 0.5 cycle to 1 min. An induction motor at rest can be modeled as a transformer with the secondary winding short circuited. Thus when full voltage is applied, a heavy inrush current (of 6 to 10 times the rated value) is drawn from the power system that causes voltage sag. As the motor accelerates and attains the rated speed, the inrush current decays and the system voltage recovers [4].

Voltage sag can cause mal-operation of voltage sensitive devices such as computers, relays, programmable logic controllers etc. Also because of the highly inductive nature of the motor circuit at rest, the power factor is very low, usually of the order of 10 to 20 %. Thus reactive power demand at the starting of I.M. is very high and it reduces as motor picks up the speed. There are several solutions to minimize this problem; the most common are reactor start, auto transformer start, star-delta, capacitor start, soft starter, frequency variable driver (FVD) etc. All these methods except capacitor start are based on a motor terminal voltage reduction to decrease the rotor current, reducing the line voltage drop. Problem with this method of starting is that the motor torque is directly proportional to the square of the supply voltage hence decrease in the motor terminal voltage will cause the motor torque to decrease, which may be insufficient for driving the required load. Soft starter and frequency variable driver methods are the most expensive and complex, requiring more expert maintenance. In capacitor start system, reactive current required by the motor during acceleration is supplied by capacitors which reduce the source current. This in turn reduces the magnitude of voltage sag in the system. Capacitor start method has a lower cost in comparison with other methods however one has to consider the transitory effects of switching of capacitor banks [5, 6].

22.2 Objectives

This topology has the following distinctive characteristics [7, 8]:

- It maintains the power factor at the PCC to any specified value.
- It compensates for rapid variation in reactive power or voltages.
- Maximum compensation time is 20 ms.
- No transients or harmonics are allowed to be present due to fast selective instants of switching in well co-ordinate manner.
- It is adaptive in the sense that the amount of the compensation is determined and provided on a cycle by cycle basis.
- It can compensate each phase independently which makes ideal for unbalanced systems.
- Capacitors are sized in binary sequential ratio for minimum size of switching steps.
- The control strategy is error activated to match with the load reactive power for the chosen time interval.
- It eliminates possible over compensation and resulting leading power factor.
- It is flexible to choose required number of steps as per the resolution.
- Resolution can be made small with more number of steps.
- Simple in principle, elegant in usage and of low cost.
- Possible to incorporate the idea presented in the controllers for large size transformers at substations.

22.3 Theme

The Proposed work demonstrates the power quality problem due to reactive power in the system. In this proposed scheme TBSC Compensator is presented. TBSC Bank means Thyristor Binary Switched Capacitor Bank [9, 10].

- Normally we are using SVC (TSC + TCR) for reactive power compensation. But SVC has some drawbacks due to TCR.
- TCR generates Harmonics.
- More losses in TCR.
- SVC required additional filter for Harmonic reduction which is generated by TCR.
- Cost of SVC is more due to TCR and Filter.

To avoid these drawbacks we design a new compensator that is TBSC Compensator. Also we are added one additional concept that is Transient Free Switching of TBSC Bank. With the help of TBSC Compensator we can achieve almost stepless control of reactive power.

- The automatic power factor controllers available at present in the market do not offer exact matching with reactive power required by the load during specified interval which leads to inaccuracies in power factor control and reactive power compensation. In order to overcome these limitations and to ensure precise control, the bank is arranged in binary sequential steps to enable reactive power variation with least possible resolution and all most stepless way.

22.4 Proposed Work

The objective of this work is to present "An Innovative Transient Free TBSC Compensator for Dynamic Reactive Load and Voltage Sag Mitigation" scheme for reactive power compensation and power quality improvement in power system.

The proposed scheme consists of Thyristor Switched Capacitor (TSC) banks in binary sequential steps known as Thyristor Binary Switched Capacitor (TBSC) [11, 12]. This TBSC facilitates step-less control of reactive power closely matching with load requirements so as to maintain desired power factor.

The schematic diagram of the capacitor bank in binary sequential steps through thyristor and with respective current limiting reactors is shown in Fig. 22.1 TBSC compensator connected at the point of common coupling (PCC) for reactive power compensation is shown in Fig. 22.1 and the operating principle of each equipment is analyzed in the following sections.

TBSC consists of an anti-parallel connected thyristor as a bidirectional switch in series with a capacitor and a current limiting small reactor. Transient free switching of capacitors is obtained by satisfying following two conditions [13]:

- Firing the thyristors at the negative/positive peak of supply voltage.
- Precharging the Capacitors to the negative/positive peak of supply voltage.

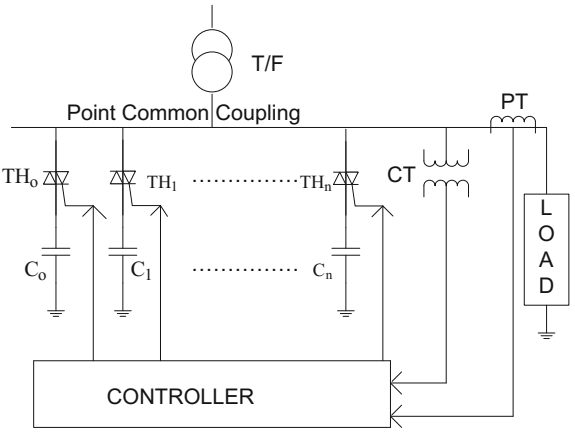


Fig. 22.1 TBSC compensator

TBSC current is sinusoidal and free from harmonics, thus eliminating the need for any filters. Small-series inductor is placed in series with capacitor. It serves following purposes.

- It limits current transients during overvoltage conditions and when switching at incorrect instants or at the inappropriate voltage polarity.
- The chosen inductor magnitude gives a natural resonant frequency of many times the system nominal frequency. This ensures that the inductance neither creates a harmonic-resonant circuit with the network nor dampers the TBSC control system.

In the proposed scheme, capacitor bank step values are chosen in binary sequence weights to make the resolution small. If such 'n' capacitor steps are used then 2n different compensation levels can be provided [3].

$$Q = 2^{n}C + 2^{n-1}C + \cdots + 2^{2}C + 2^{1}C + 2^{0}C$$

- TBSC Closed Loop Operation:

A block diagram of reactive power compensator using TBSC banks is shown in Fig. 22.2 Reference reactive power, Q_{Ref} is calculated from the desired power factor. Actual reactive power at PCC, Q_{Actual} is calculated by sensing voltage and current at PCC by P.T. and C.T. respectively. Error between Q_{Ref} and Q Actual is given to PI Controller. A Discrete PI Controller is used. Output of PI Controller is given to ADC and its output is given to TBSC banks in such a way that no transients occur. In this way closed loop operation of TBSC banks for reactive power compensation is achieved.

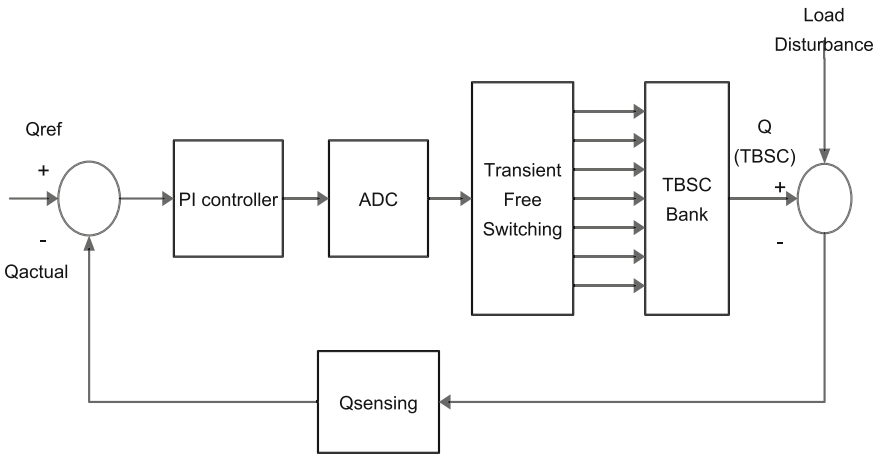


Fig. 22.2 TBSC closed loop operation

22.5 Simulation Results

- *Binary Current Generation*:

The Fig. 22.3 shows the binary operation of the TBSC compensator proposed in Fig. 22.1 The total compensating current from phase "R" (total ic), is being increased step by step. The capacitor currents from the branches Bl (ic1), B2 (ic2), B4 (ic4), and B8 (ic8) are shown in Fig. 22.3 respectively. In Fig. 22.3 the total compensating current for the phase "R" (total ic) is displayed (total ic = ic1 + ic2 + ic4 + ic8).

The total compensating current from phase "R" (total ic), is being increased step by step. The capacitor currents from the branches Bl (ic1), B2 (ic2), B4 (ic4), and B8 (ic8) are shown in Fig. 22.3 respectively. In Fig. 22.3 the total compensating current for the phase "R" (total ic) is displayed (total ic = ic1 + ic2 + ic4 + ic8). It can be noted that harmonics or inrush problems are not generated, and that the current total ic seems to vary continuously. The transitions during connection and disconnection are quite clean.

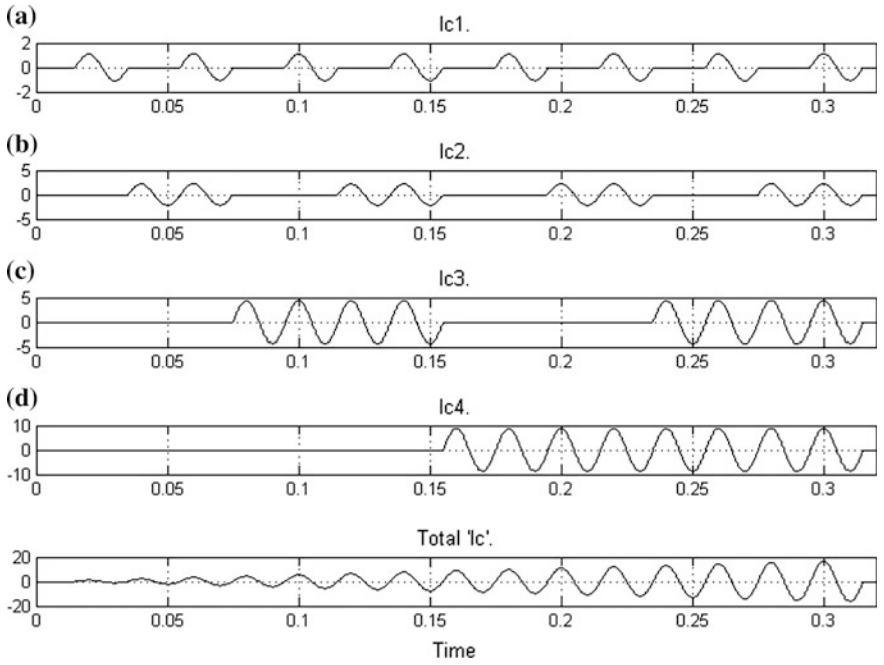


Fig. 22.3 Compensating current for phase "R" (Ic1) Current through B1 **b** (Ic2) Current through B2 (Ic3) Current through B4 **d** (Ic4) Current through B8

- *TBSC Compensator for fast varying dynamic load*:

Continuously changing reactive power, Q_L is obtained by simulating three phase dynamic load. The nature of load variation is as shown in Fig. 22.4.

Minimum reactive power Q_{min}, maximum reactive power Q_{max}, and base reactive power Q_{Base} can be varied by changing the parameters of three phase dynamic load. In all simulations Q_{Ref} is set to zero since it is assumed that desired P.F. is unity at all times.

- *TBSC Closed Loop Operation for dynamic load*:

Discrete PI controller with KP = 0.565 and KI = 25 is used. 5 bit ADC is used in simulation. Parameters of Three-phase dynamic load block are adjusted in such a way that Q_L varies continuously from $Q_{Min.}$ = 2.5 KVAR to $Q_{Max.}$ = 77 KVAR with base load $Q_{Base.}$ = 40 KVAR. This variation takes place in five seconds. Waveforms of load reactive power Q_L, reactive power given by TBSC, $Q_{comp.}$ (TBSC) and actual reactive power Q_{Actual} at PCC are shown in Fig. 22.5.

From simulation results it is seen that $Q_{comp.}$(TBSC) closely follows Q_L shown in Fig. 22.5, and actual reactive power Q_{Actual} at PCC is approximately +500 to −500 VARs at all discrete switching instances. The small error is due to the binary switching arrangement of TBSCs. These errors can be minimized by adding more number of capacitor banks in TBSC.

- *TBSC Compensator for voltage sag mitigation of induction motors*:

Waveforms of I.M. reactive power demand Q_{Motor} and reactive power given by Q_{TBSC} are shown in Fig. 22.6. From simulation results it is seen that Q_{TBSC} closely follows Q_{Motor} and actual reactive power Q_{Actual} at PCC is approximately zero at all

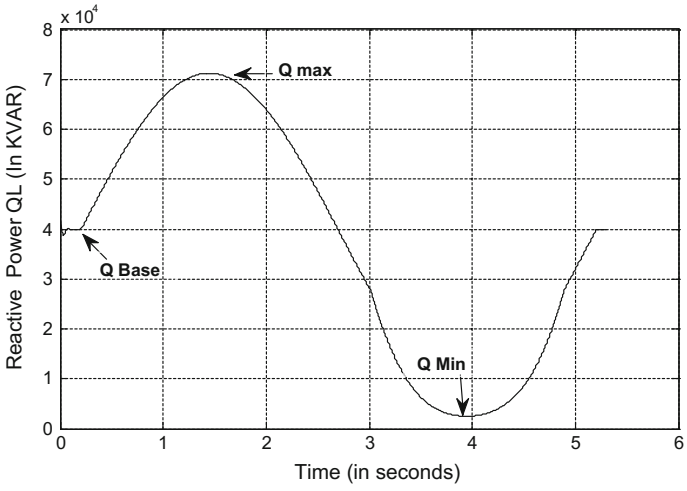


Fig. 22.4 Simulation of three phase dynamic load

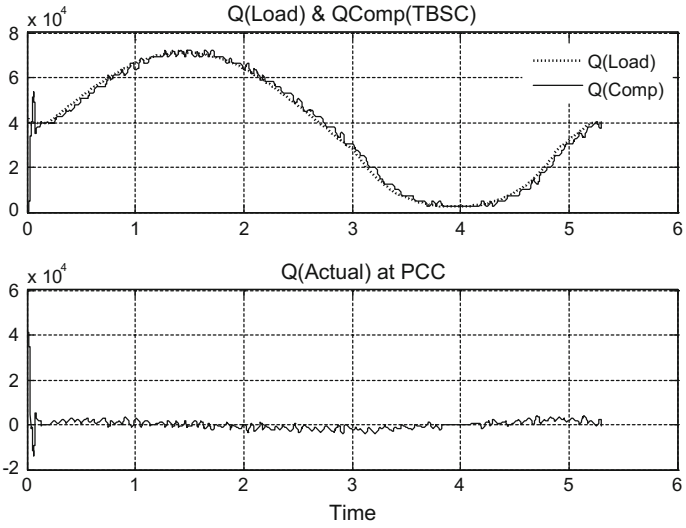


Fig. 22.5 Simulation result of TBSC operation

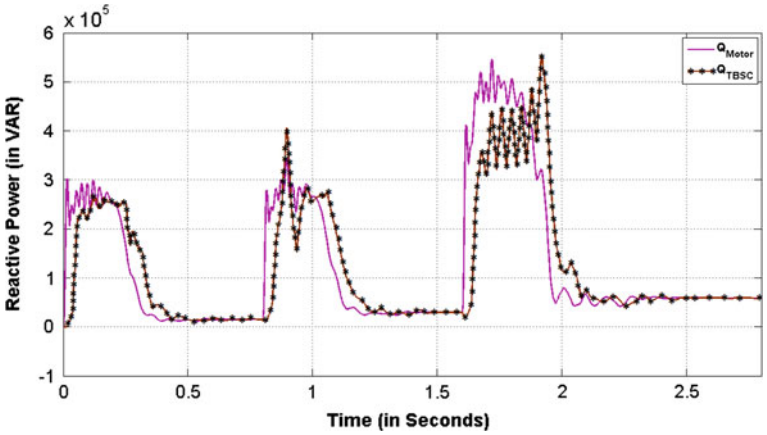


Fig. 22.6 Waveforms of Q_{Motor} and $_{QTBSC}$

times. Thus power factor is maintained near unity at all time. The small error is due to the binary switching arrangement of TSCs.

Figure 22.7 shows the comparison of motor line voltage with and without TBSC compensator Current waveforms through all TBSC banks and total compensating current (of R phase) are shown in Fig. 22.7 which are free from harmonics and have negligibly small transients only at few switching instants Table 22.1.

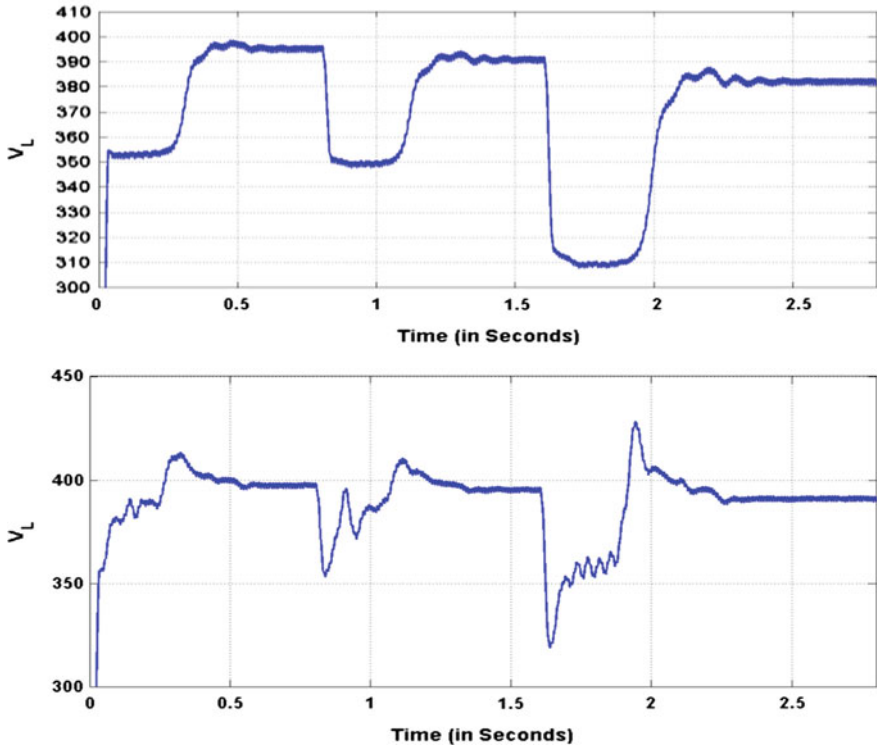


Fig. 22.7 Motor line voltage without TBSC compensator (*Top*) and with TBSC compensator (*Bottom*)

Table 22.1 Comparisons of results with and without TBSC compensator

Sr. No.	Parameter	Without TBSC Compensator			With TBSC Compensator		
		I.M.1 (50 h.p.)	I.M.2 (50 h.p.)	I.M.3 (100 h.p.)	I.M.1 (50 h.p.)	I.M.2 (50 h.p.)	I.M.3 (100 h.p.)
1	Switching instant (in sec)	0.0	0.8	1.6	0.0	0.8	1.6
2	% Voltage sag	11.14	11.64	21.77	2.01	5.3	7.92
3	Reactive power at starting (in KVAR)	250	250	380	Closely matches with the required value		
4	Starting current (in A)	500	500	1000	300	300	700

22.6 Hardware Results

The aggregate behavior of typical load at a substation/on a distribution transformer is somewhat unpredictable since the load changes in its nature, composition, and magnitude from time to time during the day.

A comprehensive microcontroller is designed and developed for analyzing the performance of TBSC Compensator. The TBSC consists of capacitor bank in three binary sequential steps being operated.

All the above components are fabricated, tested and implemented as a prototype hardware model, at 230 V, 50 Hz, 1phase AC supply. The load was increased from 0 Var to 700 Var. The details of the system performance with and without TBSC Compensator are given in the Table 22.2 and Table 22.3 respectively.

It is observed that without compensation p.f. varies from 0.334 to 0.343 while with compensation it is in between 0.634 to 0.9.

- *Comparison of experimental results*:

Graph shown below it shows the comparison of experimental results of Current, Reactive Power, Apparent Power and Power Factor.

Graph 1 shows the comparison between current without compensation and with compensation.

Graph 2 shows the comparison between Reactive Power without compensation and with compensation.

Graph 3 shows the comparison between Apparent Power without compensation and with compensation.

Graph 4 shows the comparison between current without compensation and with compensation.

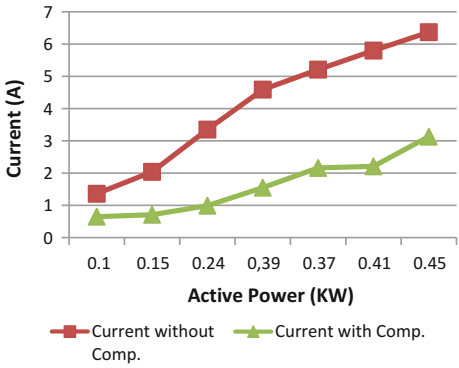
Table 22.2 System Performance without TBSC Compensator

Sr. no.	Voltage (V)	Current (A)	Active power (KW)	Reactive power (KVAR)	Apparent power (KVA)	Power factor (p.f)
1.	239.5	1.36	0.1	0.28	0.3	0.334
2.	238.8	2.04	0.15	0.41	0.44	0.339
3.	238	3.35	0.24	0.67	0.71	0.335
4.	237	4.59	0.39	0.91	0.96	0.333
5.	236.1	5.21	0.37	1.02	1.08	0.331
6.	235.7	5.8	0.41	1.12	1.21	0.340
7.	234.9	6.37	0.45	1.21	1.31	0.343

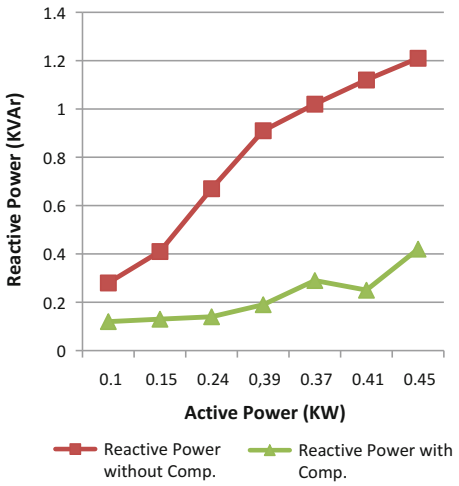
Table 22.3 System Performance with TBSC Compensator

Sr. no.	Voltage (V)	Current (A)	Active power (KW)	Reactive power (KVAR)	Apparent power (KVA)	Power factor (p. f)
1.	236.2	0.65	0.1	0.12	0.15	0.634
2.	235.9	0.709	0.14	0.13	0.18	0.773
3.	237.7	0.99	0.19	0.14	0.22	0.868
4.	235.1	1.55	0.28	0.19	0.34	0.828
5.	235.2	2.16	0.36	0.29	0.46	0.785
6.	234	2.21	0.39	0.25	0.47	0.839
7.	231.9	3.13	0.49	0.42	0.65	0.800

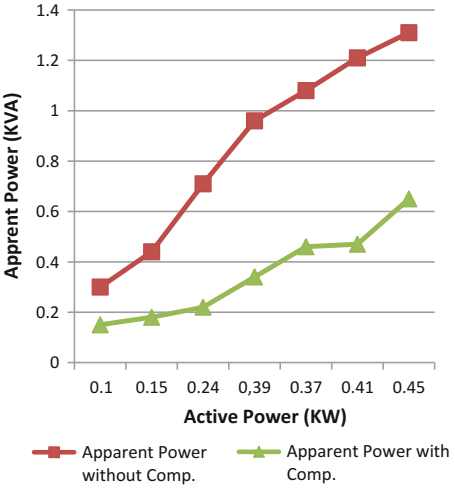
Graph 1 Reduction in load current



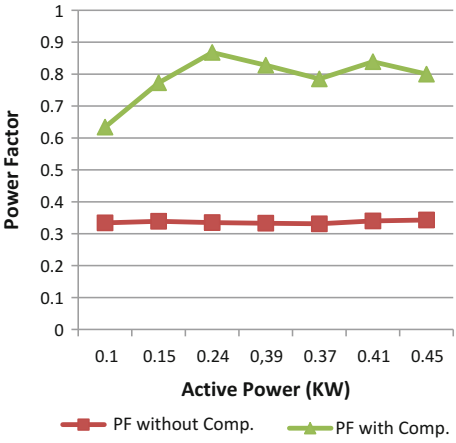
Graph 2 Reduction in load reactive power



Graph 3 Reduction in apparent power



Graph 4 Improvement in load power factor



22.7 Conclusion and Future Work

- *Simulation Conclusion based on TBSC Compensator for Fast Varying Dynamic Load*:

A topology using a TBSC has been presented. The TSC bank step values are chosen in binary sequence weights to make the resolution small. Current flowing through TBSC as well as source is transient free. Harmonic content in source current is negligibly small. By coordinating the control of TBSC, it is possible to obtain fully stepless control of reactive power. Also one can operate the system at any desired power factor. Proposed topology can compensate for rapid variation in reactive power on cycle to cycle basis. It gives following benefits:

- Maintaining the power factor near about unity (Graph 4).
- Minimum feeder current and loss reduction (Graph 3).
- Improvement in distribution feeder efficiency.
- Improvement in the voltage at load end.
- Relief in maximum demand (Graph 3) and effective utilization of transformer capacity.
- Conservation of energy takes place.
- *Conclusion based on TBSC Compensator for Induction Motor Load simulation*:

A topology for direct online starting of induction motors using TBSC compensator is presented. TSC bank step values are chosen in binary sequence weights to make the resolution small in order to achieve almost stepless reactive power compensation. Harmonic contents in source current are negligibly small. With the use of TBSC compensator; voltage sag magnitude gets reduced as well as voltage profile is improved. Controller operates in a closed loop to determine the number of capacitor units to be switched in the system. At the time of starting of I.M.s higher capacitor banks are switched in the system while once the motor reaches the rated speed only few lower capacitor banks will remain connected at the PCC. Thus at all times power factor is maintained near unity.

- *Future Scope*:
  - Fast varying dynamic reactive power compensation.
  - Voltage sag mitigation at the time of Induction Motor starting.
  - Where the farmer's use small rating and high rating (irrigation) HP motors.
  - Also it can be used for, at the time of long transmission and high rating of transformer charging.
  - Also it can be used for, to avoid resonance condition

Acknowledgements This work was supported by Department of Electrical Engineering, Annasaheb Dange College of Engineering and Technology, Ashta, Maharashtra, India.

References

1. Patil DR, IAENG M, Gudaru U (2008) Senior member IEEE, a comprehensive microcontroller for SVC Wit capacitor bank in binary sequential step minimizing TCR capacity. 978-1-4244-1762-9/08/$25.00 c2008 IEEE
2. Patil DR, Gudaru U (2012) The experimental studies of transient free digital SVC controller with thyristor binary compensator at 125 KVA distribution transformers. In: Lecture notes in engineering and computer science: proceedings of the world congress on engineering 2012, WCE 2012. London, U.K, pp 1053–1060, 4–6 July 2012
3. Patil DR, Gudaru U (2011) An innovative transient free adaptive SVC in stepless mode of control. Int Sci Index 5(5). http://www.waset.org/Publication/6880
4. Stout JH (1978) Capacitor starting of large motors—industry applications, In: IEEE Std 1159–1995, IEEE transactions on IEEE recommended practice for monitoring electric power quality, vol IA–14, Issue 3, May 1978

5. de silveira EP, Pires RC, de Almeida ATL, José A, Rezek J (2009) Direct on line starting induction motor with thyristor switched capacitor based voltage regulation. In: IEEE, pp 1124–1129
6. Irfan I, Swapnil M, Patil D, Gudaru U (2013) Senior Member IEEE, Patil DR Member, IAENG, A closed loop TBSC compensator for direct online starting of induction motors with voltage sag mitigation. In: Lecture notes in engineering and computer science: proceedings of the world congress on engineering and computer science 2013, WCECS 2013, vol I. San Francisco, USA, pp 271–279, 23–25 Oct 2013
7. Patil S, Shinde Y, Khushal S, Gudaru U (2013) Senior member IEEE, Patil DR member, IAENG, Transient free TBSC compensator for dynamic reactive load with closed loop control. In: Lecture notes in engineering and computer science: proceedings of the WCECS 2013, WCECS 2013, vol I. San Francisco, USA, pp 300–304, 23–25 Oct 2013
8. Mujawar I, Mujawar I, Swapnil DP, Patil DR (2014) Member, IAENG, U. Gudaru, senior member IEEE, TBSC-TCR compensator simulation: a new approach in closed loop reactive power compensation of dynamic loads. In: Lecture notes in engineering and computer science: proceedings IMECS 2014, vol II. Hong Kong, pp 721–727, 12–14 Mar 2014
9. Dixon Juan, Valle Yamilledel et al (2003) A full compensating system for general loads, based on a combination of thyristor binary compensator, and a PWM-IGBT active power filter. IEEE Trans Ind Electron 50(5):982–989
10. Dixon Juan, Morán Luis, Rodríguez José, Domke Ricardo (2005) Reactive power compensation technologies, state of–the-art review. Proc IEEE 93(12):2144–2164
11. Maffrand J, Dixon W, Morán L (1998) Binary controlled, static VAR compensator, based on electronically switched capacitors. In: Proceedings of the IEEE PESC'98. pp 1392–1396
12. Dixon JW, del Valle Y, Orchard M, Ortizar M, Moran L, Maffrand C (2003) A full compensating system for general loads based on a combination of thyristor binary compensator and a PWM-IGBT active power filter. IEEE Trans Ind Electron 50(5):982–989, Oct. 2003
13. Patil S, Kumbhar N, Mulla AM, Patil DR (2015) Binary current generation with transient free switching for TBSC Compensator. In: Lecture notes in engineering and computer science: proceedings of the world congress on engineering and computer science 2015. San Francisco, USA, pp 243–248, 21–23 Oct 2015

Chapter 23
Sensorless Speed Estimation by Using Rotor Slot Harmonics for Squirrel Cage Induction Motor

Hayri Arabaci

23.1 Introduction

The induction motors are widely used in industry because of inexpensive cost and their robustness. The motors drive various vital components and loads in industrial processes. Condition monitoring for both inductions motor and combined systems has become necessary to avoid unexpected failure. Fault detection algorithms dependent on the current spectrum analysis and the detection of some harmonic components that occur according to motor speed [1–3]. In such algorithms, accuracy of speed knowledge is critical for the effective condition monitoring and fault diagnosis. Moreover, motor speed and rotor position require a feedback for various applications, but most of these systems do not have any speed sensor in industrial setups. Speed sensors such as encoders, tacho-generator and hall effects sensors are connected to the system to obtain the speed value. Majority of these sensors generate square wave signal. Speed estimation from square wave signal is typically based on several methods [4–6]:

I Measurement of a time interval between successive pulses,
II Counting of pulses during the prescribed time,
III Measurement of time duration for the variable number of pulses.

In addition to these, spectral analysis methods are used for speed estimation by square wave signal of speed sensor [4].

In these all method, the speed sensors need cabling to connect to motor. It makes the cost of the motor system increased. In addition to this, faulting probability of the speed sensors is higher then induction motor. Robustness of whole system decreases. The sensorless speed estimation is a viable alternative to avoid the

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

S.-I. Ao et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-981-10-2717-8_23

problems which associates with the system including speed sensor. Many approaches have been done to obtain the speed from electrical quantities of motor during recent years. Various motor speed estimation methods have been presented. They can be summarized in two main sections as follows [7]:

- Methods are based on motor mathematical models to deduce observers and adaptive schemes,
- Methods are based on spectral component estimation of the voltage or current.

The methods using motor mathematical models have advantage of short processing time. But, they have disadvantages depending on many time variations and requiring a lot of knowledge of motor parameters. Otherwise, current and voltage sensing are required for these methods [8]. The second mentioned methods need only voltage or only current measurement. There are also independent of electromagnetic motor parameters. But, these methods cause longer processing times according to first mentioned methods.

The current and voltage measurement equipments are generally present to protect from over voltage and over current in most industrial systems including motor. The voltage and current data could be obtained by adding some hardware. Using of voltage measurements has some problems such as distorted waveform (especially system with induction motor which is supplied by inverter). So, the algorithms, which is based on voltage spectral analysis, have not been encouraged this problem. Motor current have been preferred due to the filtering behavior of the motor stator winding. Because the speed observers of a squirrel cage induction motor are available in the stator currents, the speed could be obtained by analysis of frequency spectrum of the current [9]. Harmonic components in the spectrum are related to the number of rotor slots and named Rotor Slot Harmonics (RSH). Resolution of the frequency spectrum effects directly speed estimation accuracy. The high resolution improves accuracy, but it increases the number of computation. Different spectral analysis methods are compared in detail in [10, 11] (such as Chirp-Z Transform (CZT), Fast Fourier Transform (FFT), Hilbert Transform, and Interpolated FFT). The CZT gives good spectral resolution, but it increases the number of operations. Adequate accuracy level is obtained by using FFT. Moreover, the number of operations is lesser than using CZT. So, FFT is generally used to obtain the frequency spectrum of the current.

In the literature, authors have mainly focused on the optimization of search algorithms of the rotor slot spectral components [11–14]. Various studies are available such as fast orthogonal search algorithm, which's result data have lower resolution current spectrum, could be used to identify RSH [15]. The algorithm allows estimation of the speed for real-time performance by an embedded digital signal processor.

A scheme, which combines Hilbert transform and interpolated FFT, is used to improve the estimation accuracy of the speed [16]. Hence, the algorithm realizes direct motor torque control and soft-starting process control without any tachometer.

Eventually, most of studies have been focused on methods that are used to transform time domain to frequency domain. Determination of RSH components

and estimation of speed according to RSH are made by individual approaches in their studies. Some studies using Artificial Neural Network (ANN) [17–19] and genetic algorithm (GA) [20–22] are available in the literature. These studies use methods which are based on motor mathematical model. Otherwise the authors in [23] have shown the benefit of use in the field of GA.

An ANN approach is presented for sensorless speed estimation using RSH in [24]. To calculate the motor speed, each related RSH component is extracted from the spectrum by using ANN. 1.5 rpm average calculation error is reached.

In this paper, the motor current spectrum is obtained by using FFT due to a smaller number of operations and adequate accuracy level. RSH in the spectrum are used to estimate motor speed by using GA.

23.2 Speed Calculation Method Using Square Wave Speed Sensor

Rotary encoders measure motor speed using incoming pulses from an encoder mounted to motor shaft. They have been used in speed control of electrical motors. The accuracy of speed measurement is an essential requirement to reach a good dynamic response which can cause stability problems. Speed measurements are typically based on several methods [4–6]:

I Measurement of a time interval between successive pulses,
II Counting of pulses during the prescribed time,
III Measurement of time duration for the variable number of pulses.

The speed measurement systems are based on two methods; encoder pulses frequency and period measurement. In Method I, motor speed in rpm unit is calculated according to Eq. (23.1).

$$n_r = \frac{60}{T_{mes} \cdot N_{noc}} \tag{23.1}$$

where N_{noc} is the number of encoder notches, T_{mes} is measured time between sequential two signals of encoder. The main disadvantages of Method I are that value of T_{mes} is small in high speed and measurement time is long in low speed. So a quantization error is occurred in the calculated speed value in that area. Motor speed in rpm unit is calculated from Eq. (23.2) for Method II.

$$n_r = \frac{60}{T_{pr} \cdot N_{noc}} \cdot N_{con} \tag{23.2}$$

where N_{con} is the number of counted pulses from encoder, T_{pr} is significant fixed measurement time. In this method because $T_{pr} \cdot N_{noc}$ determines quantization

error of calculated speed values, definition of T_{pr} is very important. In Method III, motor speed in rpm unit is calculated according to Eq. (23.3).

$$n_r = \frac{60}{T_{mes} \cdot N_{noc}} \cdot N_T \tag{23.3}$$

where N_T is the number of fixed pulses of encoder. Weakness of Method I is resolved by this method in high speed. But measurement time is still long in low speed. Therefore, this method is suitable for high speed and the period measurement method is suitable for low and medium speed.

To attenuate the above speed calculation methods-dependency, an advanced speed estimation technique has been used for speed calculation [4]. In the method, sensor signal is sampled at high frequency. Windowing in particular length is performed to the sampled data. Each window data is transformed to frequency domain by using a signal processing technique as FFT. Resolution of speed-time curve can be increased by Welch method. The maximum frequency value corresponds to maximum amplitude in the spectrum is used to estimate the speed according to Eq. (23.4).

$$n_r = \frac{f_{s\max}}{S_{nn}} \cdot 60 \tag{23.4}$$

where $f_{s\max}$ is maximum frequency values, S_{nn} is the number of notch of sensor and n_r is unit of speed in rpm.

Figure 23.1 shows signal of speed sensor, the using windows, window size (M), hop size (s) and some calculated speed values.

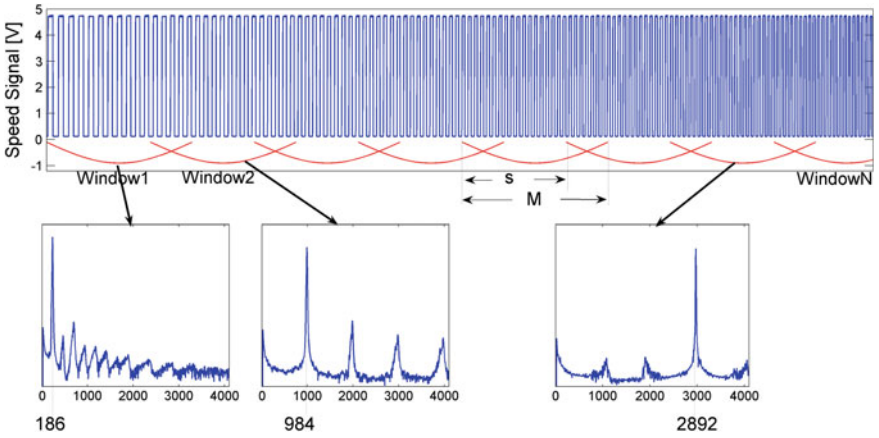


Fig. 23.1 Speed sensor signal and some maximum frequency values corresponding to maximum amplitude in the spectrum

23.3 Sensorless Speed Estimation Method

Various sensorless speed estimation techniques have been received with increasing attention for squirrel cage induction motor. These techniques can be divided two main parts. One of them is model based approaches. But, most of these approaches are sensitive to motor parameters and load type. Accuracy of these parameters directly affects performance of these methods. So, the accuracy of speed estimation is limited for the model-based approaches. In the second part is based on the spectrum analysis of stator current. The stator current harmonics are included RSH which are proportional to motor speed [8]. RSH could be used to estimate the motor speed depending on the number of rotor slots [1]. Moreover, an accurate model of the motor parameters and the load type information are not required in these methods. In this paper, the speed estimation is made by using RSH in frequency spectrum of motor current. Only this method is mentioned in this section because the proposed algorithm bases on RSH analysis of the current.

A. The speed estimation by using rotor slot harmonics

The stator current data of squirrel cage induction motor has speed observers. The motor speed can be estimated by mean of spectrum analysis of the current. The speed estimation method does not depend on any motor parameters. The method requires only the knowledge of the number of rotor slots and the number of motor pole pairs. The rotor slots produce changes on air-gap flux with a spatial distribution dependent on the number of rotor slots. These changes are rotor slot mmf harmonics and these effect the stator current in continuous variations. Frequency of RSH components is calculated by using Eq. (23.5) [1].

$$f_{sh} = f_1 \left[(k \cdot R \pm n_d) \frac{1-s}{p} \pm v \right] \tag{23.5}$$

where: $n_d = 0, \pm 1, \pm 2, \ldots$ $k = 0, 1, 2, \ldots$ $v = 0, \pm 1, \pm 3, \ldots$

f_1 is fundamental frequency component of stator current. p is the number of pole pairs of motor. R is the number of rotor slots (bars). s is per unit motor slip. v is the order of stator time harmonics presenting in the stator current. nd is known as eccentricity order and it's value is zero for healthy motor. When $n_d = 0$ and $k = 1$ values are replaced in Eq. (23.5), principal slot harmonic is obtained as in Eq. (23.6).

$$f_{psh} = f_1 \left[R \frac{1-s}{p} \pm v \right] \tag{23.6}$$

The principal slot harmonic components corresponding to v values are calculated by using s and Eq. (23.6). Then motor speed is obtained in rpm by Eq. (23.7).

$$n = \frac{60 \cdot f_1}{p}(1 - s) \tag{23.7}$$

The values of p can be calculated from nameplate information of motor. The fundamental frequency component f_1 is easily determined from the frequency spectrum. It is need that R value is informed from manufacturer of the motor. But R is always integer and motor speed is approximately known under no-load condition. So, this R values could be obtained by no-load motor test [9].

The main problem is to determine that the concerned RSH component in Ed. (23.6) occurs due to which v value. Furthermore, RSH component for whole values of v is not available in the spectrum. In addition to this, amplitude of certain RSH component may be very small. To estimate the motor speed from RSH is difficult because of above mentioned reasons. In the literature individual algorithms have been used for speed estimation [7, 9, 11, 16]. In this paper, the proposed method uses GA to extract the v value and slip knowledge by using one of the RSH components in scanning region of current spectrum.

B. The proposed sensorless speed estimation method

The proposed method uses frequency spectrum of stator current. The frequency spectrum is obtained by using FFT. Many frequency components are available in the spectrum. Squirrel cage induction motors under steady state operation run in the slip region that is between the slip corresponding to maximum torque and synchronous speed. The slip corresponding to maximum torque depends on the structure of the motor. It is generally below the 10 %. Therefore, search region of the frequency spectrum is determined by using Eq. (23.6) for s values between 0 and 10 %. The determined search region is shown in Fig. 23.2.

RSH components are searched in the marked region. But, there are many components which are amplitude are at small level. These components are eliminated by determining a threshold value. It reduces the number of operations at searching process of probable RSH. Any operation is not done for fundamental

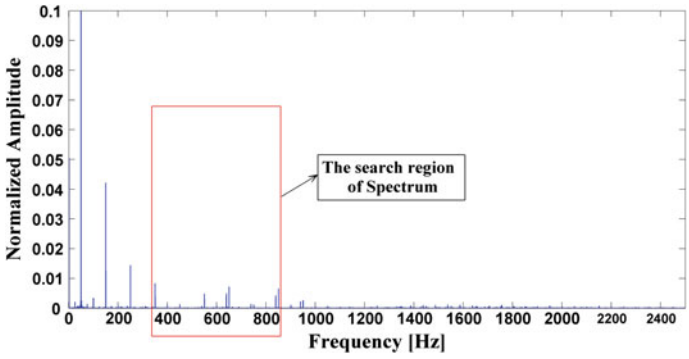


Fig. 23.2 The determined region to scan the RSH

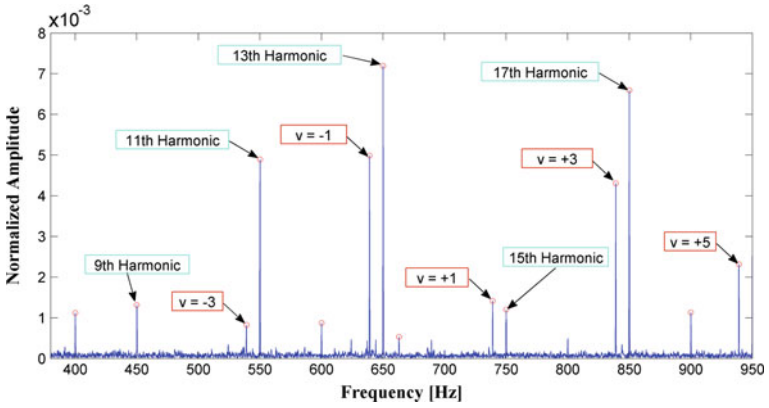


Fig. 23.3 The frequency components of scanned region for 1476.4 rpm at 50 Hz

harmonic components in this region because of easy determination. RSH components, fundamental harmonic components are shown in Fig. 23.3.

Because motor speed is known from experimental study, *v* values are directly shown in the spectrum for every RSH components in Fig. 23.3. However, in the practice motor speed is unknown and wants to find out. So, it is not certain that RSH component corresponds to which v value. For overcoming of this problem GA is used in this paper. Frequency value of any RSH component in the region is used in GA process. End of the GA process, s and v values are found out. So, motor speed is calculated by using the obtained s value and Eq. (23.7). The block diagram of the proposed method is given in Fig. 23.4.

C. GA approach in sensorless speed estimation

The methods, which are based on RSH for sensorless speed estimation, generally use similar process stages. Initially, motor current is sampled from one stator coil. Then frequency spectrum of the current is obtained by using time-frequency transformation. RSH components are determined from the frequency spectrum. Motor speed is calculated via the RSH, v value, Eqs. (23.6) and (23.7). But it needs to certainly know that the selected RSH frequency is for which the v value. This is main problem of the method. To date, individual algorithms have been used to calculate the motor speed from RSH components by Eq. (23.6). In this paper, GA is used to determine v and s values corresponding to related RSH component.

Processing steps of the used GA algorithm are given as follows:

1 Adapt the structure of chromosome according to constraints and limits of parameters
2 Form the objective function
3 Generate the initial population
4 Apply crossover, mutation, selection
5 Calculate the fitness and penalty values

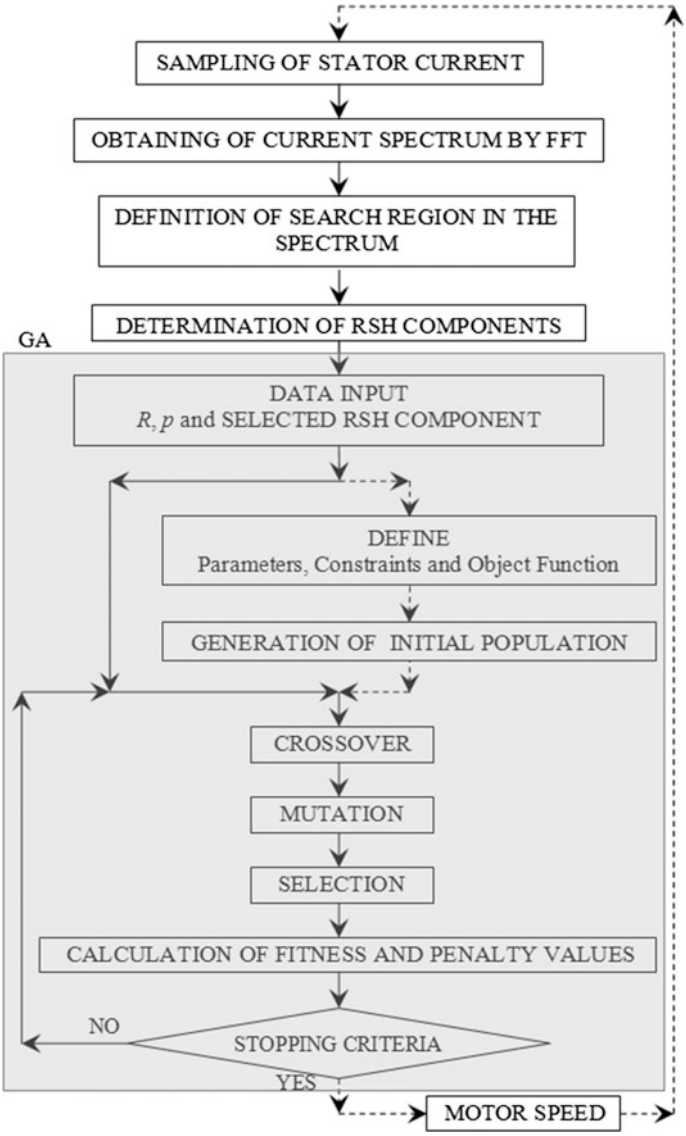


Fig. 23.4 The block diagram of the proposed method

6 Convergence and repeat of the test from step 4 until achieving the optimum solution.

In this study, aim is to find the value of v, and thus to estimate the motor speed (or s). Therefore, each chromosome consists of v and s values. Chromosome is formed to be binary code. So, the binary-code GA is used in the proposed algorithm.

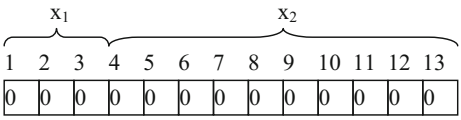


Fig. 23.5 The structure of chromosome

Values of v are odd number. Its minimum value and maximum value are −7 and 7 respectively. s is real number. Its minimum value and maximum value are 0 and 0.1 respectively. 0.0001 is enough for resolution of s. Hence, length of each chromosome is 13 bits (13 genes). First 3 bits (x1) represent v value. Other 10 bits (x2) represent s value. The structure of this chromosome is given in Fig. 23.5.

a Initial population

The initial population is generated randomly without any limit, because of definition of limit is done at forming process of chromosome.

b Constraints

Equation (23.6) is used to form the object function. Values of R and p parameter are known. But, f1 is calculated from frequency spectrum of current. So, the constraint function is obtained as shown in Eq. (23.8).

$$g(x) = f_{psh} - f_1\left[R\frac{1-x_2}{p} \pm x_1\right] \tag{23.8}$$

where f_{psh} is value of RSH component which is selected from the scanning region of spectrum. x_1 is first 3 bits of jth x chromosome of every generation and is calculated as to Eq. (23.9).

$$x_1 = 2 \cdot (x(1:3))_{10} - v_{\min} \tag{23.9}$$

where $v_{\min}$ is minimum value of the v and determined to be −7 for study. x_2 represents motor slip and it's values is calculated by last 10 bits of jth x chromosome. Value of the x_2 is calculated by using Eq. (23.10).

$$x_2 = \frac{s_{\max} \cdot (x(4:13))_{10}}{2^{10} - 1} - s_{\min} \tag{23.10}$$

where $s_{\max}$ is acceptable maximum value of the motor slip under steady state condition and approved to be 0.1 for this study. $s_{\min}$ is also determined to be 0. Penalty values are calculated for each generation. The penalty values are used for selection.

c Selection

The obtained penalty values by Eq. (23.8) are compared other individuals of related generation. The best chromosome is saved and carried next generation.

d Crossover

Crossover is an operator used in order to produce better chromosomes and increase the diversity of the population. One-point is used for crossover is applied on the population. The individual populations are chosen and paired randomly. The selected pair of genes is exchanged from one point which is determined randomly.

e Mutation

Mutation is applied on the population to avoid local optimum [25]. So the diversity is increased in the population. A number between 0 and 1 is produced randomly for mutation. This number is compared the mutation rate. If the number is under the mutation rate, mutation is applied one gene of the related chromosome. The gene is selected randomly. If the selected gene is 0, it is made 1, else 0.

f Convergence test

New individuals are achieved every loop of the procedure. This process continues until reaching a predetermined generation number or meeting an object. The most important point is to determine the stopping criterion. In the study, the generation number is chosen as stopping criterion.

23.4 Experimental Study and Results

Squirrel cage induction motor with 28 bars was used to carry out the experiments study. Its specifications were 2.2 kW, 3 phases, 2 pole pairs, 380 V and 50 Hz. The experiments were made in the laboratory by using an experimental system. The motor was loaded by generator. Loading of the motor is leveled by using resistors which are conducted to the generator. The block diagram of the used experimental system is given in Fig. 23.5. The photograph of experimental system and its components are also given in Fig. 23.6. One phase current of the motor was sampled in steady state operation for various speed levels. Verification of estimated results with real speed values is made by using an encoder combining to motor-generator shaft in the experimental system Fig. 23.7.

The motor current was sampled from Hall-effect current sensor by using data acquisition card. The sampling rate was 5 kHz. FFT with Hanning window was used to obtain frequency spectrum of this current data. RSH components were determined from the scanning region. One of these determined RSH components was used in objection function of GA. Then GA iteration was started. End of the iteration, the calculated s value was used to estimate the motor speed via Eq. (23.7).

The estimated motor speeds and real speeds are compared and given in Table 23.1. Maximum error rate in rpm is 0.064 % (0.96 rpm for 50 Hz). Finally the average error is calculated as 0.29 rpm.

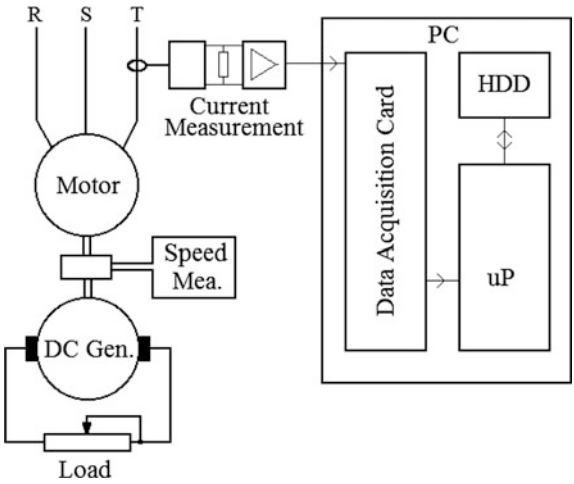


Fig. 23.6 The block diagram of experiment system

Fig. 23.7 The photograph of experiment system and components

Table 23.1 Comparison of experimental results and real speed values for different load levels

Load levels	Estimated speed values [rpm]	Real speed values [rpm]	Estimation error [rpm]	Estimation error [%]
Load 1	1468.4	1468.91	0.51	−0.034
Load 2	1468.6	1468.77	0.17	−0.011
Load 3	1468.9	1468.77	0.13	0.009
Load 4	1476.3	1475.95	0.35	0.023
Load 5	1476.4	1476.54	0.14	−0.009

(continued)

Table 23.1 (continued)

Load levels	Estimated speed values [rpm]	Real speed values [rpm]	Estimation error [rpm]	Estimation error [%]
Load 6	1476.6	1476.69	0.09	−0.006
Load 7	1485.8	1485.63	0.17	0.011
Load 8	1486	1485.63	0.37	0.025
Load 9	1486.1	1486.07	0.03	0.002
Load 10	1488.7	1489	0.3	−0.020
Load 11	1488.8	1489	0.2	−0.013
Load 12	1489	1489	0	0.000
Load 13	1492	1492.96	0.96	−0.064
Load 14	1492.2	1492.96	0.76	−0.051
Load 15	1492.4	1492.96	0.56	−0.037
Load 16	1497.3	1497.51	0.21	−0.014
Load 17	1497.4	1497.51	0.11	−0.007

23.5 Conclusion and Future Work

Motor current is main component in the sensorless speed estimation. RSH are required to estimate the speed. Frequency spectrum of motor current is obtained by one of the signal processing method as FFT. The RSH are available in the current spectrum. RSH components are estimated by Eq. (23.5). The v values in this equation are required to calculate the motor speed. But, it is not clear that the one the RSH in the spectrum represents which v value. In this work, GA is used for discriminate this couple problem. Experimental study was carried out to prove accuracy of the proposed method. The motor was loaded for various load levels in the experiments to investigate different values of motor speed. Calculation error varies for every load levels. The maximum calculation error is 0.96 rpm (0.06 %). The average calculation error is 0.29 rpm (0.02 %). Therefore, the experimental results show that the proposed method is feasible to estimate the speed in high accuracy. Furthermore, this proposed algorithm can enhance the performance and reliability of monitoring of squirrel cage induction motor without any speed sensor as in submersible induction motor applications.

In the future work, the advanced signal processing techniques, such as Estimation of Signal Parameters via Rotational Invariance Technique (ESPRIT) [14, 15] and Multiple Signal Classification (MUSIC), can be used to obtain the current frequency spectrum. These techniques have its higher frequency-resolution compared to FFT. But they have long computation time.

Acknowledgement The study has been supported by Scientific Research Project of Selcuk University.

References

1. Vas P (1993) Parameter estimation, condition monitoring and diagnosis of electrical machines. Clarendron Press, Oxford
2. Benbouzid MEH, Klimam GB (2003) What stator current processing-based technique to use for induction motor rotor faults diagnosis? IEEE Trans Energy Convers 18(2):238–244
3. Benbouzid MEH (1998) A review of induction motors signature analysis as a medium for faults detection. In: IECON '98. Proceedings of the 24th annual conference of the IEEE, vol 4, pp 1950–1955
4. Arabaci H, Bilgin O (2012) A novel motor speed calculation method using square wave speed sensor signals via fast fourier transform. Turkish J Electr Eng Comput Sci 20(1):1090–1099
5. Prokin M (1994) Extremely wide-range speed measurement using a double-buffered method. IEEE Trans Ind Electron 41(5):550–559
6. Prokin M (1991) Speed measurement using the improved DMA transfer method. IEEE Trans Ind Electron 38(6):476–483
7. Aiello M, Cataliotti A (2005) An induction motor speed measurement method based on current harmonic analysis with the chirp-Z transform. IEEE Trans Inst Meas 54(5):1811–1819
8. Rajashekara K, Kamamura A, Matsure K (1996) Sensorless control of AC motor drives, Piscataway. IEEE Press, NJ
9. Hurst KD, Habetler TG (1996) Sensorless speed measurement using current harmonic spectral estimation in induction machine drives. IEEE Trans Power Electron 11(1):66–73
10. Aiello M, Cataliotti A, Nuccio S (2001) A comparison of spectrum estimation techniques for periodic and nonstationary signals. In: IEEE instrumentation and measurement technology conference, Budapest, Hungary, pp. 1130–1134
11. Hurst KD, Habetler TG (1997) A comparison of spectrum estimation techniques for sensorless speed detection in induction machines. IEEE Trans Ind Appl 33:898–905
12. Ishida M, Iwata K (1984) A new slip frequency detection of induction motor utilizing rotor slot harmonics. IEEE Trans Ind Appl 20:575–581
13. Ishida M, Iwata K (1987) Steady state characteristics of torque and speed control system of an induction motor utilizing rotor slot harmonics for slip frequency sensing. IEEE Trans Power Electron 2:257–263
14. Blasco R, Sumner M, Asher GM (1994) Speed measurement of inverter fed induction motors using the FFT and the rotor slot harmonics. Power Electron Variable-Speed Drives, Conf Publ 399:470–475
15. McGaughey DR, Tarbouchi M, Nutt K, Chikhani A (2006) Speed sensorless estimation of AC induction motors using the fast orthogonal search algorithm. IEEE Trans Energy Convers 21(1):112–120
16. Shi D, Unsworth PJ, Gao RX (2006) Sensorless speed measurement of induction motor using Hilbert transform and interpolated fast Fourier transform. IEEE Trans Instrum Meas 55 (1):290–300
17. Beliczynski B, Grzesiak L (2002) Induction motor speed estimation: neural versus phenomenological model approach. Neurocomputing 43(1-4):17–36
18. Bharadwaj RM, Parlos AG (2003) Neural state filtering for adaptive induction motor speed estimation. Mech Syst Sign Proces 17(5):903–924
19. Toqeer RS, Bayindir NS (2003) Speed estimation of an induction motor using Elman neural network. Neurocomputing 55(3-4):727–730
20. Yan Z, Zhang Y, Zhan X (2008) Research of speed estimator based on wavelet neural network adjusted by ant colony optimization. In: IEEE international conference on mechatronics and automation, Kagawa, Japan, pp 398–403
21. Datta M, Rafid A, Glosh BC (2007) Genetic algorithm based fast speed response induction motor drive without speed encoder. In: IEEE international conference on power engineering, energy and electrical drives, Setubal, Portugal, pp. 146–151

22. Çunkaş M, Sağ T (2010) Efficiency determination of induction motors using multi-objective evolutionary algorithms. Adv Eng Softw 41(2):255–261
23. Arabaci H (2015) A genetic algorithm approach for sensorless speed estimation by using rotor slot harmonics. In: Lecture notes in engineering and computer science: proceedings of the world congress on engineering and computer science 2015, WCECS 2015, San Francisco, USA, pp 265–269, 21–23 Oct 2015
24. Arabaci H (2014) An artificial neural network approach for sensorless speed estimation via rotor slot harmonics. Turkish J Electr Eng Comput Sci 22:1076–1084
25. Martínez M, García-Nieto S, Sanchis J, Blasco X (2009) Genetic algorithms optimization for normalized normal constraint method under Pareto construction. Adv Eng Softw 40(4): 260–267

Chapter 24
Selective Harmonic Elimination Technique Based on Genetic Algorithm for Multilevel Inverters

Hulusi Karaca and Enes Bektas

24.1 Introduction

In recent years, power converters applications in industry have increased due to the developments in semiconductor technology and the different pulse width modulation techniques [1]. Especially, in high power and medium voltage applications, multilevel inverter topology has come up with a choice because power switches which enable to withstand the medium voltage have not been developed yet [2].

Multilevel inverters generating output voltage in the form of stepped wave can be categorized as Cascaded H-Bridge multilevel inverter, Diode-clamped multilevel inverter (DCMLI) and Capacitor-clamped multilevel inverter (CCMLI) structure. First structure presented in 1975 is H-Bridge topology including the series connection of H-bridge. After that, the structures of the diode clamped multilevel inverter and the capacitor clamped multilevel inverter have appeared, respectively [1–3]. In Fig. 24.1, conventional multilevel inverter structures with five-level have been given.

All of these structures have same number of power switches for each level. However, other components such as diodes and capacitors are needed additionally in diode-clamped and capacitor-clamped structures respectively as seen in Fig. 24.1. Using of extra capacitor and diodes make the system hardware bulky and may cause to unbalanced voltage problem. Especially with rising level the number of clamped diodes and clamped capacitors rises exponentially [4, 5]. In Fig. 24.2,

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

E. Bektas
Engineering Faculty, Department of Electrical and Electronics Engineering,
Cankiri Karatekin University, 18000 Cankiri, Turkey
e-mail: enesbektas@karatekin.edu.tr

S.-I. Ao et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-981-10-2717-8_24

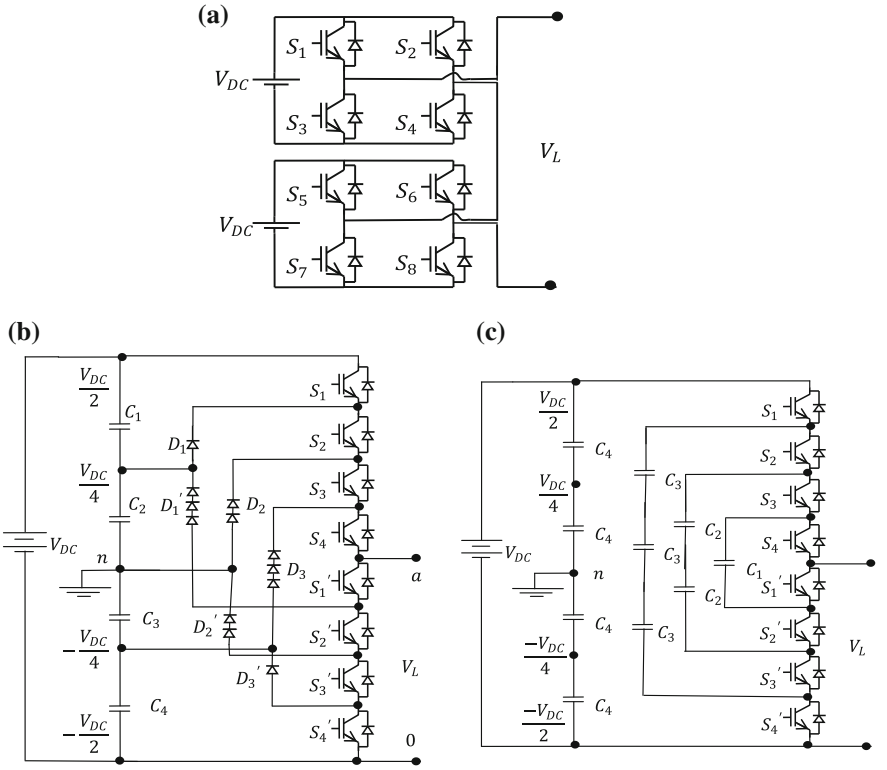


Fig. 24.1 Conventional multilevel inverter structures **a** Cascaded H-bridge topology, **b** Diode clamped topology and **c** Capacitor clamped topology

the number of each required components have been given for all conventional multilevel inverter structures.

As seen in Fig. 24.2. Cascaded H-bridge topology is more practical to implement because of lower required component number when compared to CCMLI and DCMLI.

Multilevel inverter has outstanding performances such as lower *dv/dt* for good electromagnetic compatibility, reduced voltage stress on the switches, input current with low distortion, more importantly having lower total harmonic distortion than traditional two-level inverter in output voltage and currents [2, 4, 6]. By applying the different techniques aiming at the elimination of harmonics, the THD of multilevel inverter's output voltage can be decreased. One of the most important techniques is Selective Harmonic Elimination (SHE) method that can eliminate the desired any harmonics. At the first stage, the nonlinear equations of output voltage harmonics including trigonometric terms must be obtained. In the second stage, which is the basic purpose of SHE technique, switching angles can be acquired with

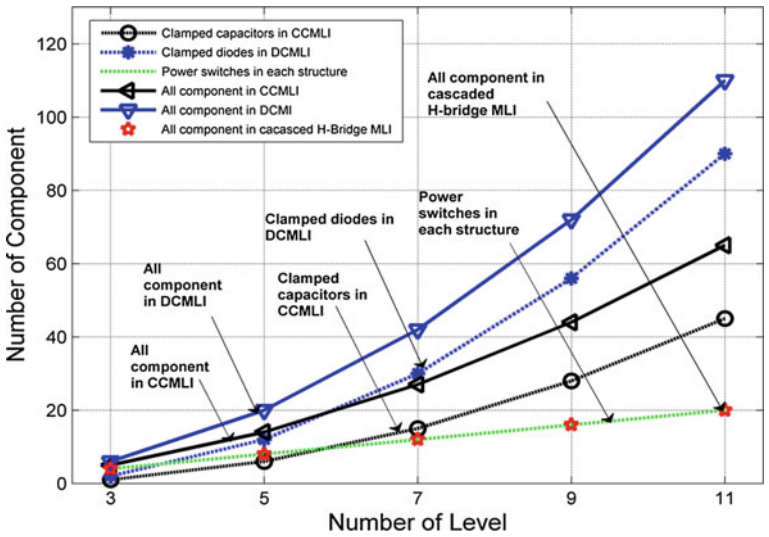


Fig. 24.2 Number of components for all conventional multilevel inverter structures

solving the harmonic equations. These equations change according to the number of harmonics' order desired to eliminate and the requested output voltage value [6, 7].

The usual methods to obtain switching angles are iterative methods such as Newton-Raphson [8] and Neural Networks using Newton-Raphson method. The iterative methods are sensible to the initial value and divergence problems are probably to arise, also the optimal switching angles may not be produced for the minimum THD [9]. These methods provide more complex solution than Genetic Algorithm due to the necessity of different look-up tables for each modulation indexes. The GA is a type of artificial intelligence approaches. The GA gives the optimum solution of the harmonic equations in contrast to the iterative methods [10, 11].

Multilevel inverters have several disadvantages. The most significant drawback is the necessary of great number of switches especially in high level number as stated in Fig. 24.2. In addition, the drivers of each level module must be isolated from each other. As THD decreases with rising level, the hardware of multilevel inverter needs more isolated power source. That leads the increasing expense of inverter according to conventional inverter. Thus, to realize the lower cost applications, it is very important to reduce the number of power semiconductor switches and drivers [2, 11]. In this paper, multilevel inverter with reduced number of switches using the isolated DC sources has been proposed. The proposed topology resembles cascaded H-bridge multilevel inverter structure. While cascaded H-bridge multilevel inverter have H-bridge modules, the proposed multilevel inverter have Half-bridge modules. SHE equations of the stepped output voltage have been solved by genetic algorithm software which eliminates desired harmonics

order without the use of GA toolbox of MATLAB. The results of simulation have clearly proved that the proposed SHE technique for multilevel inverter with reduced number of switches can eliminate the desired harmonics' order [1].

24.2 Proposed Multilevel Inverter with Reduced Number of Switches

The proposed cascaded multilevel inverter consists of level module units having half-bridge topology with two switches. Each half-bridge is connected isolated DC power supply as shown in Fig. 24.3. The overall structure of suggested topology includes two basic parts. The first part is the level module units, which produce DC voltage levels. The second is the H-bridge generating positive and negative stepped output voltage.

Regarding isolated DC voltage sources' number is k, also the number of level-modules, the maximum and minimum rates of level's output voltage are

$$V_{o\_min} = 0 \tag{24.1}$$

$$V_{o\_max} = kV_{dc} \tag{24.2}$$

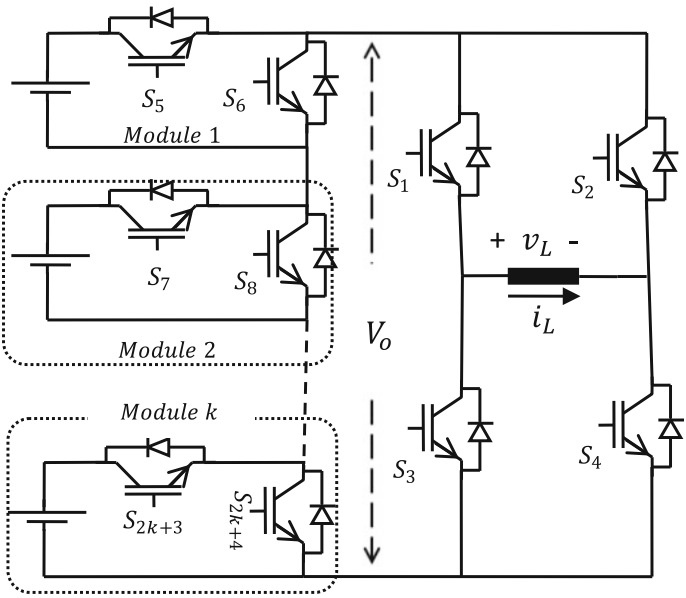


Fig. 24.3 The proposed structure of multilevel inverter with reduced number of switches

The is known as symmetric multilevel if all voltage sources equal to V_{dc} [2]. The number of output voltage levels in symmetric multilevel inverter is

$$N_{level} = 2k + 1 \tag{24.3}$$

Maximum and minimum rates of load voltage are

$$V_{L\_min} = -kV_{dc} \tag{24.4}$$

$$V_{L\_max} = +kV_{dc} \tag{24.5}$$

The number of switches is *4 k* in H–bridge inverter topology. But in the suggested topology, this value is *2 k + 4*. When compared usual cascaded multilevel inverter and suggested topology in terms of switches' number, suggested topology is more advantageous in multilevel inverter with more than level number 5 (7, 9, …).

In these types of inverter, in order to get sinusoidal output voltage, switching angles have been calculated analytically by (24.6) and (24.7) [2]. This technique aiming sinusoidal output voltage is named as SPWM (Sinusoidal Pulse Width Modulation).

$$N_\alpha = k = \frac{N_{level} - 1}{2} \tag{24.6}$$

$$\alpha_i = \sin^{-1}\left(\frac{2i-1}{N_{level}-1}\right) \quad i = 1, 2, 3, \ldots .k-1, k \tag{24.7}$$

where, N_α is the number of switching angles. In five-level inverter, for example, two switching angles with the condition.$\alpha_1 < \alpha_2 < \pi/2$ have been calculated. It is quite apparent that other angles can be obtained from these angles because of the quarter wave symmetry as shown in Fig. 24.4. S_5 and S_7 is the passkey of the five-level inverter. Therefore, α_1 and α_2 is calculated for these switches respectively. Switches' state, in the same module, cannot be simultaneously due to the fact that to elude short-circuit.

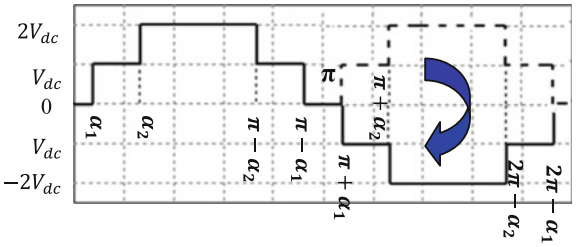


Fig. 24.4 Output voltage waveform of 5-level inverter

In Fig. 24.4, the switching angles can be seen for one cycle of output voltage wave form. Dashed line in the figure represents the voltage (shown in Fig. 24.1. as V_o) applied to full-bridge and it's frequency is twofold of output voltage [1].

24.3 Calculation of Switching Angles Using Shepwm Technique

The control of multilevel inverter is based on specifying the switching angles to synthesize a desired sinusoidal voltage waveform. In this paper, the stepped voltage waveform with odd-harmonic components is given in (24.8) for *2 k + 1* level inverter. Because of symmetric wave, output voltage has not included even-harmonic components.

For equal dc sources, Fourier series expansion of the stepped output voltage is given as,

$$V_o(\omega t) = \sum_{n=1,2,\ldots}^{\infty} \frac{4V_{dc}}{\pi n}\left[\cos(n\alpha_1) + \cos(n\alpha_2) + \cdots + \cos(n\alpha_{k-1}) + \cos(n\alpha_k)\right]\sin\omega t \tag{24.8}$$

The desired effective value of output voltage, V_1, is controlled by using the modulation index and higher harmonic order desired to abolish have to been equal to zero. It is clear that so as to eliminate harmonic-order until the number of $k-1$ and control modulation index, we need harmonic equations until the number of k. Therefore, for example, a nine-level inverter may ensure the control of modulation index and elimination of three harmonic-orders [9, 12].

The modulation index is given by:

$$M = \frac{V_1}{kV_{dc}} \tag{24.9}$$

where, V_1 is the first harmonic of the output voltage in the same time. For $(2k+1)$ level inverter, harmonic equations until the number of k is given in the following equations first of which is related to desired output voltage value (V_1).

Equation (24.10) changes according to harmonic-orders' number desired to eliminate and level number of multilevel inverter.

These equations are transcendental and nonlinear. To handle this problem that consists of minimization the harmonics, optimization techniques are applied.

$$\begin{aligned}
\cos(\alpha_1) + \cos(\alpha_2) + \cdots + \cos(\alpha_{k-1}) + \cos(\alpha_k) &= Mk\tfrac{\pi}{4} \\
\cos(3\alpha_1) + \cos(3\alpha_2) + \cdots + \cos(3\alpha_{k-1}) + \cos(3\alpha_k) &= 0 \\
&\vdots \\
\cos((N_{level}-4)\alpha_1) + \cdots + \cos((N_{level}-4)\alpha_k) &= 0 \\
\cos((N_{level}-2)\alpha_1) + \cdots + \cos((N_{level}-2)\alpha_k) &= 0
\end{aligned} \tag{24.10}$$

Conventional techniques such as Newton-Raphson method can be applied to this problem. But this method may not give optimum solution of equations above. Also, it is time consuming and complicated in accordance with GA. Therefore, in this paper the solution of the SHE equations has been solved by GA. This method uses fitness function to optimize the solution. In the solution of SHE problem, the fitness function must be THD's value given in (24.15) [1].

24.4 Genetic Algorithm Based She Technique

Genetic Algorithm is quantitative method that solves the problem and has been the most used technique to optimize a function which originates in constrained functions [10]. The algorithm solves the problem by using genetic operators, the types of which alter as to problem. And the algorithm can be adapted to an optimization problem easily. The flow-diagram of the algorithm is given in Fig. 24.5.

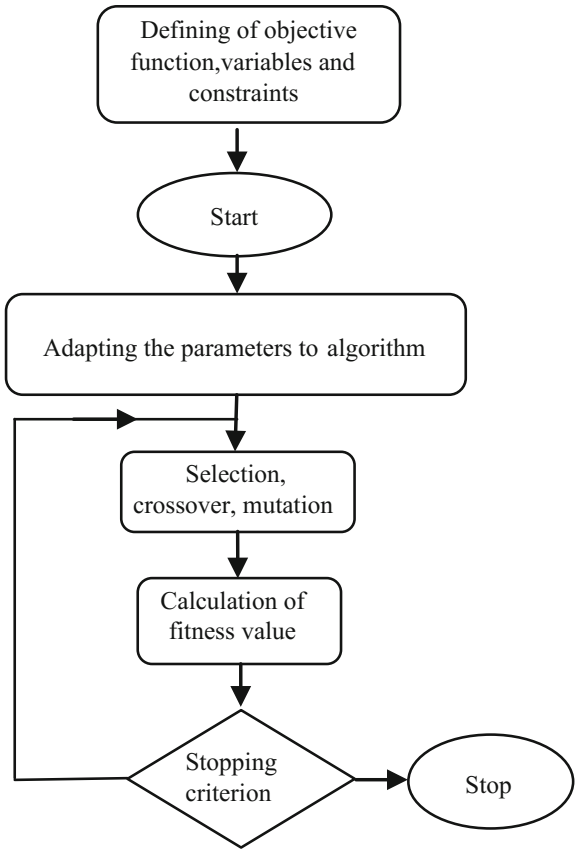


Fig. 24.5 The flow chart of genetic algorithm

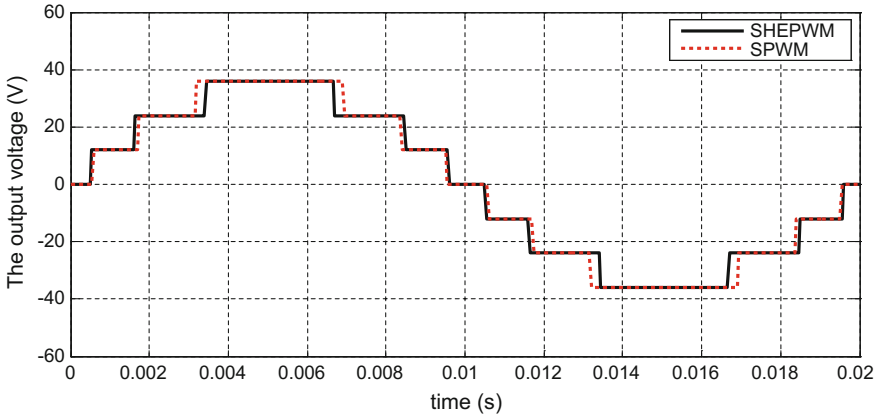


Fig. 24.6 Output voltage wave form of seven-level inverter with SPWM and SHEPWM eliminating 3rd and 7th harmonic orders

In order to solve SHE equations, the harmonic-orders desired to eliminate are depicted as constraint functions. After this state, it will be mentioned about the adaptation of SHE problem to GA for nine-level inverter. For example, constraint functions are given in (24.11) for nine-level inverter eliminating 3rd, 5th, 7th harmonic components. GA evaluates these constraint functions and as a result, generates penalty functions. Thus, GA uses best population to optimize the problem. Population's size of each iteration is constant. But, chromosomes that constitute the population vary in accordance with evaluation of penalty function.

Then crossover and mutation operators are applied. The all period is iterated as the number of iterations [13].

1., 2., 3. and 4. constraint functions are given respectively as:

$$\begin{aligned}
&\cos(\alpha_1)+\cos(\alpha_2)+\cos(\alpha_3)+\cos(\alpha_4)-Mk\frac{\pi}{4}=0\\
&\cos(3\alpha_1)+\cos(3\alpha_2)+\cos(3\alpha_3)+\cos(3\alpha_4)=0\\
&\cos(5\alpha_1)+\cos(5\alpha_2)+\cos(5\alpha_3)+\cos(5\alpha_4)=0\\
&\cos(7\alpha_1)+\cos(7\alpha_2)+\cos(7\alpha_3)+\cos(7\alpha_4)=0
\end{aligned} \tag{24.11}$$

A basic GA can generate the chromosome including genes as binary coded or real coded. In this paper, real coded genetic algorithm has been used so as to solve SHE equations.

Algorithm is started numerically with producing of variables under the condition following equation [14];

$$0<\alpha_1<\alpha_2<\alpha_3<\alpha_4<\pi/2 \tag{24.12}$$

The population changes according to type of problem, size of population is selected as 500.

Selection is reproduction of the starting population. In this paper, tournament selection is used. In tournament selection, two chromosomes are chosen randomly to become pairs. One of these pairs is specified according to the penalty function which is acquired from constraint function and incorporated in the new generation.

Penalty function evaluates closeness of the values of the chromosomes should be produced if there is any out of range of the constraint function. This range is between 0 and 0,01 in this study. This range determines precision of produced switching angles in multilevel inverter. The penalty function used in GA is given in following;

$$P(\alpha) = f(\alpha) + \sum_{k=1}^{4} r[g_k(\alpha)] \tag{24.13}$$

where, $P(\alpha)$ is penalty function, r = 500, g_k are constraint functions. The process of selection can be summarized with three sections;

- If both pairs of chromosomes are in appropriate range, that is constraint functions are not between 0 and 0,01, chromosome with small penalty function is incorporated in the new generation.
- If one of the pairs of chromosomes is in the appropriate range, that is the result of penalty function is equal to zero, this chromosome is incorporated in the new generations.
- If both pairs of chromosomes are in the appropriate range, chromosome with small fitness function is incorporated in the new generation.

One of the most notable parts of GA is crossover. Crossover performs creation of new offspring by taking into account selected chromosomes.

Several crossover techniques can be performed GA such as linear, arithmetic, flat, uniform and heuristic etc. [15]. In this paper, arithmetical technique is used. In this technique, a number p which is named as crossover parameter is produced between 0 and 1 for each mutual gene of selected pairs of chromosomes. If this produced number is bigger than the cross rate, mutual genes are put into process of crossover as to arithmetically crossover Eq. (24.14). For example, if only 3rd genes are in the crossover for $p = 0{,}2$, chromosomes are changed as in Table 24.1 and the arithmetically crossover equation is given by:

Table 24.1 Aritmetically crossover sample of GA solving SHE equations

Before crossover	Values of genes (rad)			
	Gene 1	*Gene 2*	*Gene 3*	*Gene 4*
Chromosome 1	0,21	0,44	0,8	1,4
Chromosome 2	0,36	0,77	0,91	1,36
After crossover	Values of genes (rad)			
	Gene 1	*Gene 2*	*Gene 3*	*Gene 4*
Chromosome 1	0,21	0,44	0,822	1,4
Chromosome 2	0,36	0,77	0,888	1,36

$$\begin{aligned} Gene_{3,CH1} &= pGene_{3,CH2} + (1-p)Gene_{3,CH1} \\ Gene_{3,CH2} &= pGene_{3,CH1} + (1-p)Gene_{3,CH2} \end{aligned} \tag{24.14}$$

After crossover, mutation occurs. There have been many types of mutation operator such inversion, scramble, reversing [16]. In real-coded GA, in order to perform uniform mutation, a number is produced between 0 and 1 for each gene of each chromosome. If produced number for gene is lower than the mutation rate, gene is reproduced according to limits [17].

The most important point for the GA is evaluation of each chromosome. The aim of the study is to eliminate desired harmonic orders of output voltage in multilevel inverter, so fitness function must be associated to THD equation which may change based on harmonic orders and the number of harmonic orders desired to be equal zero [10]. For example, if third, fifth, seventh harmonic orders are desired to eliminate in nine-level inverter, fitness function is given by:

$$Fitness = \frac{\sqrt{\sum_{n=3,5,7}\left(\frac{1}{n}\sum_{k=1}^{4}\cos(n\alpha_k)\right)^2}}{\sum_{k=1}^{4}\cos(n\alpha_k)} \tag{24.15}$$

Consequently, the GA calculates the fitness function in each iteration and defines chromosomes as to fitness value. GA must select the chromosomes to minimization of desired harmonic orders. Therefore, the values of fitness functions should be equal to zero [1].

24.5 Simulation Results

In this study, Selective Harmonic Elimination technique has been applied to seven-level and nine-level inverters. Simulation results have been given to assess the effectiveness of GA based SHE. The SHE equations were solved by GA and simulation was developed in Matlab&Simulink. The fundamental frequency of the output voltage is 50 Hz and switching frequency of the semiconductor power switches in the level module units and H-bridge is 100 Hz and 50 Hz, respectively.

In Table 24.2 and Fig. 24.7, the eliminated harmonic orders are shown for seven-level inverter. Similarly, Table 24.3 and Fig. 24.9 give the results of nine-level inverter. Additionally, Fig. 24.6 shows the output voltage waveforms of seven-level inverter having 12 V isolated DC source. The output voltage's maximum rate is 36 V because of the three modules in seven-level inverter.

In Table 24.2, it has been aimed at eliminate 3rd and 7th harmonic orders. While the eliminated harmonic orders are almost equal the zero, 5th harmonic component has increased from % 0,01 to % 4,28 when it is compared to SPWM. To overcome this, 5th harmonic order has been eliminated in nine-level inverter as shown in

Table 24.2 Seven level inverter % harmonics value in SPWM and elimination of 3rd and 7th harmonic orders in SHEPWM and switching angles of SHEPWM

M = 1	Amplitude of Harmonics %	
Switching angles for SHEPWM $\alpha_1 = 0,144, \quad \alpha_2 = 0,493,$ $\alpha_3 = 1,059\, rad$	SHEPWM	
Harmonic order	SPWM	3rd and 7th harm. elimination
1	100	100
3	**1,5**	**0,16**
5	0,01	4,28
7	**2,25**	**0,17**
9	3,33	4,83
11	1,68	4,69

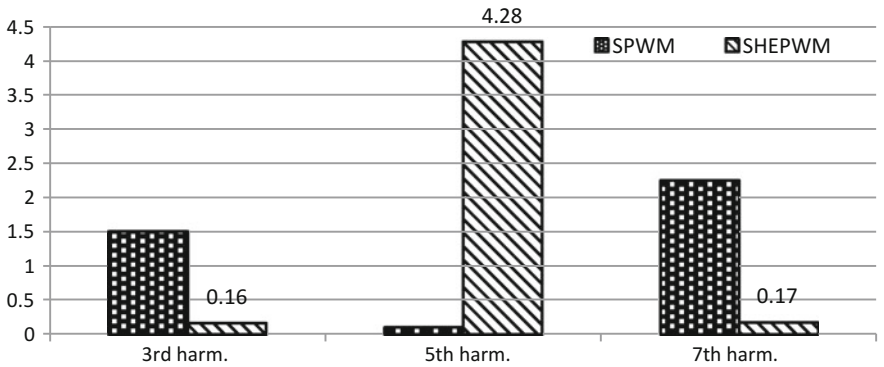


Fig. 24.7 Elimination of 3rd and 7th harmonic orders in seven-level inverter

Table 24.3 Nine-level inverter % harmonics value in SPWM and elimination of 3rd, 5th and 7th harmonic orders in SHEPWM and switching angles of SHEPWM

M = 1	Amplitude of Harmonics %	
Switching angles for SHEPWM $\alpha_1 = 0,207, \alpha_2 = 0,311,$ $\alpha_3 = 0,661, \alpha_4 = 1,041\, rad$	SHEPWM	
Harmonic order	SPWM	3rd, 5th, 7th harm. elimination
1	100	100
3	**0,79**	**0,08**
5	**0,38**	**0,22**
7	**0,7**	**0,09**
9	1,71	4,48
11	2,62	1,98

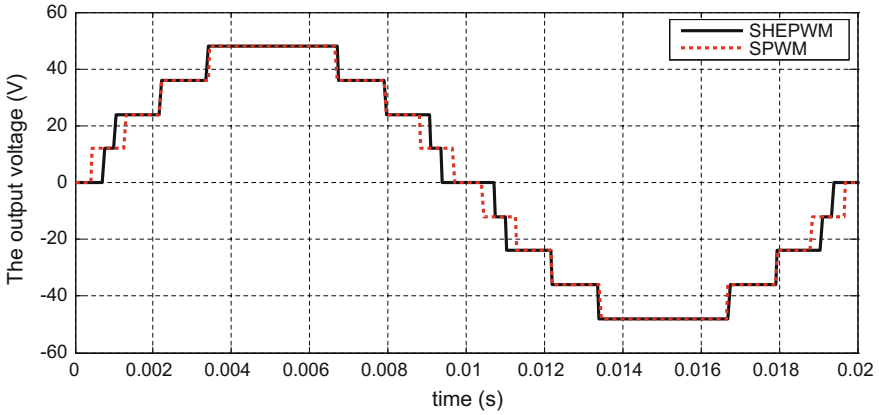


Fig. 24.8 Output voltage waveforms in nine-level inverter with SPWM and SHEPWM eliminating 3rd, 5th and 7th harmonic orders

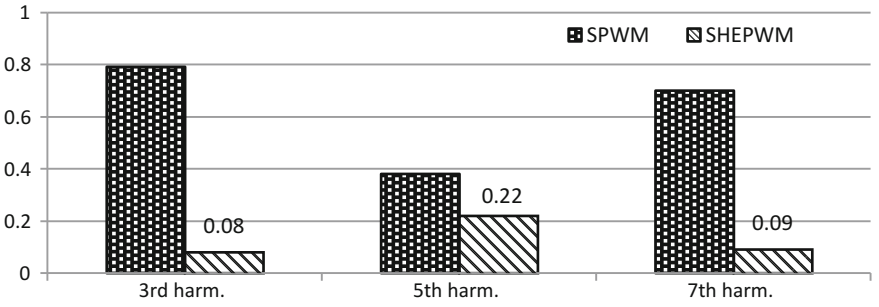


Fig. 24.9 Elimination of 3rd, 5th and 7th harmonic orders in nine-level inverter

Table 24.3 and also Fig. 24.9 clearly. In addition to this, Fig. 24.8 illustrates the output voltage waveforms of nine-level inverter. The output voltage's maximum rate is 48 V with four voltage step in positive side. The whole step consists of four voltage step in positive and four voltage step negative side and zero.

GA based SHE technique enables to eliminate any harmonic orders. To demonstrate this, 5th, 7th, 11th harmonic orders are eliminated in nine-level inverter. In Table 24.4 and Fig. 24.10, the harmonic orders are shown for nine-level inverter. When compared Tables 24.3 and 24.4, it is understood that with GA based SHE technique, harmonic orders desired to eliminate can be changed thanks to adaptability of SHE equations to any harmonic order. With the angles given in Table 24.4, three-phase simulation of the multilevel inverter, until 13th harmonics is eliminated because three-phase systems have no 3rd, 9th, 15th, etc. harmonic orders. The line and phase voltage waveform is given in Fig. 24.11.

Table 24.4 Nine-level inverter % harmonics value in SPWM and elimination of 5, 7 and 11th harmonic orders in SHEPWM and switching angles of SHEPWM

M = 1	Amplitude of harmonics %	
Switching angles for SHEPWM $\alpha_1 = 0,182, \alpha_2 = 0,412,$ $\alpha_3 = 0,742, \alpha_4 = 1,089\ rad$	SHEPWM	
Harmonic order	SPWM	5th, 7th, 11th harm. elimination
1	100	100
3	0,79	4,71
5	**0,38**	**0,14**
7	**0,7**	**0,12**
9	1,71	3,2
11	**2,62**	**0,1**

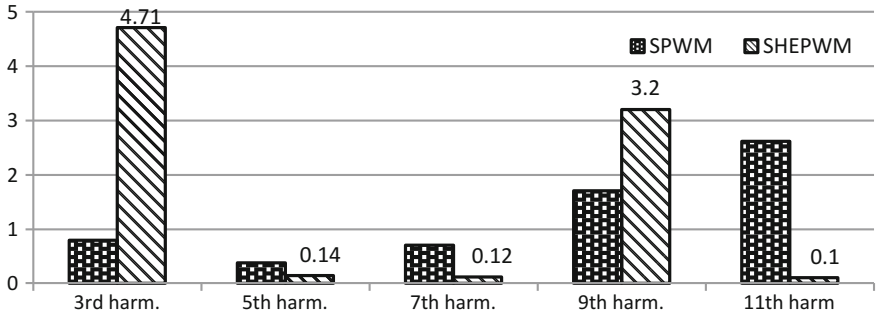


Fig. 24.10 Elimination of 5, 7 and 11th harmonic orders in nine-level inverter

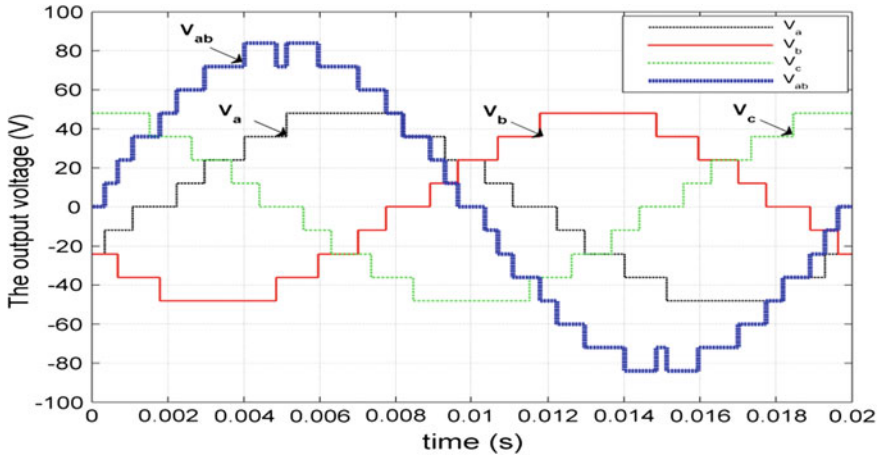


Fig. 24.11 Phase and line voltage of three-phase nine-level inverter eliminating 5, 7 and 11th harmonic orders

24.6 Conclusion

GA based selective harmonic elimination in cascaded multilevel inverter with reduced number of switches has been presented. The advantage of proposed inverter topology is to having a structure with the reduced number of switches. Therefore, the hardware cost of proposed multilevel inverter is lower and it is compact. Also, switching techniques applied to conventional multilevel inverter can be used with this topology.

SHE equations of the output voltage of seven-level and nine-level multilevel inverter have been solved by the developed GA software without the use of GA toolbox of MATLAB and applied to proposed inverter topology. The obtained results have clearly proved the effectiveness of the proposed multilevel structure and GA based SHE method.

References

1. Karaca H, Bektas E (2015) GA based selective harmonic elimination for multilevel inverter with reduced number of switches. In: Lecture notes in engineering and computer science: proceedings of the world congress on engineering and computer science 2015, WCECS 2015, vol 1. San Francisco, USA, pp 204–209, 21–23 Oct 2015
2. Karaca H (2013) A novel topology for multilevel inverter with reduced number of switches. In: Lecture notes in engineering and computer science: proceedings of the world congress on engineering and computer science 2013, WCECS 2013, vol. 1. San Francisco, USA, pp 350–354, 23–25 Oct 2013
3. Babaei E, Hosseini SH (2009) New cascaded multilevel inverter topology with minimum number of switches. Energy Convers Manag 50:2761–2767
4. Rodriguez J, Lai JS, Peng FZ (2002) Multilevel inverters: survey of topologies, controls, and applications. IEEE Trans Ind Appl 49(4):724–738
5. Bektas E, Karaca H (2015) Harmonic minimization technique for multilevel inverter using cascaded H-bridge modules. In: Proceedings of international scientific conference, UNITEC 2015, vol. 1. Gabrovo, Bulgaria, pp 139–143, 20–21 Nov 2015
6. Fei W, Du X, Wu B (2010) A generalized Half-Wave symmetry SHE-PWM formulation for multilevel voltage inverters. IEEE Trans Ind Electron 57(9):3030–3038
7. Kumar J, Das B, Agarwal P (2008) Selective harmonic elimination technique for a multilevel inverter. In: Proceedings of the 15th national power systems conference IIT, Bombay, India, pp 608–613
8. Tolbert LM, Chiasson JN, Du Z, McKenzie KJ (2005) Elimination of harmonics in a multilevel converter with nonequal DC sources. IEEE Trans Ind Appl 41(1):724–738
9. El-Naggar K, Abdelhamid TM (2008) Selective harmonic elimination of new family of multilevel inverters using genetic algorithms. Energy Convers Manag 49:89–95
10. Kumar D, Dr S, Pattnaik Mrs, Singh V (2012) Genetic algorithm based approach for optimization of conducting angles in cascaded multilevel inverter. Int J Eng Res Appl (IJERA). 2(3):2389–2395
11. Bouhali O, Bouaziz F, Rizoug N, Talha A (2013) Solving harmonic elimination equations in multi-level inverters by using neural networks. Int J Inf Electr Eng 3(2):191–195
12. Baskaran J, Thamizharasan S, Rajtilak R (2012) GA based optimization and critical evulation SHE methods for three-level inverter. Int J Soft Comput Eng (IJSCE) 2(3):321–326

13. Gkoutioudi K, Karatza HD (2012) A simulation study of multi-criteria scheduling in grid based on genetic algorithms. In: Proceedings of the 10th IEEE international symposium on parallel and distributed processing with applications
14. Chiasson JN, Tolbert LM, McKenzie KJ, Du Z (2004) A unified approach to solving the harmonic elimination equations in multilevel converters. IEEE Trans Power Electron 19 (2):478–490
15. Mendes JM (2003) A comparative study of crossover operators for genetic algorithms to solve job shop scheduling problem. WSEAS Trans Comp 12(4), 164–173
16. Soni N, Kumar T (2003) Study of various mutation operators in genetic algorithms. Int J Comput Sci Inf Technol 5:4519–4521
17. Wenyi Z, Xiaodan M, Zhenhua L (2014) The simulation research for selective harmonic elimination technique based on genetic algorithm. In: Proceedings of the 33rd Chinese Control Conferences, Nanjing, China, 28–30 July 2014

Chapter 25
Proposal of a Modulated Extended Cumulative Exposure Model for the Step-Up Voltage Test

Takenori Sakumura and Toshinari Kamakura

25.1 Introduction

In past decades, accelerated life testing (ALT) is one of the most useful methods to find the lifetime of industrial materials (e.g., electrical insulation) in short time [7]. Using failure data from ALT, we can estimate reliability of items such as mean lifetime and some quantile of the lifetime distribution at the service stress. For example, when the major factor of the deterioration of the insulation is the electrical stress, the power law,

$$h(v) = \frac{K}{(v - v_{th})^p} \tag{25.1}$$

has been empirically used to estimate the lifetime of the insulation at the service stress, as shown in Fig. 25.1. Here, K is a constant parameter, p is the degradation rate, v is the electrical stress, and v_{th} is the threshold stress below which the failures will not occur [1].

A Weibull distribution is often used for the reliability distribution in the power law model. Then, a Weibull power law [7] with the threshold stress can be assumed [2]:

$$F(t; v_i) = 1 - \exp\left[-\left\{\frac{t}{h(v_i)}\right\}^{\beta}\right] = 1 - \exp\left[-\left\{\frac{(v_i - v_{th})^p}{K} t\right\}^{\beta}\right]. \tag{25.2}$$

T. Sakumura (✉) · T. Kamakura
The Department of Industrial and Systems Engineering, Chuo University,
1-13-27 Kasuga, Bunkyoku, Tokyo 112-8551, Japan
e-mail: sakumura@indsys.chuo-u.ac.jp

T. Kamakura
e-mail: kamakura@indsys.chuo-u.ac.jp

S.-I. Ao et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-981-10-2717-8_25

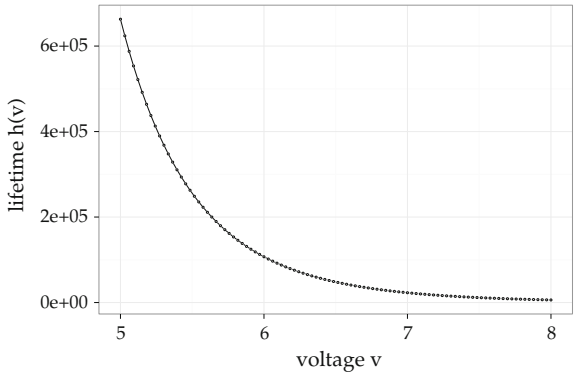


Fig. 25.1 Illustration of the power law. The parameter values used are $K = 0.632122 \times 20^{10}$, $p = 10$, $v_{th} = 0$

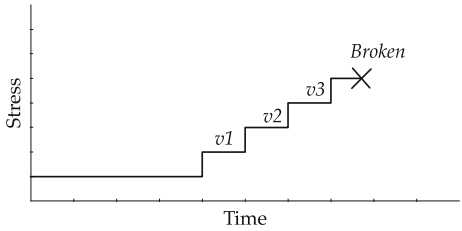


Fig. 25.2 Step-stress test. In step-stress loading, a specimen is subjected to successively higher levels of stress. The stress on a specimen is increased step by step until it fails. (× failure and v1, v2, and v3 are step-stress.)

Here, t is the time to failure, v_i is the applied stress at level i, and β is the shape parameter in the Weibull distribution. We can obtain these parameters using maximum likelihood estimation method. If the failure data from ALT is the type II censored data, the likelihood function is expressed as,

$$L = \prod_{i,j} f(t_{i(j)}; v_i)^{\delta i(j)} \times \left\{1 - F(t_{i(j)}; v_i)\right\}^{1-\delta i(j)}, \tag{25.3}$$

where, f is a pdf of F, $i(j)$ denotes that sample j is broken at level i, and $\delta i(j)$ is a indicator function.

To accelerate ALT much faster, step-stress accelerated life testing (SSALT) is considered to be a special case of ALT. In SSALT, the stress levels are increased during the test period in a specified discrete sequence [5], i.e. the step-stress test as shown in Fig. 25.2.

The first report to use the maximum likelihood estimation method to SSALT combined with the Weibull power law was made Nelson [6], where the cumulative exposure model (CEM) is also proposed. The CEM is often used to express the failure probability models and is widely accepted in reliability fields because accumulation of fatigue to each stress level is considered to be reasonable.

In electrical insulation tests, however, another model is sometimes referred to. That is, the accumulation of fatigue is assumed to vanish each time the applied stress is reduced to zero. This indicates that the insulation materials possess a self-restoring feature when a rest time is given between the consecutive stress imposed. This model is called the memoryless model (MM) [4].

Hirose and Sakumura [4] proposed the extended cumulative exposure model (ECEM), which includes features of the CEM and the MM simultaneously. When there is accumulation of fatigue, the weight on the CEM may be large; however, when no accumulation of fatigue occurs, the weight on the MM may be large.

The fact that the ECEM includes the CEM and the MM deserves special emphasis, but this model is difficult for estimating for parameters owing to recursive calculations. Thus, we proposed an another modulated ECEM model [8]. We assume vanishing of the fatigue accumulation as returning the imposed time.

25.2 Step-Stress Test

The step-stress test, also called the step-up voltage test, is described as follows. (1) A stress $v_1 = \Delta v$ is imposed on the insulation for time t_1. (2) If the insulation is not broken during that period, then $v_2 = v_1 + \Delta v$ is loaded. (3) This procedure is repeated until the insulation is broken, as shown in Fig. 25.2. (4) The resultant breakdown voltage is recorded as v_i when it fails, during that stress loading period between v_i and v_{i+1}.

25.3 Proposed Modulated Extended Cumulative Exposure Model Based on Time-Scale Changing

We briefly review the cumulative exposure model (CEM), the memoryless model (MM), and the extended cumulative exposure model (ECEM). Then, we propose an another modulated ECEM that would be resulted in time reducing. In this section, we assume the power law under the stress level v_i be expressed as Eq. (25.1).

25.3.1 Cumulative Exposure Model, CEM

The CEM [6] connects the cumulative distribution functions by transforming the stress loading durations recursively. For example, at the very first step, we assume,

$$F_2(s_1) = F_1(t_1). \tag{25.4}$$

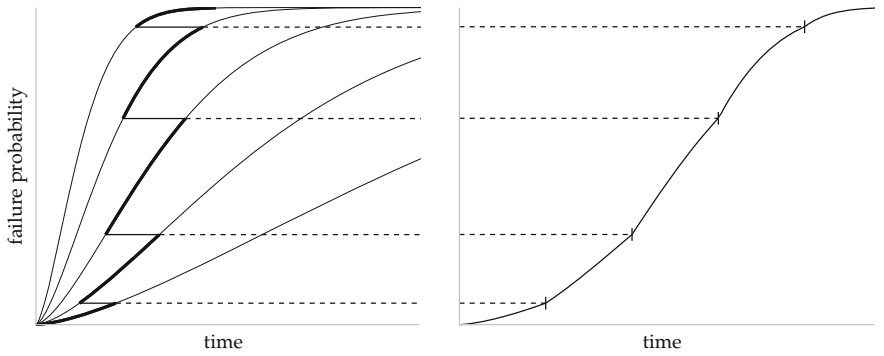


Fig. 25.3 Cumulative exposure model (CEM); *left* is the Weibull distribution $F_i(t)$ of specimens failing by time t under constant stress v_i for step i and *right* is the cumulative distribution $G(t)$

This means that the fatigue accumulation, $F_1(t_1)$, succeeds the next stress imposed. Then, s_1 is determined by

$$s_1 = (t_1 - t_0)\left(\frac{v_1 - v_{th}}{v_2 - v_{th}}\right)^p. \tag{25.5}$$

In general, $F_i(s_{i-1})$ and s_{i-1} are

$$F_i(s_{i-1}) = F_{i-1}(\Delta t_{i-1} + s_{i-2}), \tag{25.6}$$

$$s_{i-1} = (\Delta t_{i-1} + s_{i-2})\left(\frac{v_{i-1} - v_{th}}{v_i - v_{th}}\right)^p, \tag{25.7}$$

where, Δt_i expresses the time duration $t_i - t_{i-1}$ at a constant stress application v_i. Thus, a consistent continuous cumulative distribution function (CDF) is defined as,

$$G(t) = 1 - \exp\left\{-\varepsilon(t)^\beta\right\}, \tag{25.8}$$

$$\varepsilon(t) = \frac{1}{K}\left[(v_i - v_{th})^p(t - t_{i-1}) + \sum_{j=1}^{i-1}\left\{(v_j - v_{th})^p \Delta t_j\right\}\right]. \tag{25.9}$$

This probability model will be shown in Fig. 25.3.

25.3.2 Memoryless Model (MM)

We assume that all the fatigue accumulations vanish everytime the applied stress is reduced to zero. This indicates that the insulation materials reveal a self-restoring feature when a rest time is given between the consecutive stress imposed. This is

called the memoryless model (MM) here. Considering that all the fatigue accumulations vanish when the applied stress is reduced to zero, the memoryless model is exactly the same model as that presented in Eq. (25.2) after all.

25.3.3 Extended Cumulative Exposure Model (ECEM)

The extended cumulative exposure model (ECEM) is assumed partial fatigue accumulation instead of full fatigue accumulation as shown in Eq. (25.10); the rate α expresses the fatigue accumulation rate. For example, at the first step,

$$F_2(s_1) = \alpha F_1(t_1), \quad (0 \le \alpha \le 1). \tag{25.10}$$

This means that some fraction, $(1-\alpha)F_1(t_1)$, will vanish, and that the remaining fatigue accumulation, $\alpha F_1(t_1)$, will succeed the next stress imposed. Then, s_1 is determined by

$$s_1 = \frac{K}{(v_2 - v_{th})^p} \left[-\log\{1 - \alpha F_1(t_1)\} \right]^{1/\beta}. \tag{25.11}$$

In general, $F_i(s_{i-1})$ and s_{i-1} become

$$F_i(s_{i-1}) = \alpha F_{i-1}(\Delta t_{i-1} + s_{i-2}), \tag{25.12}$$

$$s_{i-1} = \frac{K}{(v_i - v_{th})^p} \left[-\log\{1 - \alpha F_{i-1}(\Delta t_{i-1} + s_{i-2})\} \right]^{1/\beta}, \tag{25.13}$$

and the corresponding continuous CDF is calculated recursively,

$$G(t) = 1 - \exp(-\varepsilon(t)^\beta), \tag{25.14}$$

$$\varepsilon(t) = \frac{(v_i - v_{th})^p}{K}(t - t_{i-1}) + \left[-\log\{1 - \alpha F_{i-1}(\Delta t_{i-1} + s_{i-2})\} \right]^{1/\beta}. \tag{25.15}$$

This probability model will be shown in Fig. 25.4.

We can see that this model is an extension of the CEM and the MM. When $\alpha = 0$, the model is reduced to the MM. When $\alpha = 1$, the model is reduced to the CEM. Thus, this model includes the CEM and the MM together. When $0 < \alpha < 1$, we can control the partial fatigue accumulation by a constant α.

25.3.4 Modulated Extended Cumulative Exposure Model

Here, we propose the modulated ECEM based on time scale changing. We consider that vanishings of the fatigue accumulations are defined as the imposed time revers-

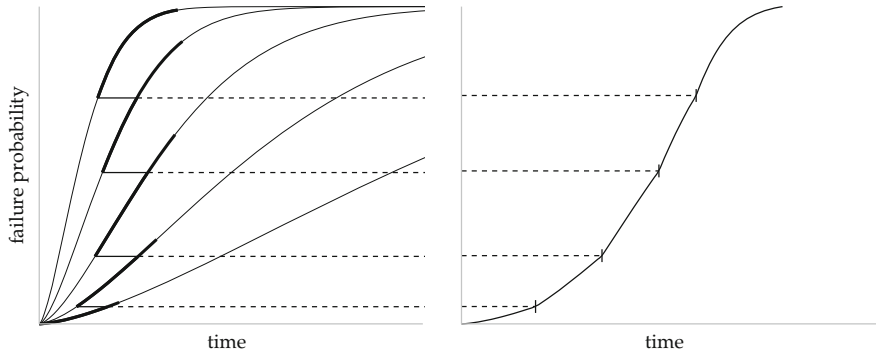


Fig. 25.4 Extended Cumulative Exposure Model (ECEM) at $\alpha = 0.8$; *left* is the Weibull distribution $F_i(t)$ of specimens failing by time t under constant stress v_i for step i and *right* is the cumulative distribution $G(t)$

ing. The rate γ expresses the imposed time rate. For example, at the very first step, we assume,

$$F_2(s_1) = F_1(\gamma t_1), \quad (0 \le \gamma \le 1). \tag{25.16}$$

The above Eq. (25.16) means that the imposed time will be $t_1 - (1-\gamma)t_1$ back to the time points $(1-\gamma)t_1$ and that the fatigue accumulation corresponding to the imposed time $F_1(\gamma t_1)$ is taken over to the next stress imposed. Then, s_1 is determined by

$$s_1 = \gamma t_1 \left(\frac{v_1 - v_{th}}{v_2 - v_{th}} \right)^p, \tag{25.17}$$

and the corresponding continuous CDF becomes

$$\begin{aligned} G(t) &= F_2(t - t_1 + s_1) \\ &= 1 - \exp\left[- \left\{ \frac{1}{K} \left[(v_2 - v_{th})^p (t - t_1) + \gamma (v_1 - v_{th})^p \Delta t_1 \right] \right\}^\beta \right], \end{aligned} \tag{25.18}$$

where, Δt_i expresses the time duration $t_i - t_{i-1}$ at a constant stress application v_i. In general, $F_i(s_{i-1})$, and s_{i-1} become

$$F_i(s_{i-1}) = F_{i-1}(\gamma(\Delta t_{i-1} + s_{i-2})), \tag{25.19}$$

$$s_{i-1} = \sum_{j=1}^{i-1} \gamma^j \Delta t_{i-j} \left(\frac{v_{i-j} - v_{th}}{v_i - v_{th}} \right)^p. \tag{25.20}$$

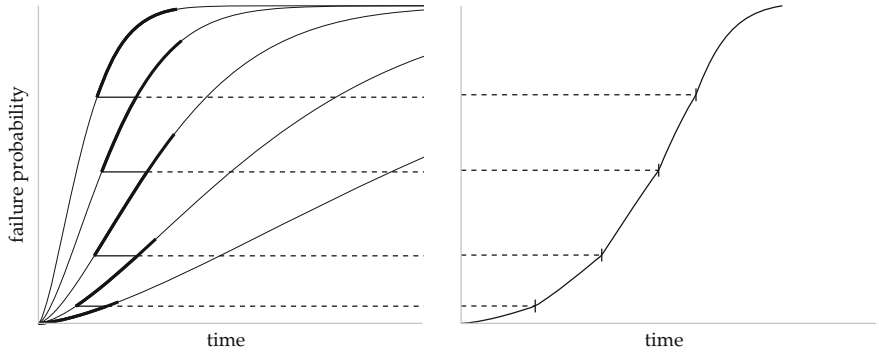


Fig. 25.5 Modulated Extended Cumulative Exposure Model Based On Time Scale Changing at $\gamma = 0.8$; *left* is the Weibull distribution $F_i(t)$ of specimens failing by time t under constant stress v_i for step i and *right* is the cumulative distribution $G(t)$

The corresponding continuous CDF is

$$\begin{aligned} G(t) &= F_i(t - t_{i-1} + s_{i-1}) \\ &= 1 - \exp[-\varepsilon(t)^{\beta}] \end{aligned} \tag{25.21}$$

$$\varepsilon(t) = \frac{1}{K}\left[(v_i - v_{th})^p (t - t_{i-1}) + \sum_{j=1}^{i-1} \gamma^j (v_{i-j} - v_{th})^p \Delta t_{i-j}\right]. \tag{25.22}$$

Here, for numerical stability in computation, we transformed $1/K = k^p$. Then Eq. (25.22) becomes

$$\varepsilon(t) = \left[\{k(v_i - v_{th})\}^p (t - t_{i-1}) + \sum_{j=1}^{i-1} \gamma^j \{k(v_{i-j} - v_{th})\}^p \Delta t_{i-j}\right]. \tag{25.23}$$

This probability model will be shown in Fig. 25.5.

We can see that this model is another extension of the CEM and the MM as the ECEM is. When $\gamma = 0$, the model is equivalent to the MM, that is, Weibull distribution with power law, Eq. (25.2). When $\gamma = 1$, the model is equivalent to the CEM. It is clear from the Eqs. (25.2), (25.8), (25.9), (25.21) and (25.22). Thus, this model includes the CEM and the MM together like the ECEM.

25.4 Simulation Study

To check if we can estimate the parameters well, we perform a simulation study. The simulation condition is $v_{th} = 0$, $p = 10$, $K = 0.632122 \times 20^{10}$, and $\beta = 0.5,\ 1,\ 1.5$, which is a mimicked case in typical solid electrical insulation [3]. The starting stress

Table 25.1 Basic statistics for the estimated parameters in simulation study

β	γ		$\hat{\gamma}$	$\hat{p}$	$\hat{k}$	$\hat{\beta}$
1	0	mean	1.06e-05	10.1	0.0524	0.997
		sd	3.64e-07	0.521	0.000442	0.0557
		rmse	7.18e-07	0.525	0.000445	0.0558
	0.5	mean	0.529	10	0.0522	0.989
		sd	0.0195	0.41	0.000338	0.045
		rmse	0.0352	0.413	0.000373	0.0463
	1	mean	0.949	10.1	0.0554	0.949
		sd	0.0296	0.264	0.00128	0.0474
		rmse	0.0587	0.286	0.00335	0.0694
0.5	0	mean	1.06e-05	9.94	0.0525	0.505
		sd	3.53e-07	0.547	0.000893	0.0265
		rmse	7.35e-07	0.55	0.000903	0.0271
	0.5	mean	0.549	9.92	0.0522	0.5
		sd	0.0441	0.48	0.000993	0.0235
		rmse	0.066	0.487	0.001	0.0235
	1	mean	0.964	10	0.0551	0.504
		sd	0.0259	0.307	0.00114	0.0264
		rmse	0.0445	0.307	0.00302	0.0266
1.5	0	mean	1.06e-05	10.1	0.0524	1.49
		sd	4.01e-07	0.501	0.000284	0.0828
		rmse	7.09e-07	0.509	0.000285	0.0834
	0.5	mean	0.521	10.1	0.0522	1.47
		sd	0.0178	0.324	0.000225	0.0688
		rmse	0.0279	0.341	0.000249	0.074
	1	mean	0.944	10.2	0.0557	1.36
		sd	0.0303	0.257	0.00123	0.0662
		rmse	0.0634	0.3	0.00353	0.156

The upper values are presented as mean, the middle as standard deviation, and the lower as root mean squared error (RMSE)

is 0 and the stress time duration to each stress is 1 (unit time). We set three cases $\gamma = 0,\ 0.5,\ 1$. The number of items, N, is 200 which is sufficiently large enough for estimation. The number of replications in simulation is 10,000 for each case.

The basic statistics (mean, standard deviation and the root mean square error (RMSE)) for the estimated parameters are shown in Table 25.1.

As a typical case, we present Fig. 25.6 which shows the estimated parameter frequency distributions for $\hat{\gamma}$, $\hat{p}$, and $\hat{k}$, and $\hat{\beta}$ when $\gamma = 0.5$, $p = 10$, $k = K^{-1/p} = 0.05234677$, $\beta = 1.0$. From Table 25.1 and Fig. 25.6, we can see that the parameters are well estimated.

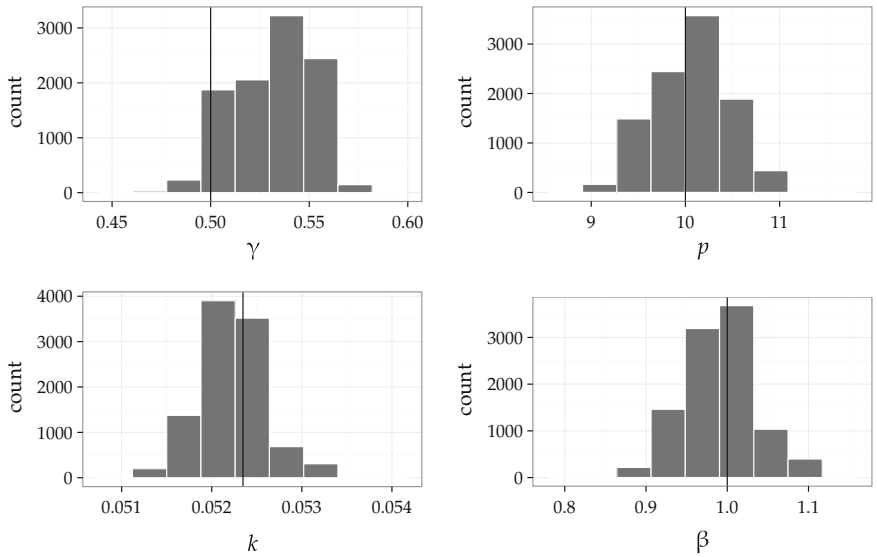


Fig. 25.6 Frequency distributions for estimated parameters in a simulation study. $\gamma = 0.5$, $n = 10, k = 0.05234677, \beta = 1.0$

25.5 Experimental Case

In this section, we apply our model to the insulation data in Nelson (1980) [6]. The data is the step-stress test data on Cable I [6]. The number of specimen is 21. In steps 1 through 4, each specimen was held for 10 min each at 5, 10, 15, and 20 kV. In steps 5 through 14, a specimen has the same hold time at each voltage (15 min, 60 min (1 h), 240 min (4 h), or 960 min (16 h)). The final step and the total time to failure are recorded when a specimen broke down. Each specimen have a thickness which used to calculate its stress as the voltage divided by the thickness (See [6] for more detail).

Figure 25.7 shows the plot of the maximum log likelihood versus γ, where $p = 19.937, k = 1/1616.4$, and $\beta = 0.75597$ are evaluated by Nelson (1980) [6]. As seen, we suggest that $\gamma = 0$, that is, the MM should be used. This result is consistent with the result of the ECEM by Hirose and Sakumura [4].

25.6 Discussion

Our proposed model, called the modulated ECEM, is similar to the ECEM, however our model has the good property that the Jacobian and Hessian Matrices can be obtained, thus, we can use the Newton-Raphson Method for estimating parameters. Therefore, the calculation speed is much faster.

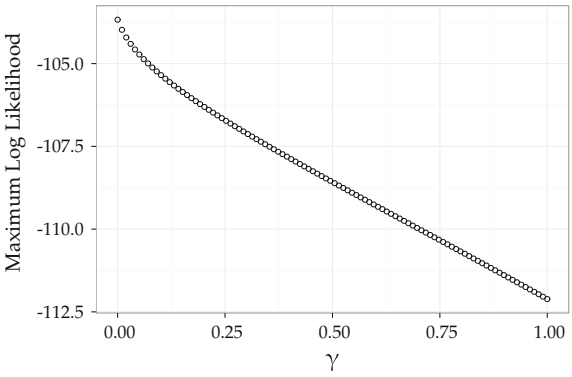


Fig. 25.7 Maximum log likelihood versus γ for Nelson's data with $p = 19.937$, $k = 1/1616.4$, $\beta = 0.75597$ estimated by Nelson (1980) [6].

In this section, we discuss about the relationship to the fatigue accumulation rate α and the imposed time rate γ. Now, we assume that Eqs. (25.10) and (25.16) is equal, that is,

$$\begin{aligned}\alpha F(t_1) &= F(\gamma t_1) \\ &= 1 - \exp[-\{(k(v_1 - v_{th}))^p \gamma t_1\}^\beta] \\ &= 1 - \exp[-\{(k(v_1 - v_{th}))^p t_1\}^\beta \gamma^\beta] \\ &= 1 - \left\{1 - F(t_1)\right\}^{\gamma^\beta}.\end{aligned}$$

Therefore,

$$\alpha = F(t_1)^{-1}\left[1 - \left\{1 - F(t_1)\right\}^{\gamma^\beta}\right]. \tag{25.24}$$

Otherwise,

$$\gamma = \left[\frac{\log(1 - \alpha F(t_1))}{\log(1 - F(t_1))}\right]^{1/\beta} \tag{25.25}$$

We can confirm from Eqs. (25.24) and (25.25); if $\gamma = 1$, then $\alpha = 1$ and if $\gamma = 0$, then $\alpha = 0$.

25.7 Conclusion

Failure data obtained from ALT are used to estimate reliability of items. In this paper, we have dealt with the case where the stress is discretely (stepwise) increases, i.e., the step-stress test. The cumulative exposure model (CEM) is often used to express the failure probability model in step-up accelerated life test. Contrary to this, the

memoryless model (MM) is also used in electrical engineering because accumulation of fatigue is not observed in some cases. In general, the extended cumulative exposure model (ECEM) includes features of both models. The CDF of ECEM is defined recursively, thus, parameter estimation is difficult.

In this paper, we have proposed a new model, the another modulated extended cumulative exposure model based on time scale changing. This model is similar to the ECEM, but we can obtain Jacobian and Hessian matrices for parameter estimation easily. To investigate whether we can estimate the parameters well, we have conducted a simulation study, and we have successfully obtained the reasonable parameters. A simulation study supports the validity of the proposed model.

Acknowledgements This work was supported by JSPS Grant-in-Aid for Young Scientists (B) Grant Number 15K21379.

Appendix

Parameter Estimation

Figure 25.8 shows a profile log-likelihood parameterized by four parameter in the modulated ECEM. By the figure, the optimum point can be found. In searching for the optimum point of the log-likelihood function, it is necessary to use iterative methods because of its nonlinearity. Optimization schemes without derivatives are many proposed, but the Newton-Raphson method is finally indispensable for problems appeared in this paper. Because, the likelihood function is extraordinary flat.

Likelihood Function

Here, N is the sample size, and $i(j)$ denotes that sample j is broken at level i. To obtain the parameters, we pursue the maximizer of the parameters in the likelihood function,

$$L = \prod_{j=1}^{N} g(t_{i(j)})^{\delta i(j)} \times \left\{1 - G(t_T)\right\}^{1-\delta i(j)}, \tag{25.26}$$

when time of breakdown is continuously observed or unobserved by censoring to the right (type II); in the latter case, t_T is the truncation time. When we only observe the number of failures at each stress level, the likelihood function becomes to

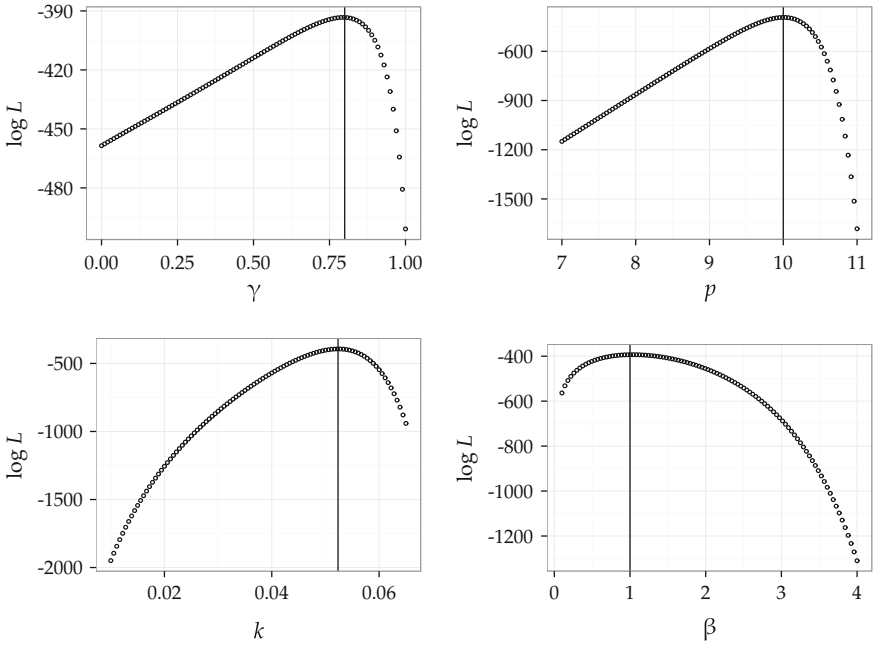


Fig. 25.8 Profile log-likelihood function for the modulated ECEM; model parameters are $\gamma = 0.8$, $p = 10$, $k = 0.05234677$, and $\beta = 1$, the sample size $N = 100$, and condition values $\Delta t = 1$, $\Delta v = 0.1$, and $v_{th} = 0$

$$L = \prod_{j=1}^{N} \left[G(t_{i(j)}) - G(t_{i(j)-1}) \right], \tag{25.27}$$

which provides the case for grouped data. Furthermore, if the truncation time $t_T(j)$ is provided with r failures, the likelihood function is

$$L = \prod_{j=1}^{N} \left[G(t_{i(j)}) - G(t_{i(j)-1}) \right] / G(t_T(j)), \tag{25.28}$$

In this paper, we deal with data from step-stress test for Eq. (25.27).

The 1st and 2nd Derivatives

We denote Eq. (25.27) as follows,

$$L = \prod_{j=1}^{N} \left[G(t_{i(j)}) - G(t_{i(j)-1}) \right] = \prod_{j=1}^{N} \left[\exp\{-\varepsilon(t_{i(j)-1})^{\beta}\} - \exp\{-\varepsilon(t_{i(j)})^{\beta}\} \right]$$
$$= \prod_{j=1}^{N} x_j, \tag{25.29}$$

then the derivatives of the log-likelihood function are,

$$\frac{\partial \log L}{\partial \theta_1} = \sum_{j=1}^{N} \frac{1}{x_j} \frac{\partial x_j}{\partial \theta_1}, \tag{25.30}$$

where, θ_1 denotes γ, p, k, and β. Note that $k = K^{-1/p}$. And,

$$\frac{\partial x_j}{\partial \theta_1} = \exp(-\varepsilon^{\beta}_{i(j)-1}) \left(-\frac{\partial \varepsilon^{\beta}_{i(j)-1}}{\partial \theta_1} \right) - \exp(-\varepsilon^{\beta}_{i(j)}) \left(-\frac{\partial \varepsilon^{\beta}_{i(j)}}{\partial \theta_1} \right), \tag{25.31}$$

where, $\varepsilon_{i(j)} = \varepsilon(t_{i(j)})$. The second derivatives are,

$$\frac{\partial^2 \log L}{\partial \theta_1 \partial \theta_2} = \sum_{j=1}^{N} \frac{\partial}{\partial \theta_1} \left(\frac{1}{x_j} \frac{\partial x_j}{\partial \theta_2} \right)$$
$$= \sum_{j=1}^{N} \left\{ -\frac{1}{x_j^2} \frac{\partial x_j}{\partial \theta_1} \frac{\partial x_j}{\partial \theta_2} + \frac{1}{x_j} \frac{\partial^2 x_j}{\partial \theta_1 \partial \theta_2} \right\}, \tag{25.32}$$

where,

$$\frac{\partial^2 x_j}{\partial \theta_1 \partial \theta_2} = \exp(\varepsilon^{\beta}_{i(j)-1}) \left\{ \left(-\frac{\partial \varepsilon^{\beta}_{i(j)-1}}{\partial \theta_1} \right) \left(-\frac{\partial \varepsilon^{\beta}_{i(j)-1}}{\partial \theta_2} \right) + \left(-\frac{\partial^2 \varepsilon^{\beta}_{i(j)-1}}{\partial \theta_1 \partial \theta_2} \right) \right\}$$
$$- \exp(\varepsilon^{\beta}_{i(j)}) \left\{ \left(-\frac{\partial \varepsilon^{\beta}_{i(j)}}{\partial \theta_1} \right) \left(-\frac{\partial \varepsilon^{\beta}_{i(j)}}{\partial \theta_2} \right) + \left(-\frac{\partial^2 \varepsilon^{\beta}_{i(j)}}{\partial \theta_1 \partial \theta_2} \right) \right\}. \tag{25.33}$$

Here, we express $\varepsilon^{\beta}_{i(j)}$ and $\varepsilon^{\beta}_{i(j)-1}$ as ε^{β} for simplicity and its derivatives for parameter θ as follows;

$$\varepsilon_{\theta_1} = \frac{\partial \varepsilon}{\partial \theta_1}, \quad \varepsilon_{\theta_1 \theta_2} = \frac{\partial^2 \varepsilon}{\partial \theta_1 \partial \theta_2},$$

then, the 1st derivatives are

$$\frac{\partial \varepsilon^\beta}{\partial \gamma} = \beta \varepsilon^{\beta-1} \varepsilon_\gamma, \quad \frac{\partial \varepsilon^\beta}{\partial p} = \beta \varepsilon^{\beta-1} \varepsilon_p, \quad \frac{\partial \varepsilon^\beta}{\partial k} = \beta \varepsilon^{\beta-1} \varepsilon_k, \quad \frac{\partial \varepsilon^\beta}{\partial \beta} = \varepsilon^\beta \log \varepsilon,$$

and the 2nd derivatives are

$$\begin{aligned}
&\frac{\partial^2 \varepsilon^\beta}{\partial \gamma^2} = \beta(\beta-1)\varepsilon^{\beta-2}\varepsilon_\gamma^2 + \beta\varepsilon^{\beta-1}\varepsilon_{\gamma\gamma}, \quad \frac{\partial^2 \varepsilon^\beta}{\partial p^2} = \beta(\beta-1)\varepsilon^{\beta-2}\varepsilon_p^2 + \beta\varepsilon^{\beta-1}\varepsilon_{pp}, \\
&\frac{\partial^2 \varepsilon^\beta}{\partial k^2} = \beta(\beta-1)\varepsilon^{\beta-2}\varepsilon_k^2 + \beta\varepsilon^{\beta-1}\varepsilon_{kk}, \quad \frac{\partial^2 \varepsilon^\beta}{\partial \beta^2} = \varepsilon^\beta \{\log \varepsilon\}^2, \\
&\frac{\partial^2 \varepsilon^\beta}{\partial \gamma \partial p} = \beta(\beta-1)\varepsilon^{\beta-2}\varepsilon_\gamma\varepsilon_p + \beta\varepsilon^{\beta-1}\varepsilon_{\gamma p}, \quad \frac{\partial^2 \varepsilon^\beta}{\partial \gamma \partial k} = \beta(\beta-1)\varepsilon^{\beta-2}\varepsilon_\gamma\varepsilon_k + \beta\varepsilon^{\beta-1}\varepsilon_{\gamma k}, \\
&\frac{\partial^2 \varepsilon^\beta}{\partial p \partial k} = \beta(\beta-1)\varepsilon^{\beta-2}\varepsilon_p\varepsilon_k + \beta\varepsilon^{\beta-1}\varepsilon_{pk}, \quad \frac{\partial^2 \varepsilon^\beta}{\partial \gamma \partial \beta} = \varepsilon^{\beta-1}\varepsilon_\gamma \{1 + \beta \log \varepsilon\}, \\
&\frac{\partial^2 \varepsilon^\beta}{\partial p \partial \beta} = \varepsilon^{\beta-1}\varepsilon_p \{1 + \beta \log \varepsilon\}, \quad \frac{\partial^2 \varepsilon^\beta}{\partial k \partial \beta} = \varepsilon^{\beta-1}\varepsilon_k \{1 + \beta \log \varepsilon\},
\end{aligned}$$

Now, $\xi_{ij} = k(v_{i-j} - v_{th})$ and Eq. (25.23), then

$$\varepsilon(t_i) = \sum_{j=0}^{i-1} \left[\gamma^j \xi_{ij}^p \Delta t_{i-j} \right].$$

Finally, we can obtain the 1st and 2nd derivatives as follows;

$$\begin{aligned}
&\varepsilon_\gamma = \sum_{j=0}^{i-1} \left[j\gamma^{j-1} \xi_{ij}^p \Delta t_{i-j} \right], \quad \varepsilon_p = \sum_{j=0}^{i-1} \left[\gamma^j \xi_{ij}^p (\log \xi_{ij}) \Delta t_{i-j} \right], \\
&\varepsilon_k = \sum_{j=0}^{i-1} \left[\gamma^j p k^{n-1} \xi_{ij}^p \Delta t_{i-j} \right], \quad \varepsilon_{\gamma\gamma} = \sum_{j=0}^{i-1} \left[j(j-1)\gamma^{j-2} \xi_{ij}^p \Delta t_{i-j} \right], \\
&\varepsilon_{pp} = \sum_{j=0}^{i-1} \left[\gamma^j \xi_{ij}^p (\log \xi_{ij})^2 \Delta t_{i-j} \right], \quad \varepsilon_{kk} = \sum_{j=0}^{i-1} \left[\gamma^j p(p-1) k^{p-2} \xi_{ij}^p \Delta t_{i-j} \right], \\
&\varepsilon_{\gamma p} = \sum_{j=0}^{i-1} \left[j\gamma^{j-1} \xi_{ij}^p (\log \xi_{ij}) \Delta t_{i-j} \right], \quad \varepsilon_{\gamma k} = \sum_{j=0}^{i-1} \left[j\gamma^{j-1} p k^{p-1} \xi_{ij}^p \Delta t_{i-j} \right], \\
&\varepsilon_{pk} = \sum_{j=0}^{i-1} \left[\gamma^j p k^{n-1} \xi_{ij}^p (\log \xi_{ij}) \Delta t_{i-j} + \gamma^j k^{-1} \xi_{ij}^p \Delta t_{i-j} \right].
\end{aligned} \tag{25.34}$$

We can obtain the Jacobian and Hessian metrices from Eqs. (25.7)–(25.34).

References

1. Hirose H (1993) Estimation of threshold stress in accelerated life-testing. IEEE Trans Reliab 42(4):650–657
2. Hirose H (1996) Theoretical foundation for residual lifetime estimation. Trans. IEE Jpn 116-B(2):168–173
3. Hirose H (2003) The relation between accelerated life tests by brand-new insulation and step-up voltage tests by working insulation. In: Proceedings of the 7th international conference on properties and applications of dielectric materials (Cat. No.03CH37417), vol 3, pp 978–981
4. Hirose H, Sakumura T (2012) The extended cumulative exposure model (ECEM) and its application to oil insulation tests. IEEE Trans Reliab 61(3):625–633
5. Meeker WQ, Escobar LA (1998) Statistical methods for reliability data. Wiey
6. Nelson W (1980) Accelerated life testing—step-stress models and data analyses. IEEE Trans Reliab R-29(2):103–108
7. Nelson W (1990) Accelerated testing: statistical models, test plans, and data analyses. Wiley-Interscience
8. Sakumura T, Kamakura T (2015) Modulated extended cumulative exposure model with application to the step-up voltage test. In: Proceedings of the world congress on engineering and computer science 2015. Lecture notes in engineering and computer science, 21–23 Oct 2015, San Francisco, USA, pp 255–260

Chapter 26
Gesture Recognition Sensor: Development of a Tool with Playful Applications to Evaluate the Physical and Cognitive Skills of Children Through the Use of Bootstrap Algorithm

Anthony Bryan Freire Conrado, Vinicio David Pazmiño Moya, Danni De La Cruz, Johanna Tobar and Paúl Mejía

26.1 Introduction

The gesture recognition sensor (GRS) works with concentration points, where different actions and positions of the object can be determined. They contribute to the development of ludic games, based on technology that promote, conserve and strengthen innate physical activity that children possess from birth [1].

There are currently a high level attention deficient in children, because in their homes there are factors that do not allow you to focus on their activities, due to these problems is to develop a tool on the basis of educational games which help both to learning as to combat with the concentration and a deficit of early attention.

Camera is an important subject at the time of work with the device Kinect, a performance important to create the 3D geometry [2] is to get the proper resolution, as mentioned Dal Mutto, Carlo Zanuttigh in his book Time-of-Flight Cameras and Microsoft Kinect [3]. The hardware allows you to work with a camera at 30 fps with VGA (640×480) and obtain data of length, width and depth to create a scenario in three dimensions.

Zeng [4] mentions in his article, the importance of the 3D cameras and how they are creating new opportunities for revolutionizing multimedia information. He also

A.B.F. Conrado (✉) · V.D.P. Moya · D. De La Cruz · J. Tobar · P. Mejía
University of the Armed Forces, Sangolqui, Ecuador
e-mail: abfreire@espe.edu.ec; anth_freire15@hotmail.com

V.D.P. Moya
e-mail: vdpazmino1@espe.edu.ec; vinidavid_17@hotmail.com

D. De La Cruz
e-mail: drde@espe.edu.ec

J. Tobar
e-mail: jbtobar@espe.edu.ec

P. Mejía
e-mail: phmejia@espe.edu.ec

S.-I. Ao et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-981-10-2717-8_26

expresses the form in which corporal language is used for video games, saying the sensor that recognizes gestures allows the computer to directly sense the third dimension (depth) of the players and the surroundings, which makes the job much easier.

The use of a hardware as Kinect also requires knowledge of their respective software which allows you to handle all your libraries and functions to program the device, thus giving the importance manipulate the most important factor after the cameras, working with the data obtained, in fact the points of concentration of Kinect call skeleton-programmer, which creates the most basic of the program, this skeleton virtual perform actions and events necessary for the implementation of the applications [5].

The 2012 investigation of Garrido [6] demonstrates the great importance that this gesture recognition technology that Microsoft uses for the development of games for Neuro-rehabilitation for children who have problems with slow learning. Also in this investigation is been mentioned the use of libraries for the development of applications as well as the XNA platform for the programming and the potential this language has. The results demonstrated allow us to have a wide idea of the aspects in which this helps a specific handicap, in which we recognize the importance of these games. Patetsianakis et al. [7] also speaks about another type of technology. The GRS have a group of microphones that allow the recognition of instructions using voice. The monitoring of the sensor of the person, who is utilizing it, requires standards for optimal use, the minimum distance that varies from 6.5 to 7.8 feets allow the GRS to capture the entire body of the player and the information from the persons voice.

The Kinect sensor is a multi-purpose device because of its wide scope, it could be different uses according to the need of the user, Wenbing et al. [8] performs a monitoring for results in the progress of rehabilitation with the use of Kinect, showing a broad overview to allow results verify and quantify the aid in children attention deficit and concentration in their daily activities.

Kinect technology is a device that allows you to develop high-impact therapeutic games in children's fashion that is achieved to stimulate physical and mentally, allows a cognitive development by the game and learning, two activities that apply jointly.

26.2 Influence Area

26.2.1 Factors that Influence the Evolutionary Development of the Child

In the present technological advances and research in the different fields of formation have put the gaze on social needs and have said that the integral attention since the first years of life determines the later stages and the construction of a true personality as the basis of a productive society. Therefore, the timely attention to children at an

early age responds to the wide social conscience, to develop their skills and physical skills, motor and cognitive impairment.

The early development of the child and the initial education refers to the process of change through which children are enabled to handle levels increasingly complex movement, coordination and reasoning, of feeling and expressing emotions and relationships and interact with other people and with the natural environment.

26.2.2 Attention Deficit in Ecuadorian Children

The attention-deficit hyperactivity disorder (ADHD) is a growing problem in Ecuador due to the factors involved since their first years of life, this disorder is characterized by a pattern high of motor activity, impulsivity and difficulties in the performance in the attention. The article by Ramos et al. [9] presented an investigation which had as objective to describe the prevalence of ADHD in a sample of students of Quito-Ecuador, with a methodology epidemiological to establish the prevalence of the disorder in question.

26.2.3 The Education with Games

The education in children been based on the curriculum of early and basic education [10] because they are developed according to the childrens necessity. The use of tools helps in the study fields that the children develop while they are growing up.

Learning in children through the visual-motor development focuses principally on the following study fields:

* The relation between the natural and cultural means.
* Logic-mathematic relations.
* Discovering the characteristics of the elements of the natural world explored through the senses.
* Understanding the basic notions of quantity facilitating the development of thinking skills.

This type of learning allows the identification of the animals characteristics, colors, shapes, outlines and in this way learn to recognize these elements in the atmosphere in a graphic and visual way.

Children respond to visual and ludic methodologies of learning in a better way. This work hopes to provide an innovative didactic tool that contributes to the development of the skills and visual-motor abilities in children between 6–8 years old, through activities and games that involve the total physical movement. In this way, knowledge been obtained and retained through the world that surrounds us. Using new technology that efficiently contributes a decrease in a sedentary lifestyle and an improvement in childrens health.

26.3 Evaluation Techniques

Palou Pere research [11] compares the quality of life between physical activity and sedentary lifestyle that modern children live; we clearly observe how the lack of dynamic games affect the cardiorespiratory health of children.

He mentions that exploring the relation between Quality of Life relating to Health (CVRS), they are important resources for the adaptation of society and healthy development of the child, physically as well as psychological. Due to these reasons, it is of paramount importance the contribution of these ludic games and technologies that will help reduce sedentary lifestyles from an early age.

Presented in this way an algorithm that allows to quantify the development of the activities of the children, as well as the methodology of PC gaming for the development of their skills.

26.3.1 Bootstrap Algorithm

Is an algorithm that allows for the treatment of samples of a versatile by removing a number of the total population for its analysis. Allows for the construction of confidence intervals at the time of approximating the distribution of a statistical process, provides estimates of the statistical error, impose few restrictions on the random variables analyzed and establishing itself as a procedure of a general nature, regardless the statistical considered [12].

The fundamental affirmation bootstrap is that a frequency distribution of statistics calculated on the basis of samples is an estimate of the distribution of sampling of the population [13].

26.3.2 Methodology Using Gesture Recognition Sensor

SDK (Software Development Kit) allows the creation of applications using native methods, and in this way work with physical gesture recognition, allowing for an improvement in the interaction of the program and user [14].

The SDK controls the recognition of the skeletons through the SkeletonFrame Ready, and the recognition of articulations through the points of focus joints. If the sensor does not detect any body, it simply sends the image of depth to the host device, like an Xbox or computer. The software that executes the host device contains the logic to decode the information and recognize the elements in the image with human shape characteristics. To accomplish this objective, the software has been programmed with a wide variety of body shapes, it uses the aligning of diverse parts and movements to identify and monitor them (Fig. 26.1).

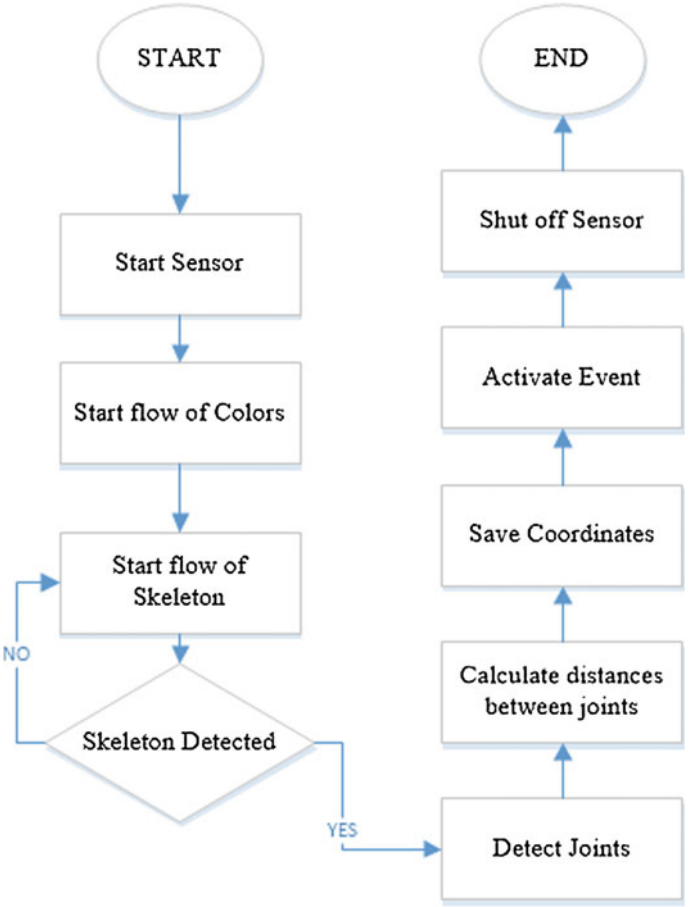


Fig. 26.1 Principal Programing Algorithm. When starting, the sensor immediately activates the control of color flow and the flow of skeleton information. With the last, it is been hoped that the user will position himself in front of the sensor to recognize the joints and calculate the distances between the points of focus. Afterwards, the coordinates are been registered with each of the points, principally the *left* and *right* hand, to execute the predefined event of the action that is being made

26.3.3 Detection of Skeleton

With the use of the SkeletonFrameReady method, a detection of the skeletons can been captured through the sensor on the camera for gesture recognition. The skeletons will been detected, but only those that provide the sufficient information and data, such as hand and head positions, and will be established as the principal or closest to the sensor, maintaining at a margin any other skeleton that may be detected.

26.3.4 Position of Points of Interest

After having the skeleton on screen, the coordinates of point of interest are been shown (hands, feet, head, arms, etc.) to achieve this; we turn to the JointID, which uses the Joints vector to obtain the elements that we want and to be recognized in a precise way (Fig. 26.2).

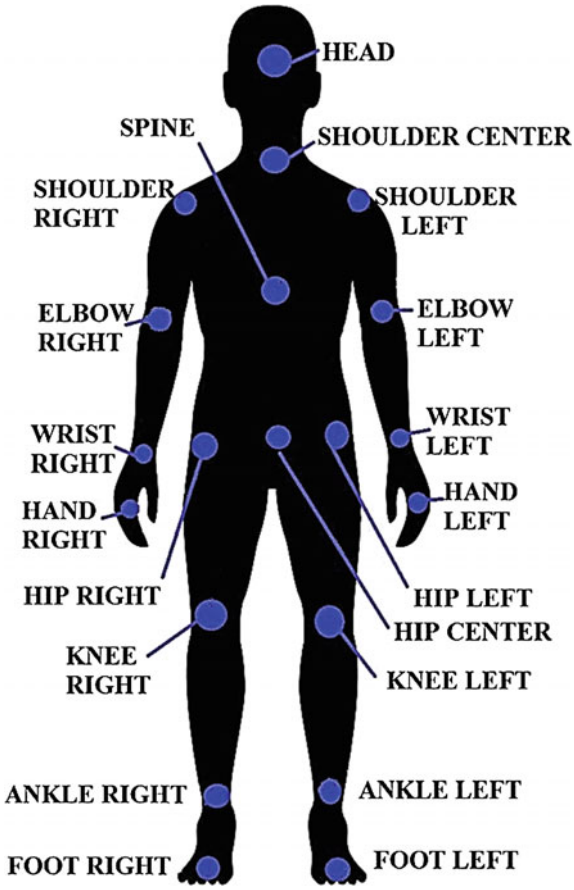


Fig. 26.2 Concentration points in human body. Following the distances and coordinates in (X, Y) of the concentration points of the human body, the recognition of gestures that the body is making is been made for example, lifting a hand. In this case we can measure the distances that exist between the hands be it the *left* or *right*, and the head; and in this way detect if the hand is lifting

26.3.5 Gesture Recognition

26.3.6 Remote Communications Network to Store in the Database

It implements a local network between various computers with its respective Kinect connected to a personal computer that fulfils the function of a remote server, with the ease of which shall not be taken different databases for the monitoring activity of children, this will send the data to the remote server and the data is stored in a single database to have a better control of all the children who are playing at that time.

The database has been implemented in the engine of MySQL databases in the same way only as server without cost, the data for each of the users are filled in an event that ran next to finish each of the games, where the tutors are responsible to qualify those errors had children and will be stored in the database remotely on the computer intended to be server on the LAN.

26.4 Applications Running

This section presents all applications created in gesture recognition sensor (GRS) and clear operation exercises developed and recreational games.

26.4.1 Learning the Colors

26.4.1.1 Description

This option hopes to teach the user about colors, through object selection. In the interface, we see three different color cups as shown in Fig. 26.3, each cup must be placed on the plate according to its corresponding color, after each hand movement over the cup we can select the desired cup and allow to move it and take it to its correct position. You can change hands (between the left and right) to play the game, the cup lands on the corresponding plate, and if placed correctly, a confirmation will appear demonstrating its correct place. Besides appearing on the right hand side, the cups change position randomly with each start of the game.

26.4.1.2 Obtained Results

The result obtained with children from 6–8 years, is that correct interaction with gestures had children, the children were able to correctly select the cup, besides being able to take her proper course, teachers and instructors they were satisfied with the

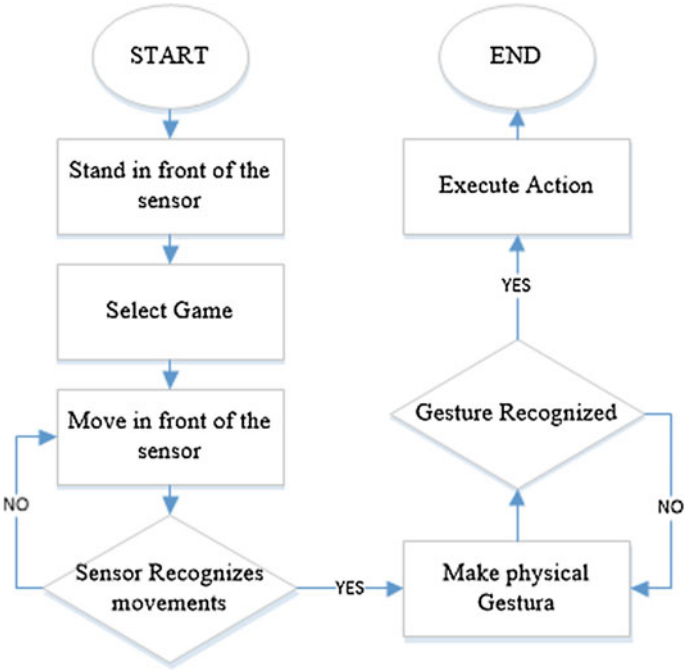


Fig. 26.3 Principal functionality Algorithm. The algorithm that is presented is the basic operation of all applications developed, in which the user is presented in front of the sensor, followed selects the appropriate game and it begins with the movement to be recognized until the program identifies you and finally you perform the corresponding action

game, because it is innovative to have an interface that allows interaction between the child and the objects found on the screen; proper operation of the program was against the rapid movement of children. The children saw fully interactive gambling, they amused themselves by using the game, the children stood, crouched moved, the game improves movement of children and allows them to improve their laterality and learning colors by special methods and teaching that are perfect for the age in which they find themselves (Fig. 26.4).

26.4.2 Learning the Numbers

26.4.2.1 Description

In the demonstrated interface, we can see numbers 1 through 5, which change positions randomly each time the game is started; the user selects the number, whether it be with the right or left hand, if the hand disappears from the screen, the number

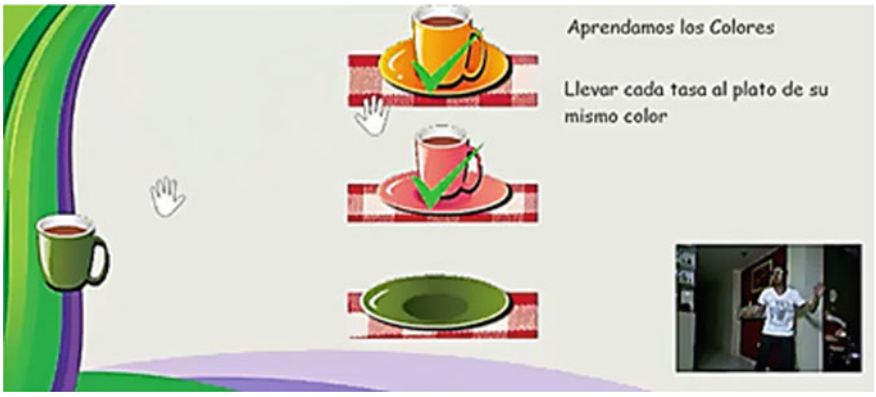


Fig. 26.4 Learning the colors interface

disappears from the list as well. The numbers must be selected in numerological order. Each time the hand passes by a certain number, there will be a number that appears on the right side that represents the number of objects where the hand is currently positioned.

26.4.2.2 Obtained Results

As in any game at first practice it is needed to understand the dynamics and controls in addition to the commands needed to run the game as effectively. The game improvement in children's ability to recognize numbers within images that can hide them, making it difficult to recognize them and at the same time improving the ability of children to recognize hidden objects and expediting their mental and brain capacity; counting the number of objects displayed on the right side of the image improves the ability to count and recognize the number of objects to be present in an image (Fig. 26.5).

26.4.3 Learning the Figures

26.4.3.1 Description

For this game you must find the object that appears only as an outline, within a variety of other objects. With the left or right hand, you must select an image that is similar to the outline on the right and separate the left hand with the band, so we can determine if the chosen object is correct, if the object is correct, the outline will be indicated as correct, otherwise an X is shown. After pressing the continue button, a new outline will be randomly shown so that the game will be more dynamic and

Fig. 26.5 Learning the numbers interface

interactive for the child. The application improves visual-motor abilities as it allows free interaction with the interface without the necessity of a control or device.

26.4.3.2 Obtained Results

Children select the figure that create similar to the shape shown, with this can improve the ability to recognize shapes and to classify or filter within a group of objects, although if you do not have a similarity between them, encourage the child the ability to discern between one object and another, increasing their intellectual capacity as well as their dynamics, laterality through movements made by children to select the object you want. The generated every time a new image allows the child not only based on his memory, but exercise your brain t his intellect against the evidence raised (Fig. 26.6).

26.4.4 Learning Animals

26.4.4.1 Description

Another important feature of the sensor is its phrase and word recognition, with this we can develop an application that allows identification of animals shown on screen. This improves the visual capturing ability of children, as well as contributing to the development of word pronunciation.

Fig. 26.6 Learning the figures interface

Fig. 26.7 Learning the animals interface

26.4.4.2 Obtained Results

The children had great interaction with the Kinect and recognition of sounds, the child is pronounced in syllables, so we decided to take another kind of sentence recognition (Fig. 26.7).

26.4.5 Avatar

26.4.5.1 Description

With the gesture recognition sensor, there is communication between the game development platform and the user. An avatar is implemented to recognize the movements of gestures from the user. With the aim being to call the attention of the child and make it easy to get up and simulate the movements of a super hero, controlling its movements (Fig. 26.8).

Fig. 26.8 Avatar

26.5 Results Analysis Through the Use of the Algorithm Bootstrap

It is considered a population of 30 children from the age of 6 to 8 years through 2 techniques different, the first is the use of techniques conventional and the written tests in order to check the knowledge of children and the second is the use of the tool ludic that allows you to get the number of errors of children from the games raised, in this section use the matlab code with the bootstrap algorithm [15].

26.5.1 Learning the Colors

As you can see in figure, for the game, it is estimated that the average of the population with respect to the percentage of learning with the use of games entertainment, increase in their frequency and therefore represents a percentage of learning the 90.03 % with respect to the use of techniques conventional, as written test that only has a result of learning of 11.12 % (Fig. 26.9).

26.5.2 Learning the Numbers

In the same way, it gets the outcome of the game of the numbers taking into account that increases the frequencies to obtain the highest result of learning from 9 % to 90 %, considering a improvement for the learning of children (Fig. 26.10).

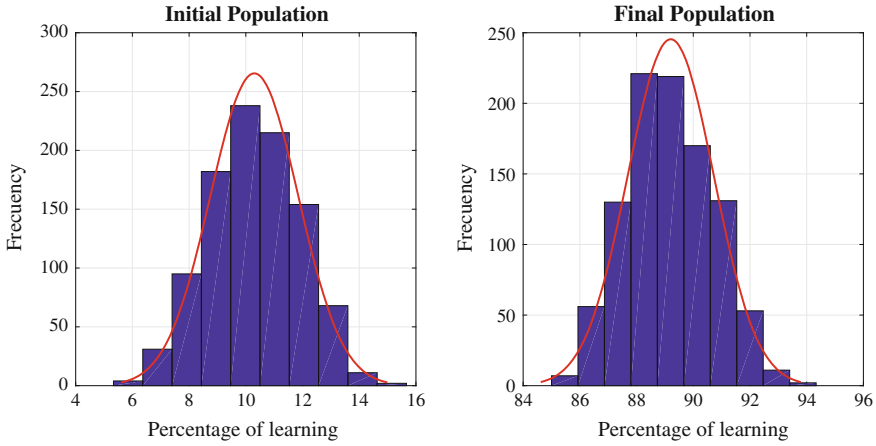


Fig. 26.9 Percentage of learning about the colors game

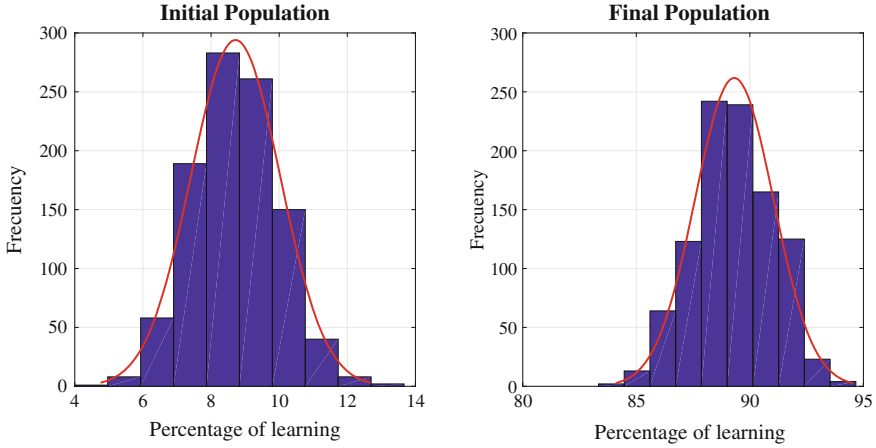


Fig. 26.10 Percentage of learning about the numbers game

26.5.3 Learning the Figures

The game is striking in the sense of handle different figures only standing in front of the sensor, for this game it was considered that the minute game will take measure of the successes that had the child with the interface, although it was a short time, was obtained an increase in the percentage of learning with regard to the use of conventional techniques of 11 % to 80 %, showing the importance of the games for learning (Fig. 26.11).

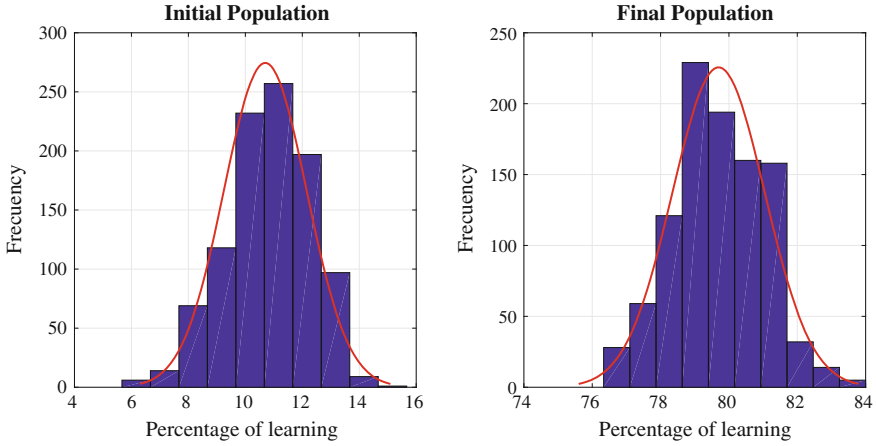


Fig. 26.11 Percentage of learning about the figures game

26.6 Conclusions

- The gesture recognition sensor is an excellent option when involving didactic games due to the fact that it s a strong technological tool not only in hardware and software, thanks to its great versatility to develop applications and programs dedicated to the necessities of the user.
- As has been observed in all the figures that represent the percentage of learning, the use of games to teach different topics such as numbers or colors, exceeds in approximately 70 % to the use of conventional techniques in order to teach these topics, it is recommended to use the technology, as has been demonstrated in the research, help in a large percentage to learning and therefore to the intellectual development of the child in the stage that you need it most.
- Due to the movement detector and gesture recognition being relatively new on market, there can be an implementation of games that are being dedicated to these types of activities, in the case of this Project, ludic games that are oriented towards the development of the physical and emotional abilities of the child.
- The mode of play is dynamic, unlike other ludic games currently available, there is diminished sedentary lifestyle and accessibility for the child at the moment of learning to play.
- Through these games, the form of learning of a child is much more advantageous, as in ages 6 to 8; the activities that involve any movement will help them relating sociably to others as well.
- The use of a LAN for you to deploy a remote database for each one of the players, implies that the teachers have a control over all the data that are coming from each of them, allowing the monitoring and the correct classification of data, considering a more effective solution to implement individual databases in each one of the computers with the games.

References

1. Freire AB, Pazmiño VD, de la Cruz DR, Tobar JB, Mejía PH (2015) Ludic tool that contributes to the physical and motor development based on sensor technology for gesture recognition. In: Lecture notes in engineering and computer science: proceedings of the world congress on engineering and computer science 2015. San Francisco, USA, pp 338–342, 21–23 Oct 2015
2. Khoshelham K (2012) Accuracy and resolution of kinect depth data for indoor mapping applications. Faculty of geo-information science and earth observation, University of Twente. ISSN: 1424-8220
3. Dal Mutto C, Zanuttigh Cortelazzo GM (2012) Time-of-flight cameras and microsoft kinect, springer briefs in electrical and computer engineering. Springer, New York
4. Zeng W (2011) Microsoft kinect sensor and its effect. University of Missouri. Published by the IEEE Computer Society
5. Tao G, Archambault PS, Levin MF (2013) Evaluation of skeletal tracking in a virtual reality rehabilitation system for upper limb hemiparesis. Interdisciplinary Research Center in Rehabilitation, Laval, Canada
6. de la Garrido FD (2012) (6 june of, 2012) Kinect applications for neurorehabilitation. Polytechnic University of Catalonia, Thesis work
7. Patetsianakis E, Ksylakis E, Triantafyllidis G (2014) A kinect-based framework for better user experience in real-time audiovisual content manipulation 2014 international conference on telecomunications and multimedia. Telecom ParisTech, Technological Educational Institute of Crete, Aalborg University
8. Wenbing Z, Espy DD, Reinthal AM, Feng H (2014) A Feasibility study of using a single kinect sensor for rehabilitation exercises monitoring: a rule based approach. Departament of electrical and computer engineering, Cleveland State University
9. Ramos C, Bolaños M, Ramos D (2015) Prevalencia del trastorno por déficit de atención con hiperactividad en estudiantes ecuatorianos. Revista científica y tecnológica UPSE, vol III, N 1, pp 13–19 (Dic. 2015)
10. Mara Cristina Espinoza Salas (2014) Stefanie naranjo oro santacruz. Early Education Curriculum. Ministry of Education
11. Palou P, Vidal J, Ponseti X, Cantallops J, Borrs PA (2012) Relations between quality of life, physical activity and cardiorespiratory fitness in sedentary children
12. Solanas A, Sierra V (1995) Boostrap: fundamentos e introducción a sus aplicaciones. Universidad de Barcelona, n 55, pp 143–154
13. Mooney CZ, Duval RD, Duval R (1993) Bootstrapping: a nonparametric approach to statistical inference. Sage, no 94–95
14. Hacking the Kinect. Autores: Kramer J, Burrus N, Echtler F, Herrera DC, Y M Parker. Editorial: Technology in Action. Chapter 2: Hardware. Page 13
15. Andrea C, Edison S, de la Danni C, Mejía P, Paredes (2016) CLEI Electronic J, 19(1), Paper 4, April 2016. http://www.clei.org/cleiej/papers/v19i1p4.pdf

Chapter 27
Design of Learning Ontology and Service Model Through Achievement Standards Profile Modeling

Hyunsook Chung and Jeongmin Kim

27.1 Introduction

Up until now, most learning support systems have provided collection and storage functions of learning resources and supported the listing and searching of the learning resources on the basis of learning objects, i.e., school subjects. This paper discusses a different perspective from the conventional one, namely, a method of linking the curriculum, teaching/learning plans, and learning resources with each other on the basis of achievement standards [1, 2]. According to the 2009 Revised National Curriculum of Korea, achievement standards is defined as "a statement of the ability and characteristics of knowledge, skills, and attitude that students must achieve through learning, to be presented as a practical basis of teaching/learning and assessment." In other words, achievement standards provide practical guidelines about what has to be taught and assessed by teachers and what has to be studied and achieved by students. They are used as criteria in teaching/learning and assessment according to the characteristics and the aim of the education curriculum [3].

The establishment and enforcement of the national achievement standards started from the 7th Education Curriculum and has been expanded and improved in terms of quantity and quality through the 2007 and 2009 versions of the Revised Education Curriculum [4]. Particularly, the curriculum development policy of the

H. Chung
Department of Computer Engineering, Chosun University, Pilmoon-daero 309, Dong-gu, Gwangju 61452, South Korea
e-mail: hsch@chosun.ac.kr

J. Kim (✉)
Department of Computer Engineering, Daejin University, Hokuk-ro 1007, Pocheon-si, Gyeonggi-do 11159, South Korea
e-mail: jmkim@daejin.ac.kr

S.-I. Ao et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-981-10-2717-8_27

2009 Revised Education Curriculum strengthened the quality of the education curriculum through school curriculum assessment, school subject general assessment, and national achievement assessment. Accordingly, teaching method research, development, and support activities are continuously carried out to develop and deploy the achievement standards and achievement levels by school subject and provide various assessment methods, processes, and tools [5].

The form of content achievement standards for each school subject is a combination of knowledge, skills, and attitudes that students must achieve through learning, and the behaviors and characteristics of learners; it is composed using behavioral terminologies to describe study contents from the perspective of students and clearly reveals the behaviors that must be achieved. For example, "Lesson 9052. When the temperatures of two objects in contact are different, the heat transfer between the two objects to reach thermal equilibrium can be explained" is one of achievement standards defined in the "Matter and energy" unit in middle school science.

The 2009 Revised Education Curriculum was applied to the first and second grades of elementary school and the first year of middle school in 2013, and it will be extended to the other grades and school subjects year by year for the application of the achievement standards curriculum across all grades and subjects by 2016. Therefore, it is necessary to build a system that can support achievement standards-oriented teaching/learning, with the aim to structuralize, share, and interconnect the achievement standards data for all school subjects.

It is necessary to convert the contents of achievement standards distributed in a document format for each subject into a data structure that a machine can process through conversion to data and hierarchical structuralization. However, the conversion into a database through an entity relationship diagram (ERD) design can stop at simply supporting data storage and search. By implementing RDF/OWL-based linked data, an interconnection-oriented data cloud between various types of learning information and resources can be established and an achievement standard–based intelligent teaching/learning service can be implemented. Also, through the connection with foreign achievement standards linked data, advanced services can be implemented, such as comparison, review, improvement, and linkage of achievement standards [6].

This paper is organized as follows. In Sect. 27.2, the concept and trend of linked data are introduced, and in addition, the Achievement Standards Network (ASN), which is the achievement standards repository of the U.S., is introduced. In Sect. 27.3, the comparison between ASN and Korean Achievement Standards is described. In Sect. 27.4, the design and utilization method of the linked data-based achievement standard profile is introduced, and in Sect. 27.5 a conclusion is provided.

27.2 Related Work

Linked Open Data (LOD) or simply linked data, are created by interconnecting data by assigning a dereferenceable URI (Unique Resource Identifier) to them and meaning links (e.g., sameAs) as a method of building a Web of data. In other words, linked open data can be understood as a combination of open data and linked data. Open data mean that the data are open on the Web and anyone can access or reuse them through the HTTP protocol, and linked data are independent data linked by meaning relationships, which can expand and mash up knowledge [7, 8].

Liked open data do not indicate a new technology; they are implemented with Web standard technologies such as HTTP, URI, RDF, and SPARQL, which are conventional Semantic Web-based technologies. The core of the Semantic Web's success lies in the construction of an ontology-based inferable knowledge base, but it is a very slow and complex work that requires much time and effort. Therefore, compared to the vision of the Semantic Web, realistic success cases and uses are very slow and because of this, the Semantic Web is sometimes regarded as a failure [9].

At present, studies are being carried out in South Korea and overseas for methods of opening and linking public data on the basis of linked data and utilizing them industrially. The U.S. and the European Union are already preparing legal and institutional systems for a continuous open policy of public data and supporting uses and reuses in the private sector. South Korea has also shown much interest and effort in the construction of linked data for open and reuses of private or public data [10].

The ASN is a linked data repository of school subject learning achievement standards established by U.S. national and state education offices, and provides the hierarchical structure definition of achievement standards and the mapping relationships between mutually different achievement standards for identical school subjects [11]. Furthermore, it provides browsing and search functions of achievement standards and supports the connection of achievement standards with teaching/learning and assessment. The metadata of the ASN profile defines the relationships between the achievement standards by modeling the K-12 achievement standards document and the statement described in the document, in an RDF format. Figure 27.1 shows the mapping structure of the ASN and explains the mapping between the achievement standards of each state, such as Texas and California, as well as the linkage with learning resources.

At present, the ASN repository stores more than 700 achievement standards documents for K-12 education established in each U.S. state; and also interconnects and stores not only the achievement standards of private education institutions in addition to the federal and state government data, but also the official achievement standards data of national, state, and regional education institutions in Australia. Currently, there are about 340,000 stored and accessible achievement standard items in the ASN repository [12].

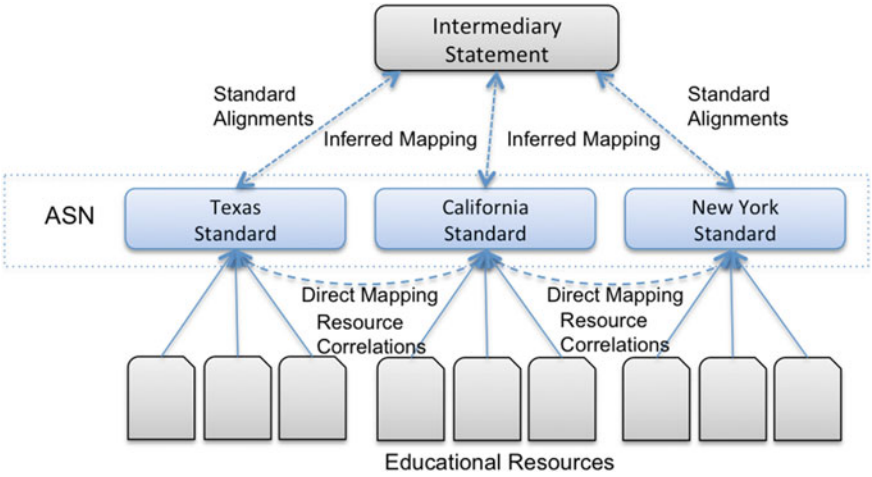


Fig. 27.1 ASN's mapping of achievement standards and linkage with educational resources

In the ASN achievement standards profile, the relationship between the achievement standards document and achievement standards is defined by a hierarchical structure. That is, the <*dcterms:isPartOf*> relationship is defined between the achievement standards and the achievement standards document, and the <*gem:isChildOf*> relationship is defined between the achievement standards, generating a taxon path.

27.3 Comparison Between ASN Framework and Korean Achievement Standards

According to WG4 N2025 'The Achievement Standards Network (ASN)—Framework a Linked Data Approach to Learning Objectives and Educational Content', standards document in terms of curriculum standard and statement in terms of achievement statement have a structural relationship like Fig. 27.2. Through the case study on comparison document scheme between ASN Framework and Korean education system, researchers could find common similarity (in fact, almost same) that achievement statement as set of competency requirements is derived from curriculum standard.

Korean experts tried to reflect the same concept of semantic relationship between outcomes stated in curriculum standard and achievement statement. Because the Korean case study doesn't transform to Linked Open Data format yet, cross subject reference property in achievement statement is just expected relationship, and even if textbook publisher and teachers generally derived outcome statement from achievement statement published by government level, which specific topic is

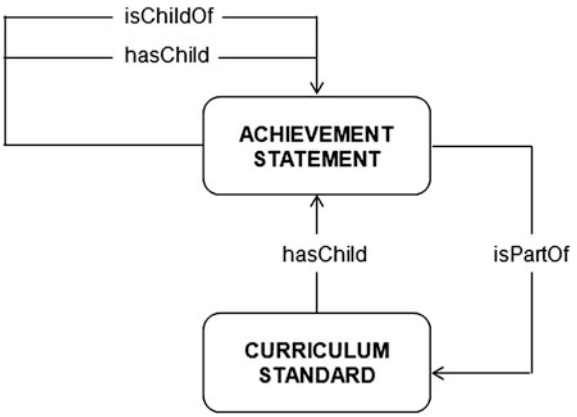
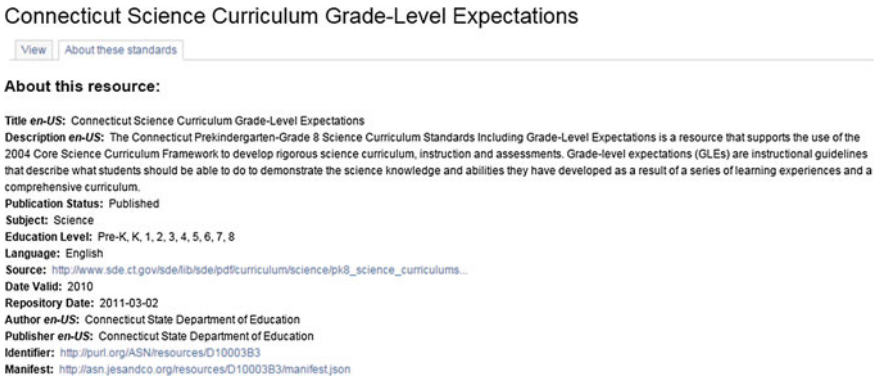


Fig. 27.2 Structural relationships between achievement standards and statements

Properties of Matter — How does the structure of matter affect the properties and uses of materials? PREKINDERGARTEN			
PK.1 - Objects have properties that can be observed and used to describe similarities and differences			
Core Science Curriculum Framework	**Preschool Curriculum Framework**	**Grade-Level Expectations** *Students should be able to:*	**Preschool Assessment Framework**
PK.1.a. Some properties can be observed with the senses, and others can be discovered by using simple tools or tests.	Cognitive Development: Logical-Mathematical/Scientific Thinking - 1. Ask questions about and comment on observations and experimentation; 2. Collect, describe and record information; 3. Use equipment for investigation; 4. Use common instruments to measure things; 5. Demonstrate understanding of one-to-one correspondence while counting; 6. Order several objects on the basis of one attribute; 7. Sort objects by one or more attributes and regroup the objects based on a new attribute; 8. Engage in a scientific experiment with a peer or with a small group.	1. Use senses to make observations of objects and materials within the child's immediate environment. 2. Use simple tools (e.g., balances and magnifiers) and nonstandard measurement units to observe and compare properties of objects and materials. 3. Make comments or express curiosity about observed phenomena (e.g., "I notice that…" or "I wonder if…"). 4. Count, order and sort objects (e.g. blocks, crayons, toys) based on one visible property (e.g., color, shape, size). 5. Conduct simple tests to determine if objects roll, slide or bounce.	**COG 1** Engages in scientific inquiry **COG 3** Sorts objects **COG 5** Compares and orders objects and events **COG 6** Relates number to quantity

Fig. 27.3 Example of Connecticut science grade-level expectations

derived from something in achievement statement is not open to public yet. This is the reason why both properties '*crossSubjectReference*' and '*derivedFrom*' are enclosed in parentheses. In this section, the document scheme comparison is introduced between the source of ASN Framework and Korean curriculum standard and achievement statement. Firstly, we picked up one of the standard of ASN Framework, Connecticut science grade-level expectations, and analyze the structure of the document depicted in Fig. 27.3. All data of the standard are registered into ASN Framework and can be browsed each information on the web such as Fig. 27.4.


Connecticut Science Curriculum Grade-Level Expectations

View | About these standards

About this resource:

Title *en-US*: Connecticut Science Curriculum Grade-Level Expectations
Description *en-US*: The Connecticut Prekindergarten-Grade 8 Science Curriculum Standards Including Grade-Level Expectations is a resource that supports the use of the 2004 Core Science Curriculum Framework to develop rigorous science curriculum, instruction and assessments. Grade-level expectations (GLEs) are instructional guidelines that describe what students should be able to do to demonstrate the science knowledge and abilities they have developed as a result of a series of learning experiences and a comprehensive curriculum.
Publication Status: Published
Subject: Science
Education Level: Pre-K, K, 1, 2, 3, 4, 5, 6, 7, 8
Language: English
Source: http://www.sde.ct.gov/sde/lib/sde/pdf/curriculum/science/pk8_science_curriculums...
Date Valid: 2010
Repository Date: 2011-03-02
Author *en-US*: Connecticut State Department of Education
Publisher *en-US*: Connecticut State Department of Education
Identifier: http://purl.org/ASN/resources/D10003B3
Manifest: http://asn.jesandco.org/resources/D10003B3/manifest.json

Fig. 27.4 Example of Connecticut science curriculum standard and statement

In case of Korea, two kinds of document are used to describe the curriculum standard and a core achievement statement. First, in the curriculum standard area of content per education level (grade group) is decided and achievement criteria and exploration activity per educational level is guided in the document. Second, through the achievement statement document each criteria of the curriculum standard match to statement of the achievement criteria described in Fig. 27.5.

The standard curriculum for science subject depicted in Fig. 27.5 can be interpreted as follows:

- **Title (*en*)**: Science subject curriculum
- **Description (*en*)**: *(omission)*
- **Publication Status:** Published

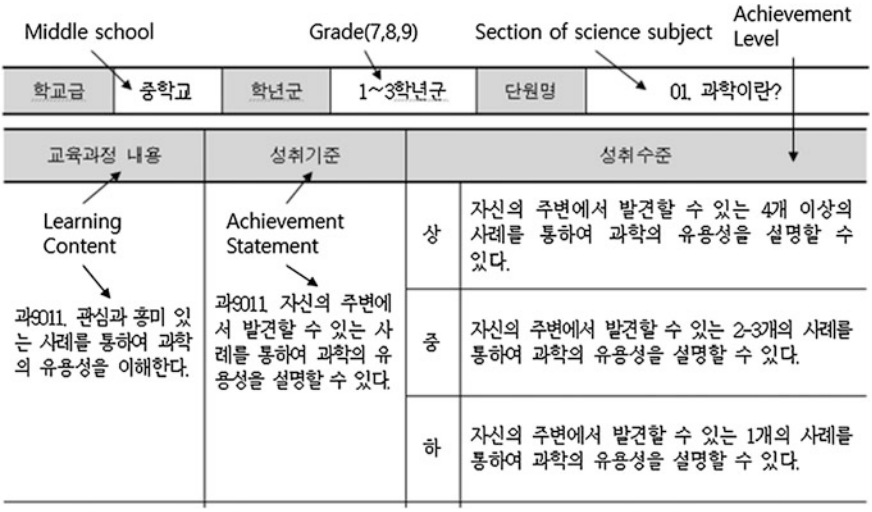


Fig. 27.5 Example of curriculum and achievement criteria for science subject (in Korea)

- **Subject:** Science
- **Education Level:** K-3, 4, 5, 6, 7, 8, and 9
- **Language:** Korean
- **Source**: http://ncic.re.kr/nation.dwn.ogf.inventoryList.do
- **Date Valid:** 2011
- **Repository Date:** 2012-12
- **Author (*en*)**: Ministry of Education
- **Publisher (*en*)**: Ministry of Education

In summary, through comparison for document scheme of countries researchers could get insight for similarity of concept model and captured entities and properties to describe curriculum standard and achievement statement. To reach and/or develop International Standard level, this kind of comparison is required for evaluation for the concept and terminology.

27.4 Achievement Standards Profile Modeling

In this section, we design a learning ontology, which conceptualizes curriculum, achievement standards, syllabus, and bloom taxonomy. In addition, we provide a service model for adaptive learning path generation based on the achievement standards ontology.

27.4.1 Class Definition

Our learning ontology is comprised of for concept groups of entities, such as curriculum, achievement standards, syllabus, and bloom taxonomy. In our model from Fig. 27.6, *KnowledgeArea*, *Subject*, *CourseType*, *Course, Instructor*, and *LearningResource* classes are included in the curriculum concept group. *StandardsDocument*, *Statement*, and *AchievementLevel* classes are included in the achievement standards concept group. The syllabus concept group includes *Syllabus*, *LearningSchedule*, *LearningEvaluation*, and *LearningActivity* classes. Finally, the bloom taxonomy concept group has *BloomTaxonomy*, *BloomKnowledge*, *BloomSkill*, and *BloomAttitude* classes.

From the internal structure analysis of Korean achievement standards documents in Sect. 27.3, we have identified three classes for describing metadata of the documents themselves and achievement statements represented in the documents. These three classes are specified in Tables 27.1, 27.2, and 27.3 respectively. Following list shows the description of some classes in the learning ontology.

- **KnowledgeArea:** describes an academic field of curriculum.
- **Subject:** is a part into which learning can be devided.

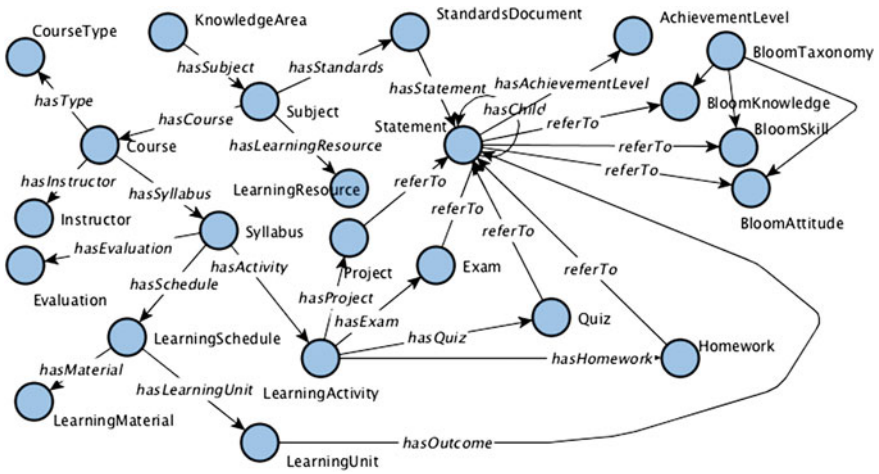


Fig. 27.6 Classes and relationships of learning ontology

Table 27.1 StandardDocument class

Description	Definition
Class name	StandardDocument
Class label	Achievement standards document
Class type	n/a
Class URI	http://kr.kasn.org/KASN/schema/StandardDocument
Child class	n/a
Explanation	It refers to an achievement standards document defined for each pertinent school subject. It is a class for specifying a document itself that describes the achievement standards, achievement levels, and assessment methods for each school subject. It defines the metadata of the document and a link to the actual document as attributes

Table 27.2 Statement class

Description	Definition
Class name	Statement
Class label	Achievement statement
Class type	n/a
Class URI	http://kr.kasn.org/KASN/schema/Statement
Child class	n/a
Explanation	It is a class that defines an individual achievement standard defined in the achievement standards document. The achievement standards consist of curriculum content achievement standards and detailed achievement standards. Since more detailed and specific achievement standards are defined when going down to a child level, this class has a self-inclusion relation to define the hierarchical structure of achievement standards

Table 27.3 AchievementLevel class

Description	Definition
Class name	AchievementLevel
Class label	Achievement level
Class type	n/a
Class URI	http://kr.kasn.org/KASN/schema/AchievementLevel
Child class	n/a
Explanation	It is a class that defines the achievement level of achievement standard. The achievement level is defined in three types of "good/fair/poor". Considering the expandability, good/fair/poor is defined as an instance of this class instead of as an attribute. That is, for one achievement standard, three (good, fair, poor) achievement standard instances are set up at maximum

- **Standards Document:** is a description of an achievement standards document.
- **The statement:** describes achievement statements defined in the achievement standards document.
- **AchievementLevel:** is a description about good, fair, and poor achivement levels for each achievement statement.
- **Syllabus:** describes teaching and learning plans of a course through referring to the related achievement statements.
- **LearningUnit:** describes a topic to be educated in an education course. It has one or more statements as learning outcomes.
- **LearningActivity:** describes the activities of students for learning and training in an educational course. Each activity has one or more statements to achieve after completion of the learning activity.
- **Bloom Taxonomy:** describes the learning objectives in terms of knowledge, skill, and attitude classification.

27.4.2 Property Definition

For each class of achievement standards, we identified critical literals from achievement standards documents and define data properties as described in Tables 27.4, 27.5, and 27.6.

Table 27.7 shows the relationships between Statement class and the other classes in learning ontology. The relationships, *hasChild* and *isChildOf*, are used to form a hierarchical structure of statements. The *LearningActivity* class has *referTo* relationship to the *Statement* class in order to describe learning outcomes of activities.

Table 27.4 Properties of StandardDocument class

Property	Type	Description
standardID	String	Unique identifier of standards document
standardTitle	String	Title of standards document
curriculumTitle	String	Course title of standards document
publicationStatus	String	Publication status of standards document(published, draft, revised)
jurisdiction	String	Jurisdiction controlling standards
createdDate	Date	Published date
linkedDate	Date	The date of registration into the linked data repository
docLink	URI	URI of the stored standards document
comment	String	Comment

Table 27.5 Properties of Statement class

Property	Type	Description
statementID	String	Unique identifier of a statement
statementNotation	String	Statement code notation
authorityStatus	Core/Gen	Core or general statement
conceptKeyword	String	Concept words defined in a statement
chapterTitle	String	Chapter or section title of a statement
statement	String	Statement
bloomTaxonomy	URI	Knowledge, skill, or attitude related to a statement
localSequence	String	Sequence no
comment	String	Comment

Table 27.6 Properties of AchievementLevel class

Property	Type	Description
achiLevelID	String	Unique identifier of an achievement level
statementURI	String	URI of a statement referenced by an achievement level
achiLevelLabel	Good/fair/poor	Category of an achievement level
achiLevelContent	String	Description of an achievement level
comment	String	Comment

27.4.3 Adaptive Learning Service Modeling

In this section, we design an adaptive learning method based on achievement standards as one of services using our learning ontology. To enable adaptive learning of students, our approach sets interlinks between learning activities and achievement statements as well as between learning units and achievement statements at first. For example, each question on a midterm exam should be connected to the specific statement to represent that the question measures whether the related

Table 27.7 Relationships between Statement class and the other classes

Relationship	Domain	Range	Direction
hasStatement	StandardDocument	Statement	→ Statement
hasChild	Statement	Statement	→ Statement
hasOutcome	LearningUnit	Statement	→ Statement
referTo	Homework Project Quiz Exam	Statement	→ Statement
hasAchievementLevel	Statement	AchievementLevel	Statement →
isChildOf	Statement	Statement	Statement →
isPartOf	Statement	StandardDocument	Statement →
referTo	Statement	BloomKnowledge BloomSkill BloomAttitude	Statement →

achievement is fulfilled or not. Likewise, other activities, like assignments, quiz, and term projects, should be connected to the related statements to measure the degree of learning achievement. Figure 27.7 shows an instance of the interconnection of statements and learning activities and learning units. In this figure, learning objectives denote achievement statements and learning subject denotes learning units.

We present mathematical expressions of the data model for implementing the relationships between statements and learning activities and between statements and learning units in expressions, (27.1), (27.2), and (27.3). In expressions (27.1) and

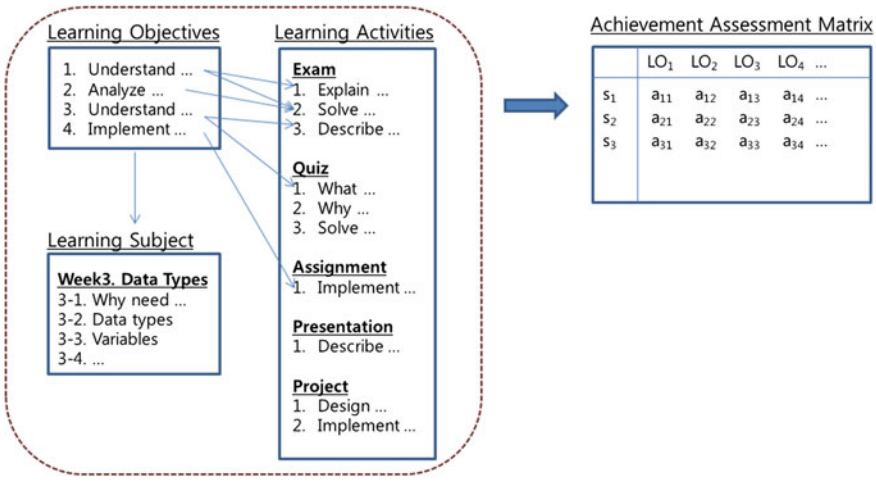


Fig. 27.7 Achievement standard-based adaptive learning

(27.2), s_i denotes an achievement statement. The symbols, a_i and u_i, denote a learning activity and a learning unit respectively. Expression (27.3) is a matrix in which each element is a normalized value ranged 0 and 1. Each value represents an achievement score of a student.

$$SLA = \{(s_i, a_i) | 1 \le i \le n, 1 \le j \le m\} \tag{27.1}$$

$$SLU = \{(s_i, u_k) | 1 \le i \le n, 1 \le k \le l\} \tag{27.2}$$

$$EVAL = \begin{bmatrix} x_{11} & \cdots & x_{1n} \\ \vdots & \ddots & \vdots \\ x_{m1} & \cdots & x_{mn} \end{bmatrix} \tag{27.3}$$

After creation of a matrix of each student, we evaluate the degree of achievement of learning units for each student by identifying the number of scores below the pre-determined threshold. Learning units with lower achievement scores can be suggested to the student for learning again. Following Table 27.8 shows an algorithm for identifying learning units with lower scores than the threshold from the matrix.

From the achievement matrix evaluation, we can figure out students who have a lower achievement in their learning. These unfulfilled students can be guided to study the unachieved topics by the generation of adaptive learning sequence. The adaptive learning sequence can be generated in different granularity. The high-level learning sequence can be created using syllabuses and their relationships. The

Table 27.8 An algorithm for identifying learning units, which have lower achievement scores

```
ALGORITHM 1. Finding learning units with low scores
    procedure findLearningUnits
    begin
           threshold = get_threshold()
           list_of_units = nil

           for all scores in matrix M
                  if M[i,j] < threshold
                        add_unit(list_of_units, M[i,j])

           align_units(list_of_units)
           suggest_learning_sequence(list_of_units)
    end
```

Table 27.9 Definition of binding rules for RDF/OWL specification

Description	Definition
KnowledgeArea	<owl:Class rdf:about = "http://kasn.org/schema/core/KnowledgeArea"/> <kasn:KnowledgeArea>
Subject	<owl:Class rdf:about = "http://kasn.org/schema/core/Subject"/> <kasn:Subject>
StandardDocument	<owl:Class rdf:about = "http://kasn.org/schema/core/StandardDocument"/> <kasn:StatementDocument>
Statement	<owl:Class rdf:about = "http://kasn.org/schema/core/Statement"/> <kasn:Statement>
AchievementLevel	<owl:Class rdf:about = "http://kasn.org/schema/core/AchievementLevel"/>
standardTitle	<owl:DatatypeProperty rdf:about = "http://kasn.org/schema/core/dataproperty/#standardTitle"> <rdfs:domain rdf:resource = "http://kasn.org/schema/core/#StandardDocument"/> <rdfs:range rdf:resource = "http://www.w3.org/2001/XMLSchema#string"/> </owl:DatatypeProperty>

low-level learning sequence can be created using learning units and their relationships defined under *Syllabus* class in the learning ontology.

27.4.4 Linked Open Data Specification

In this section, we define a mapping rule for specifying RDF/OWL documents in order to build the KASN Linked Open Data repository. First, we define some namespaces to reference to upper ontologies as follows:

- xmlns:asn = "http://purl.org/ASN/schema/core/"
- xmlns:dc = "http://purl.org/dc/elements/1.1/"
- xmlns:skos = "http://www.w3.org/2004/02/skos/core#"

Second, we define a mapping rule shown in Table 27.9. This mapping rule is used for transforming entities in KASN class model into the constructs of the RDF/OWL syntax.

27.5 Conclusion and Future Work

The achievement standards linked data profile is a framework that supports the development of various teaching/learning support services based on an open approach of conversion, storage, and search with machine-readable expressions (RDF/OWL) with respect to the achievement standards and achievement levels of each elementary/middle/high school subject provided by educational institutions and organizations. This paper examined the linked data, on which interests have been focused recently, with the method of open data policy and introduces a method of producing linked data of teaching/learning achievement standards.

The national education curriculum requires explicitly providing the standards that students must achieve for each curriculum component, going beyond a simple method of listing the curriculum components; planning a class and supporting the learning activities according to these standards; and determining the achievement levels through fair assessment. Therefore, basically, a teaching/learning support system implemented in the future has to be able to implement a service model based on the achievement standards of the national education curriculum. To this end, the achievement standards linked data profile should be continuously expanded and elaborated.

References

1. Yu D, Zhang W, Chen X (2006) New generation of e-learning technologies. In: Proceedings of first international multi-symposiums on computer and computational sciences, pp 455–459
2. Chi Y (2010) Developing curriculum sequencing for managing multiple texts in e-learning system. In: Proceedings of international conference on engineering education
3. Yun H, Park S, Lee K (2008) A study on achievement standards of curriculum. RRC2008-2, Korea Institute for Curriculum and Education
4. Lee D, Kwak B, Choi S, Hur K, Cho N, Park S, Hong H, Kim J (1997) A study on curriculum development based on 7th revision of national curriculum. Research Report CR97-36, Korea Educational Development Institute
5. Park S, Beak K, Lee K, Han H, Lee S, Lee W (2013) Development of core achievement standards of k-9 curriculum based on the 2009 revised national curriculum. Korea Institute for Curriculum and Education
6. Chung H, Kim J (2015) Design of achievement standards linked data profiles connecting curricula and digital content. In: Lecture notes in engineering and computer science: proceedings of the world congress on engineering and computer science 2015, San Francisco, USA, pp 302–307, 21–23 Oct
7. Bauer F, Kaltenböck M (2011) Linked open data: the essentials. Vienna: edition mono/monochrome, p 62. ISBN 978-3-902796-05-9
8. Sletten B (2013) Resource-oriented architecture patterns for webs of data. Synth Lect Semant Web: Theory Technol 3(3):1–95
9. Berners-Lee T, Linked data: design issues. http://www.w3.org/DesignIssues/LinkedData.html
10. National Information Agency (2014) Korea linked open data practices

11. Sutton SA, Golder D (2008) Achievement Standards Network (ASN): an application profile for mapping K-12 educational resources to achievement standards. In: proceedings of international conference on dublin core and metadata applications, pp 69–79
12. Sutton SA, Golder D, Phipps J (2010) Global linking of educational resources through learning objectives, Going Global 4, London, March 24–26

Chapter 28
Web Retrieval with Query Expansion: A Parallel Retrieving Technique

Noha Mezyan and Venus W. Samawi

28.1 Introduction

The World Wide Web (WWW) is a considerable information source for people. It contains billion of pages created by millions of authors and organizations [1]. To find specific information on the Web, people usually input a query into a search engine. The search engine will sift web pages to find pages relevant to user query. Onset, this approach has worked well, where the amount of information contained in the Web wasn't huge, but for the time being, the whole situation is different. The WWW contains millions of web pages that are unstable in terms of size or in terms of content's updating. It is important to declare that search engines highly depend on Information Retrieval (IR) techniques.

IR is concerned with finding documents relevant to user needs among huge collections of documents [2]. Combining IR, Artificial Intelligence (AI), and data mining techniques provides a generic framework for web mining. Accordingly, web information retrieval (WIR) based on web content mining is considered an appropriate technique to extract information from the internet [3, 4].

Web mining discovers and extracts information from the web documents in order to improve the degree of relevance of the retrieved Web pages when a query is entered to a search engine [3, 4]. Much research used GAs to improve the search engine performance, and help in retrieving most relevant documents. In [5], a genetic search strategy (called GeniMiner) for a search engine is proposed. Two

N. Mezyan (✉)
Department of Computer Science, Al-Albayt University, P.O. Box 130040, Mafraq 25113, Jordan
e-mail: mezyan_n@yahoo.com

V.W. Samawi
Department of Computer Information System, Amman Arab University, P.O. Box 2234, Amman 11953, Jordan
e-mail: venus@aau.edu.jo

S.-I. Ao et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-981-10-2717-8_28

fitness functions are used to evaluate web pages. GeniMiner retrieves relevant information, but it consumes longer time compared with other standard search engine. Therefore, this approach is recommended to be used as a complementary to standard search engines. Vallim et al. [6] suggest using a personal agent that mines web information sources and retrieves documents according to user's query. A GA that learns the importance factors of HTML tags which are used to re-rank the documents retrieved by standard weighting schemes was proposed in [7]. SCAIR is the retrieval engine used in this work. Vizine et al. [8] introduce an automatic keyword extraction method and a GA to improve the web search. The web mining frame work in [9] was again based on genetic search. The fitness function they used in their work depends on the link quality F(L) and the page quality F(P) to calculate the mean quality (M_q). In [10], a GA based strategy is designed for finding association rules without user-specified threshold for minimum support. The relative confidence is used as fitness function. The researchers claim that the computation cost is reduced, and interesting association rule is generated. In [11], GA is applied to perform web content mining. The used fitness function was support probability ratio of each individual. The researchers propose performing cross-over using "And" operator between two selected individuals. In [12], GA for search over XML of different domains is investigated. A steady state tournament selection Microbial is used. Thada and Jaglan in [13] use GA in WIR. Mountford coefficient is used for WIR. The researchers claim that GA is fruitful in searching and ranking the retrieved documents according to their relevancy to the user query. Yet, most previous researches did not investigate the role of fitness function and genetic operators (selection, cross-over, and mutation) and their effect on WIR process from accuracy (i.e. the degree of relevance of the retrieved documents to the user need) point of view. On the other hand, although search engines sift huge number of web pages to respond to user query, a lot of relevant pages may not be reached. Another significant point is how to lessen the time needed to reach the pertinent pages. This work tackles the problem of retrieving most pertinent documents to user needs (query) in short time, and how to boost user queries. To fulfill these goals, island genetic algorithm (IGA) is used to achieve parallelism (each island is executed independently on different server) to improve retrieving time and to diversify the results. In this work, four islands are implemented. Each island has its own specifications (fitness function and selection method) to study the effect of these specifications on the accuracy of the obtained results. Moreover, using different specifications will supports the idea of variation in results (i.e. each island will retrieve different set of documents). Although search engines help in discovering resources on the Web, but it does not satisfy users with information needs when users enter incomplete queries. Automatic query expansion technique is used to solve such problem and to enhance number of retrieved documents by each island. The output of each island is passed to a decision making stage (DMS), where the obtained results (retrieved documents) from the four islands are combined and ranked (using cosine similarity measure). Finally, the proposed approach allows the user to specify similarity threshold. The documents with similarity measure $\geq$ threshold will be retrieved. In this work, threshold = 0.8. The rest of

chapter organized in 5 sections. In Sect. 28.2, island genetic algorithm is illustrated. Methodology of the suggested approach is explained in Sect. 28.3. The assessment of experimental results and evaluation are clarified in Sect. 28.4. Finally, we concluded in Sect. 28.5.

28.2 Island Genetic Algorithms (IGAs)

Various approaches were applied to make simple Genetic Algorithms run faster. One of the most promising approaches is the employment of parallel implementations (IGAs). Island genetic algorithm is a distributed model of genetic algorithm, where each island executes GA independently with its sub-population [14, 15]. In IGA, each island evolves independently using its own sub-population, and could have different genetic operators [14, 15]. Consequently, each island follows a different searching path over the search space to obtain genetic diversity (i.e. each island might reach different solutions). Moreover, the parallel nature of the IGA accelerates the process of documents retrieval. These are the main reasons for using IGA in retrieving the pertinent documents from the WWW. As well known, simple GA starts with a single population, while, IGA starts with set of populations. Each genetic island generates its population, which differs from the other islands. The main steps of a simple GA (selection, crossover, and mutation) are utilized in each subpopulation (island) independently. In IGA, chromosomes (individuals) are allowed to migrate to another island [16]. The main steps of simple GA are [17]:

1. Specify appropriate coding for each chromosome (individual).
2. Generate a population (set of randomly generated chromosomes).
3. Repeat steps 4, 5 until the termination condition is satisfied
4. Find the fitness value of each chromosome.
5. Apply crossover and mutation on the nominated chromosomes to generate the new population.

Various selection operators could be used to nominate chromosomes to next round. In this work, random tournament selection and unbiased tournament selection are used in addition to elitism [18]. Elitism is used to pass the best individual directly to the next generation to save it from loss during the reproduction operation.

28.3 Methodology

The main motivation behind this research is the need of quick search engine to retrieve relevant documents from tremendously rich knowledge repository (such as WWW). In this work, four islands are suggested to be used for retrieving

documents (as shown in Fig. 28.1). Each island has its own selection method and fitness function to achieve genetic variation. IGA could be implemented in parallel (on different servers) to speed up the searching process through the WWW. In this work, we did not perform chromosome migration between islands to reduce traffic loading, and to study the behavior of each island independently. Finally, the results achieved by the four islands are ranked and combined (using a proposed decision making approach) to retrieve the most relevant web pages. The proposed approach mainly consists of four phases (see Fig. 28.1): tokenizing phase, preprocessing phase, indexing phase, and document retrieving phase.

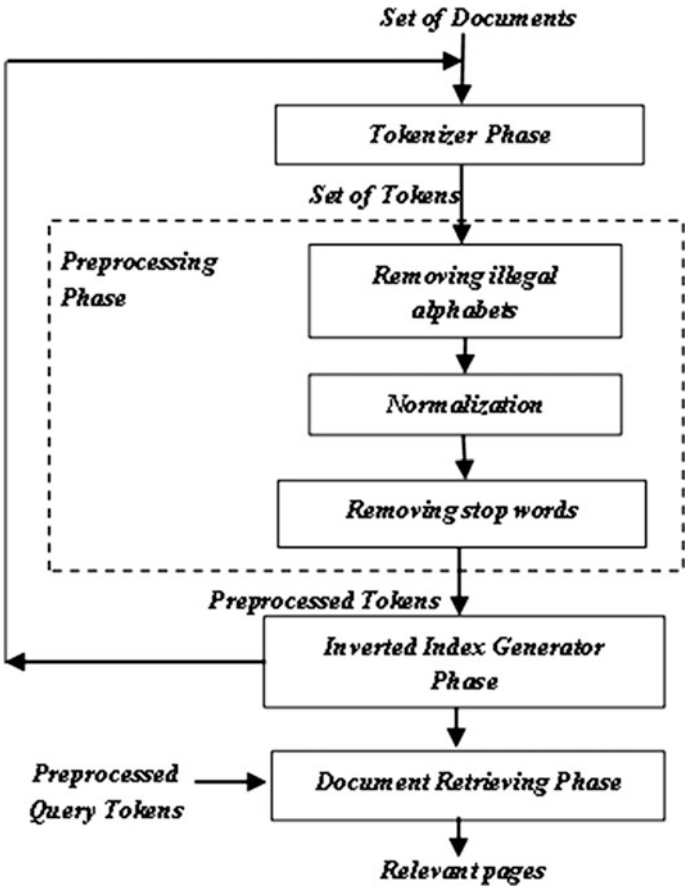


Fig. 28.1 The proposed system phases

28.3.1 Preprocessing Phase

Query preprocessing is very important phase in any information retrieval system. In this research, the following preprocessing tasks are performed:

1. *Removing illegal alphabets and irrelevant information*: every language consists of set of alphabets. In this step, remove each alphabet $\notin$ to the language alphabets (in this research, alphabet of English language). Furthermore, it is necessary to remove information that is not a part of document main topic (such as banner ads). It is awkward to be indexed as it may negatively affect the accuracy of the retrieving process.
2. *Stop-words* removal: stop words help in constructing sentences but have slight value in document retrieving (such as, about, a, is, for, the, what, etc.). Such words must be removed.
3. *Normalization (case unification)*: convert all letters to lower case.

28.3.2 Inverted Index Generator Phase

In WIR, it is not possible to scan all web pages every time to find which page contains the query terms. Alternatively, a data structure, called inverted index, is used to speed up the searching process. In this work, an inverted index table is build, where each term (token) is associated with a list of all documents that contain that term or its derivatives.

28.3.3 Document Retrieving Phase

As stated before, the proposed retrieving module depends on the IGA to attain parallelism and accelerate the retrieving process. It is mainly consist of four islands, each of these islands works independently on different server. Therefore, each island could use different index tables. In this work, to compare the behavior of the four islands under the same circumstances the same inverted index is used with the four islands. Figure 28.2 illustrates the document retrieving phase.

In this phase, the user enters the query into the proposed system. The query is tokenized and preprocessed to determine the main terms (keywords) that will be used to retrieve the pertinent documents. These terms will be passed to the genetic search stage. In this stage, an initial population of chromosomes is randomly generated, where each chromosome is a string of 0's and 1's representing the presence or absence of a query term in a document. Consequently, the length of chromosomes for different queries varies depending on number of words in the query (keywords). Each chromosome indicates set of documents (relevant or

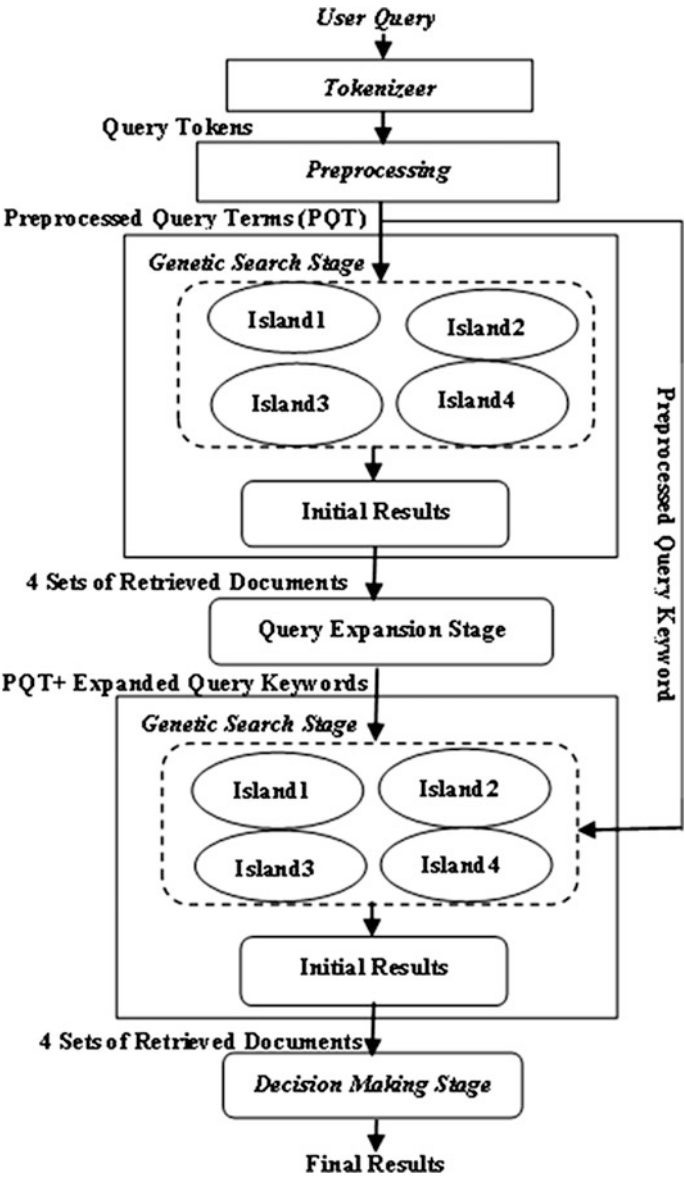


Fig. 28.2 Document retrieving phase

irrelevant) to the entered query. The fitness function determines the degree of relevance for the population depending on the number of query's keywords that occurs in a document. The next step is selecting the suitable individuals, based on their fitness values, and apply one-point crossover operator to generate new individuals, which represent the population of the new generation. Elitism is an

important process in genetic search, which is used in the four suggested islands. The best two chromosomes are passed to the new generation directly. Mutation is applied to some individuals into avoid premature convergence. In this work, mutation is performed by bit flipping.

28.3.3.1 Genetic Search Stage

This stage activates the four islands. The four islands use the same operations concerning one-point crossover operator, elitism, and mutation on individuals. On the other hand, the islands use different selection methods and fitness functions. In this work, random tournament selection method is used since it is easy to be implemented in parallel environment, and it needs insignificant time complexity [18]. In random tournament selection, k individuals (in this work, k = 2) are randomly chosen, and the one with the highest fitness value is selected to be a parent of the new individual in the next population. Tournament selection could suffer from selection bias [18]. Therefore, unbiased tournament selection is used as alternative selection method, to overcome selection bias problem. The algorithm of unbiased tournament selection is simple. For population of size S (in this work, S = 30), generate k random permutations in the range [1,S]. Arrange the k permutations as table of k rows and S columns. Each cell contains the fitness value of the corresponding individual. Compare the fitness value in each column (for the k individuals), and specify the individual with the highest fitness value per column. This will ensure that the best individual is selected k times in the mating, and worst one is not selected [18]. The used fitness functions are Jaccard's coefficient in Eq. (28.1), and Ochiai coefficient (also called Ochiai-Barkman coefficient, or Otsuka-Ochiai coefficient) [19] in Eq. (28.2).

$$J(X, Y) = \frac{|X \cap Y|}{|X \cup Y|} \tag{28.1}$$

$$O(X, Y) = \frac{|(X \cap Y|)}{\sqrt{|X| \times |Y|}} \tag{28.2}$$

where X is the query and Y is a document. Both are represented by weightless keywords (terms). |X| indicates number of elements in X. Example (1 and 2) in [20] explains how to calculate the fitness value of each chromosome according to Jaccard's and Ochiai's coefficient.

As in Jaccard's coefficient, the value of Ochiai's coefficient $\in [0, 1]$, where 0 denotes the absence of all query-keywords from the document, and 1 denotes the presence of all query-keywords in the document, and any other value between 0 and 1 denotes the presence of some of the query-keywords in the document. The specifications of the four islands are shown in Table 28.1.

Table 28.1 The specifications of the four islands are

Island #	Selection method		Fitness function	
	Random	Unbiased	Jaccard's coe.	Ochiai's coe.
1	X		X	
2	X			X
3		X	X	
4		X		X

After entering the user's query to the system, the genetic search stage is activated, where the four islands will search independently and simultaneously (as in Fig. 28.2). The last generation will generate the initial results, which will be fed to the decision making phase. As the optimal solution is not always reached, the proposed system offers an important option to the user, which is enables the user to specifying the fitness value (i.e. acceptable relevance degree to retrieve documents) of the last generation's individuals, which will be considered to compute the initial results. The initial results will be used to perform query expansion, before invoking the second genetic search stage. At the end of the second genetic search stage, a collection of documents, which represent the final result of the proposed system, are obtained. These documents are ranked depending on the degree of relevance to the entered query. In this research, vector space model is used to assign a weight to each indexed term using Eq. (28.6). The degree of relevance is calculated using cosine similarity method (see Eq. (28.7)). Equations (28.3) and (28.4) are used to find the frequency of term i in document j (tf_{ij}), and *term frequency-inverse document frequency* (idf_i) respectively. In Eq. (28.4), N represents the total number of document in the collection; df_i is the number of documents that contain the term t_i. |V| is the vocabulary size of the collection.

$$tf_{ij} = \frac{f_{ij}}{\max\{f_{1j}, f_{2j}, \ldots, f_{|V|j}\}} \tag{28.3}$$

$$idf_i = \log \frac{N}{df_i} \tag{28.4}$$

The weight for documents is calculated as in Eq. (28.5).

$$w_{ij} = tf_{ij} \times idf_i \tag{28.5}$$

In vector space model, queries q has the same representation of documents. Thus, the weight of each term i in the query q is calculated as in Eq. (28.6).

$$w_{iq} = \left(0.5 + \frac{0.5 f_{iq}}{\max\{f_{1q}, f_{2q}, \ldots, f_{|V|q}\}}\right) \times \log \frac{N}{df_i} \tag{28.6}$$

$$\cos ine(d_j, q) = \frac{\sum_{i=1}^{V} w_{ij} \times w_{iq}}{\sqrt{\sum_{i=1}^{V} w_{ij}^2} \times \sqrt{\sum_{i=1}^{V} w_{iq}^2}} \tag{28.7}$$

After assigning a weight for the indexed terms, which are presented either in the document or in the query, Eq. (28.7) is applied to calculate the relevance degree between a document and a query using cosine similarity.

28.3.3.2 Query Expansion Stage

One of the important applications of text mining which is an instance of Web content mining is query expansion, which is in turn an application of text clustering. Query expansion is a very helpful technique used to deal with the user query that consists of few keywords which may not cover all of the relevant documents in order to improve the retrieval performance of the original query [21, 22].

In this research, the initial results—which are the output of the first genetic search phase—are the input to the query expansion phase. Each term in each resulted document—from the first genetic search phase—is scanned, in order to find the term with the highest repetition among the resulted documents. Indeed the term will not be added to the original query unless it is presented in half of the resulted documents at least, since this frequency indicates the relevancy of the term to the retrieved documents according to the original entered query. Many terms may satisfy the condition of being presented in half of the resulted documents after entering the original query or even more, but the term with the highest frequency is the chosen one. After extracting the most frequently presented term from the resulted documents, it is added to the original query to construct a new set of keywords. The genetic search phase is invoked again, were the new expanded set of keywords is fed to the second genetic search phase (i.e. to the same four islands mentioned before). At the end of the second genetic search phase, a collection of documents resulted from the four islands, which represent the final result of the proposed system, are ranked based on the degree of relevance (calculated using cosine similarity measure (see Eq. 28.7)) to the entered query. The ranked documents are fed to the final phase (decision making phase) to filter these documents according to their degree of relevance to the query.

28.3.3.3 Decision Making Stage (Filtering)

This stage is the final stage. As mentioned before, each island of the four islands works independently and generates its own results. Therefore, the results of the four islands will be combined. The proposed system contains a decision making stage (DMS) that merges the results of the four islands, eliminates any duplications (in case of duplication, keep only the document with the highest rank), then order the retrieved documents depending on their ranks (cosine values). Figure 28.3

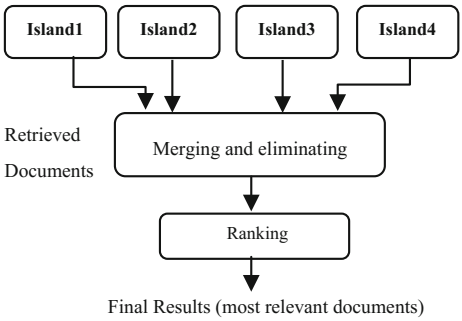


Fig. 28.3 Decision making stage

illustrates the filtering process. The suggested approach allows the user to specify similarity threshold. All documents with similarity measure ≥ threshold will be retrieved. In this work, threshold is set to 0.8.

28.4 Assessment of Experimental Results

To evaluate the proposed system, 5-fold cross validation is applied on Cranfield dataset (Cranfield dataset contains 1400 documents). Twelve queries (out of original 225 queries) are chosen for simulation purposes (Q_2, Q_3, Q_5, Q_8, Q_{28}, Q_{34}, Q_{38}, Q_{45}, Q_{47}, Q_{71}, Q_{203}, Q_{204}). These queries are chosen due to the variation in number of terms and covering different topics. Documents and queries are preprocessed, as mentioned previously. The inverted index is built depending on terms extracted from these documents. To study the behavior of each island, the same initial population is used, in addition to fixing the population size (population size = 30). As stopping condition, number of generations is set to 20. The small number of generations is due to the population size, chromosome (individual) size, and the search space. Threshold fitness values of chromosomes in the last generation are set by the user. Crossover and mutation probabilities ($P_c = 0.8$ and $P_m = 0.1$). The behaviors of the four islands are compared from number of retrieved documents (with cosine similarity ≥ 0.8). Sometimes such comparison is not enough since number of retrieved documents does not reflect the good or bad behavior of the retrieving system. Therefore, another comparison is made, which shows the average cosine similarity measure of the results of the four islands (before and after Query expansion). Figure 28.4 shows the average number of retrieved documents with degree of relevance (similarity ≥ 0.8) by each of the four islands in addition to DMS. It is clearly seen that when query expansion is used, number of retrieved documents increase since new keywords are added to the original query, which may lead to retrieve additional documents.

Figure 28.5 illustrates the average similarity measures with and without Query expansion. The behavior of $sland_3$ with query expansion on query Q_5, did not retrieve any document with similarity ≥ 0.8. This result affects the average

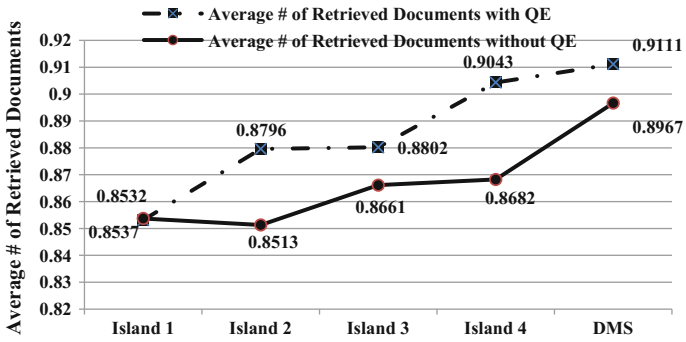


Fig. 28.4 Average number of retrieved documents with similarity ≥ 0.8

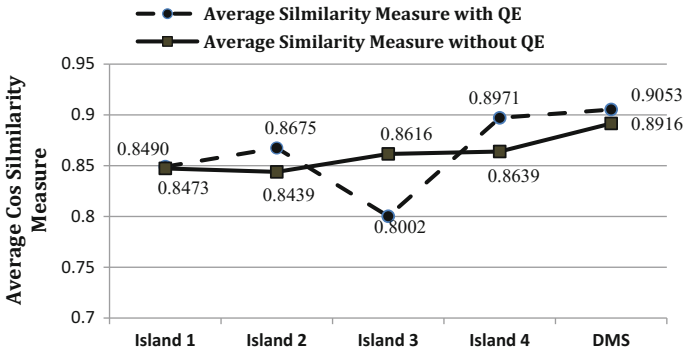


Fig. 28.5 Average similarity measure of 11 queries, similarity ≥ 0.8

performance of island$_3$. This shows that applying query expansion sometimes reduces number of retrieved documents for a certain relevance degree. This could be caused if the added term is confusing or ambiguous. Therefore, it is very important to choose appropriate query expansions technique that adds terms to enrich user query.

To answer the question "which island behaves better?", Figs. 28.4, and 28.5 illustrate that island$_4$ shows the best behavior from both average number of retrieved documents and the average cosine similarity measure of the retrieved documents (the best documents for 11 out of 12). Island$_1$ and island$_2$ both use the same selection method (random tournament selection), but island$_1$ uses Jaccard's coefficient as fitness function, while island$_2$ uses Ochiai's coefficient. Also, islands$_3$ and island$_4$, both use the same selection method (unbiased tournament selection), but island$_3$ uses Jaccard's coefficient as fitness function, while island$_4$ uses Ochiai's coefficient. By comparing the behavior of the two islands, island$_1$ shows better performance. By comparing island$_3$ and island$_4$, island$_4$ shows better performance.

This indicates that, Ochiai's coefficient shows better performance when used with unbiased tournament selection methods, while Jaccard's coefficient have better performance when used with random tournament selection method. Finally, by comparing the behavior of the suggested approach (DMS) with the behavior of the four islands, it is clearly seen that DMS outperforms the behaviors of the four islands (see Figs. 28.4 and 28.5). The average cosine similarity measure of the retrieved documents using DMS is approximately 0.9. The average number of retrieved documents by each island and DMS is illustrated in Fig. 28.6, while the cosine similarity of 11 queries out of 12 is shown in Fig. 28.7. The results of Q_{38} are very weak (only $islands_1$ and $island_4$ retrieve one document with cosine similarity measure ≥ 0.8). Therefore, it is not included in the figures. The results shown in Figs. 28.6 and 28.7 are constructed as follows:

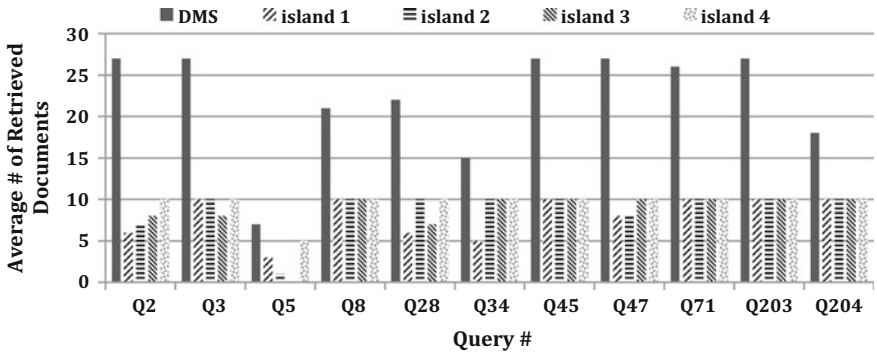


Fig. 28.6 Average # of retrieved documents for 11 query, similarity $\geq$ 0.8

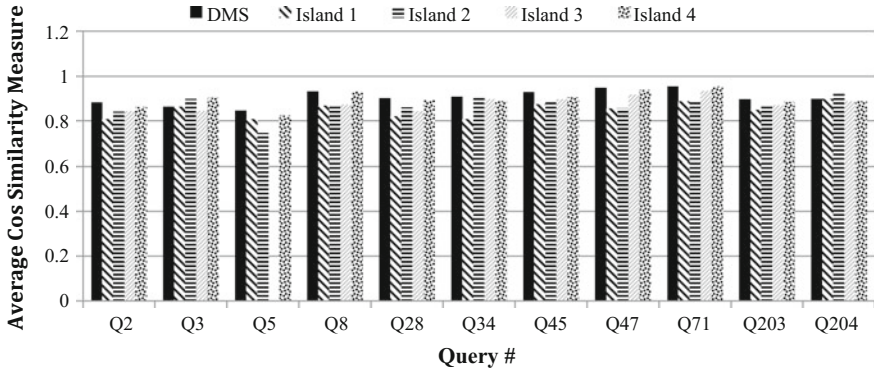


Fig. 28.7 Average similarity measure 11 queries, similarity $\geq$ 0.8

- For each island, rank the retrieved documents, and specify 10 documents with the highest similarity measure. From these 10 documents, get the documents with similarity measure ≥ 0.8.
- *For DMS*, merge the four set of retrieved documents (resulted from 2), and rank these documents.

The suggested system utilizes the genetic diversity nature of IGA. Therefore, the suggested DMS approach outperforms the behaviors of four islands. When applying DMS, more relevant documents (with similarity measure ≥ 0.8) are retrieved compared to the number of documents retrieved by each island separately (for the same similarity measure) (see Figs. 28.4a, d and 28.6). On the other hand, Figs. 28.5 and 28.7 show that DMS improves the type of retrieved documents (i.e. retrieve documents more pertinent to user needs), which evince the main objective of this work.

28.5 Conclusion and Future Work

When a user enters a query to a search engine, millions of web pages are sieved to find the pertinent pages. Consequently, the search engines might be incompetent to retrieve the most pertinent documents (or pages) in a short time. Therefore, the suggestion was using IGA since this might maintain genetic diversity, which increases the probability of reaching most pertinent documents in a short time. In this work, four islands with various fitness functions and selection methods to facilitate studying the behavior of each island and specify the technique most suitable for web retrieving. Each island will retrieve different set of relevant document, which proves genetic diversity in IGA. To achieve parallelism and improves retrieving time, the for islands work independently on different servers. The suggested approach, DMS, utilizes the genetic diversity nature of IGA. From Figs. 28.4, 28.5, 28.6 and 28.7, it is clearly seen that DMS outperforms the behaviors of the four islands from both, number of retrieved documents, and the degree of relevance of the retrieved documents to the user query, point of views. Figure 28.4 and 28.6 shows that by applying DMS, more pertinent documents are retrieved compared to the number of documents retrieved by each island separately. On the other hand, Figs. 28.5 and 28.7 show that DMS improves the type of retrieved documents (i.e. retrieve documents more relevant to user needs), which evince the main objective of this work. Query expansion technique is used to overcome the problem incomplete query, and to enhance the suggested system performance (Figs. 28.4 and 28.5) clearly show this point. As future work, semantic features will be used in additional islands to improve the relevance degree. Also, chromosome migration could be applied to check the migration effect on the speed and accuracy of the retrieving process.

References

1. Chen H, Chau M (2004) Web mining: machine learning for web applications. Ann Rev Inf Sci Technol 38:289–329
2. Manning CD, Raghavan P, Schütze H (2008) Introduction to information retrieval. Cambridge University Press
3. Herrouz A, Khentout C, Djoudi M (2013) Overview of web content mining tools. Int J Eng Sci (IJES) 2:1–6
4. Johnson F, Gupta SK (2012) Web content mining techniques: a survey. Int J Comput Appl 47:4–50
5. Picaroungne F, Monmarché N, Oliver A, Venturini G (2002) Web mining with a genetic algorithm. In: 11th international World Wide Web conference, Honolulu, Hawaii, 7–11 May 2002
6. Vallim MS, Coello JMA (2003) An agent for web information dissemination based on a genetic algorithm. In: International conference on systems, man and cybernetics. IEEE Press
7. Kim S, Zhang B-T (2003) Genetic mining of HTML structures for effective web-document retrieval. Applied intelligence, vol 18. Kluwer Academic Publishers, pp 243–256
8. Vizine AL, de Castro LN, Gudwin RR (2005) An evolutionary algorithm to optimize web document retrieval. In: International conference on integration of knowledge intensive multi-agent systems
9. Marghny MH, Ali AF (2005) Web mining based on genetic algorithm. In: Cairo: AIML 05 conference, pp 82–87
10. Yan X, Zhang C, Zhang S (2009) Genetic algorithm-based strategy for identifying association rules without specifying actual minimum support. Elsevier, Expert Systems with Applications, vol 36, pp 3066–3076
11. Sabnis V, Thakur RS (2013) GA based model for web content mining. IJCSI Int J Comput Sci Issues 10:308–313
12. Tungar D, Potgantwar AD (2014) Investigation of web mining optimization using microbial genetic algorithm. J Eng Res Appl 4:593–597
13. Thada V, Jaglan V (2014) Use of genetic algorithm in web information retrieval. Int J Emerg Technol Comput Appl Sci 7:278–281
14. Whitley D, Rana S, Heckendorn RB (1997) Island model genetic algorithms and linearly separable problems. In: Evolutionary computing: proceedings of AISB workshop, Lecture notes in computer science, vol 1305, pp 109–125
15. Belal MA, Haggag MH (2013) A structured population genetic-algorithm based on hierarchical hypercube of henes expressions. Int J Comput Appl 64:5–18
16. Engelbrecht AP (2002) England, Computational intelligence: an introduction. Wiley
17. Simon D (2013) Evolutionary optimization algorithms. Wiley
18. Yu X, Gen M (2010) Introduction to evolutionary algorithms. Springer Science & Business Media
19. Choi D, Kim J, Kim P (2014) A method for normalizing non-standard words in online social network services: a case study on twitter. In: Vinh PC et al (ed) Context-aware systems and applications, Lecture notes of the institute for computer sciences, social informatics and telecommunications engineering, vol 128. Springer International Publishing, pp 359–368
20. Mezyan N, Samawi VW (2015) Web information retrieval using island genetic algorithm. In: Lecture notes in engineering and computer science: proceedings of the world congress on engineering and computer science 2015, 21–23 Oct, 2015, San Francisco, USA, pp 325–330
21. Liu B (2007) Web Data mining: exploring hyperlinks, contents and usage data. Springer, New York
22. Rivas AR, Iglesias EL, Borrajo L (online) Study of query expansion techniques and their application in the biomedical information retrieval. Sci World J 2014 (2014). http://dx.doi.org/10.1155/2014/132158

Chapter 29
Enhanced Action Based Trust Model: A Multi-facet Model for Social Media

Manmeet Mahinderjit Singh and Yi Chin Teo

29.1 Introduction

In recent years, online social networks (OSNs) have been integrated with several popular websites such as LinkdedIn, Twitter, Google+. The positive impact of this integration is the high numbers of users worldwide, which employs these sites to share ones personal life and career advancement, keeping in touch with their friends, exchanging files, pictures and opinions and pursuing employment opportunities. Since the profiles in OSNs contain a high amount of confidential personal information, user-identity privacy attack [1] which threaten a user online's and physical well being is possible [2]. Traceability and linkability of users through updates and information shared on these websites and network jeopardize the confidentiality and integrity principles revolving this technology [2].

Privacy attack mentioned have a severe impact towards the trustworthiness of the social networking field as whole. In theory, trust refers to the willingness of a party in taking risks and to reduce debt to the lowest amount [1, 3, 4]. The properties of trust can be examined as (i) *Asymmetry (*As two friends have different belief and may have seen different behavior from each other, so trust is not identical); (ii) *Transitivity* (If A trusts B and B trusts G, it does not necessarily follow that A has to trust G); (iii) *Context Dependent (*Trust level towards an individual can be varied based on time, situation and experience) and *Personalized (*Trust is subjective. One can have different opinions regarding the trust level towards a same

M. Mahinderjit Singh (✉) · Y.C. Teo
The School of Computer Sciences, University Sains Malaysia,
11700 Gelugor, Malaysia
e-mail: manmeet@usm.my

Y.C. Teo
e-mail: tychin.ucom11@cs.usm.my

S.-I. Ao et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-981-10-2717-8_29

person) [1]. One requirement of trust models is to provide annotations of trust with confidence and without constraint.

In the field of Computer Science, several trust management models were designed through various algorithms and systems such as User-Centric Personalized Trust model [1], TISoN [5], REFEREE [6], Sporas [7], SECURE [6] and Marsh's trust model [8]. In capturing the subjective view of trust across large populations in OSN, Quinn's multi-faceted model concept has been adopted in this research. Quinn's trust model has the features of being user-centric, personalized and context dependent. Trust synonyms in the core of Quinn's model feature eight main attributes such as honesty, faith, belief, confidence, competency, credibility, reputation and reliability. The research that implements Quinn's multi-faceted in the applications of OSNs is further proof by Chieng et al. [1] with the proposed prototype MiniOSN. Based on further study done on MiniOSN by Mahinderjit Singh et al. [9] is there is no indication or ranks of the attributes listed in multi-faceted model. As a result; the survey conducted could not be generalized and the result may be argued to be biased towards the sample size and their background on the ranking of trust attributes may bias to sample size & background. Without any means to analysed the result by using the statistic tool, finding of a study done by Mahinderjit Singh et al. [10] demonstrates honesty as the most important attribute to determine one's trustworthiness is a weak determinator. As the ranking was established by adopting correlation analysis method, all eight trust attributes were ranked accordingly [10]. The adoption of the ranked trust attributes in an enhanced action based trust model will be shared in this paper.

Hence, the aim of this research is to enhance a trust model that adopts action based trust algorithm [11] by incorporating different weights of trust attributes. The contribution of this research is to present a hybrid action based trust model with a multi - facet model which can be employed in the OSN. The outcome of this research is a new trust model known as Enhanced Action-based Trust (EABT) model. The outline of the paper is as the following. Section 29.2 describes the state of art. Section 29.3 covers the design of miniOSN 2.0. Section 29.4 discuss on the proposed enhanced Action based trust with multi-facet model or EABT model and Sect. 29.5 focus on the user-acceptance of this proposed technique. Finally; conclusion and future work are presented in Sect. 29.6.

29.2 State-of-Art on Trust Management in OSN

29.2.1 Trust Management in Current OSNs

Modern OSNs like Facebook, LinkedIn, Twitter and etc. allow users to create a network to represent their social ties. These OSNs facilitate uploading of multimedia content, various ways of communication, and allowed many users to share

aspects of daily life with friends. As a user share their confidential content, they must trust the OSNs service to protect personal information even as the OSNs provide benefits from examining and sharing that information. As the consequences, OSNs has contributed to substantially increase towards the study of technology trustworthiness. Trust has been recognized as the important factors for success in obtaining user's preferences in using the OSN sites [1, 9, 10]. In present time, the features below are adopted in the trust model in social networking [1]:

(a) *Single-faceted*: Among many trust concepts, only one of them has been used to explain and define trust and to form a single-faceted model of trust so that it can back trust based on decision making that is too general, while many other significant notions of trust have been neglected.
(b) *Not personalized:* Trust in the authentic world relies on context and people are not judged by others similarly as the weight of trust traits are different too. Nevertheless, no personalized concept is prevented in the nature of the current trust model. In this concept, subjective nature and the opinions about human's trust toward people in a large population is noticed.
(c) *Trust level cannot be annotated and calculated:* In the present OSNs, friendship has not been considered in an appropriate category. Therefore, it is not possible to give a good explanation about the trust level towards various people in a context. It is not possible to compute it either. Hence, the trust value on each 'friend' is being uniformity with lists or categories, while it cannot be distinguished based on the percentage of trustiness and the way the user weighted the importance of trust traits. Overall the listed features of current OSN above are seen as the reason as per-why trust issue in OSN is worth to be tackled further. Next, some existing trust management models are discussed.

29.2.2 Exisiting Trust Management Models

29.2.2.1 Online Reputation Models

The reputation based trust [1, 12] used in most online marketplaces like eBay [13] or Amazon [14] is based on the ratings that users perform after completion of a transaction. The reputation value is computed as the sum of those ratings over the last six months for eBay. Similarly, Amazon use also a mean of all ratings to assign a reputation value. They do not provide explicit mechanisms to deal with users that provide false information. The only way to increase the reliability of the reputation value when a tremendous number of opinions that reduce false or biased information.

29.2.2.2 Marsh Trust Management

The trust model proposed by Marsh [8] only considered direct interaction. It differentiates three types of trust which are basic trust, general trust, and situational trust.

- Basic trust. Models the general trusting disposition independently of who is the agent that is in front. It is calculated from all the experiences accumulated by the agent. Good experiences lead to a greater disposition to trust, and vice versa. The author uses the notation $T\,x\,t$ to represent the trust disposition of agent x at time t.
- General trust. This is the trust that one agent has on another without taking into account any specific situation. It simply represents general trust on the other agent. It is noted as $T\,x(y)t$ representing the general trust that agent x has on agent y at time t.
- Situational trust. This is the amount of trust that one agent has in another taking into account a specific situation. The utility of the situation, its importance and the 'General trust' are the elements considered in order to calculate the 'Situational trust'.The basic formula used to calculate this type of trust is: $T\,x(y,\alpha)\,t = Ux(\alpha)t$ t $\times\ Ix(\alpha)t \times Tx(y)t$^ where x is the evaluator, y the target agent and α the situation.

29.2.2.3 Multi-context Trust Model

The mathematical core of this model leans on a theory, distributed by Marsh [8] in his founding thesis introduces so-called contexts of trust which represent the fields which are capable of trusting the entity. Dividing trust into contexts is the only reasonable way to comprise a thing as complex as trust while maintaining the possibility of flexible changes and further development. Every context is normalized into the interval from 0 to 1 to facilitate future aggregation. There are seven different trust contexts stand on functionality provided by Facebook that discussed in [15]. Among the context are (i) Interaction time span (S); (ii) Number of interactions(N); (iii) Number of characters (C); (iv) Interaction regularity (F); (v) Photo tagging (P); (vi) Group membership (G) and (vii) Common interests (L). These seven contexts should be aggregated in a way which allows us to establish an order relation. The equation below is introduced in this model to serve as a priority vector of number where Tx represents the priority for given context.

$$\mathrm{P} = (TS, TN, TC, TF, TP, TG, TL)$$

The final value of trust can be calculated with this formula below:

$$Tx = S.Ts + N.TN + C.TC + F.TF + P.Tp + G.TG + L.TL\,S + N + C + F + P + G + L$$

This method of aggregation enables to rank each context with its importance. Similarly, a completely new context may be added to the existing set and this expansion is also planned in the nearest future in [15].

29.2.2.4 Trust Inference for Social Networks (TISoN)

One of the computational trust models like Trust Inference for Social Networks (TISoN) [5] was introduced a hybrid model which implementation are based on algorithm and mathematical model. Hamdi et al. [5] introduce TISoN model to generate and evaluate trust value helps user and allow them to rate each other without any interactions. His designed a novel Trust Path's Searching(TPS) algorithm to discover the reliable trust path in a large social network then use trust inference measure(TIM) to decide how much the user will trust another (Table 29.1).

29.2.2.5 Action-Based Trust Model

In [11], the authors proposed a new trust model based on what type of content user disclosure in OSN and what action performed by the user, examples like commenting, liking, sharing a post, and tagging on an image, posting a video and so on. An algorithm is designed to calculate trust value on the basis of the actions performed by the user which lead to users from being aware in sharing sensitive content in OSNs. If the trust value of a user is showing a constant low value over some period of time, then he is suspected to be involved in malicious activities. This algorithm would be amended to consider Quinn multi-faceted trust model in OSN in order to discover a appropriate way to compute the trust.

Table 29.1 Trust models analysis

Model name	Context dependent	Mathematical/ algorithm model	Formula of model
eBay [13]	No	Based on feedback	N/A
Marsh [8]	Yes	Mathematical	$T_x(y,\alpha)^t = U_x(\alpha)^t \times I_x(\alpha)^t \times \widehat{T_x(y)^t}$
Multi-context trust [15]	Yes	Mathematical	$T_x = \frac{S.T_S + N.T_C + F.T_F + P.T_p + G.T_G + L.T_L}{S+N+C+F+P+G+L}$
TISoN [5]	N/A	Hybrids	For TIM: $t_{o\to S} = S_{MTP} \times t_{a_n MTP \to S}$
Action-based trust [11]	N/A	Algorithm	Shown in Figs. 29.3 and 29.4.
Trust network based Context Aware recommender system [16]	Yes	Mathematical	$P_{c,i} = \bar{r}_c + \frac{\sum_{p\in M} sim(c,p)(r_{p,i} - \bar{r}_p)}{\sum_{p\in M} \lvert sim(c,p)\rvert}$ $T_{c(p,i)} = 1 - \frac{\lvert P_{c,i} - r_{c,i}\rvert}{Z_{\mathrm{MAX}} - Z_{min}}$ $\mathrm{WT}_{(\mathrm{c,i})} = T_{(\mathrm{c,i})}\left[X + Y \cdot \frac{\sum_{q=1}^{m} W_q}{\sum_{q=1}^{r} W_q}\right]$

29.2.2.6 Multi-facet Trust Model with Ranked Trust Attributes

Quinn et al. [15] found that utilizing only one trust attribute in a single-faceted approach is an inadequate model of trust for use in internet environments. Current trust model "tend to use a single synonym, or definition in the use of trust such approaches can only provide a generic, non-personalised trust management solution". To address this problem of the lack of potential for personalizing trust management, a multi-faceted model of trust that is both personalisable and specialisable was proposed by Quinn et al. which can satisfy large and board population. In [15], myTrust Trust Management system defined trust into concrete concept and abstract concept with attributes of their own, where the former includes credibility, honesty, reliability, reputation and competency attributes, and the later with belief, faith and confidence attributes. Ratings are then given to each of the eight attributes, and trust is calculated as the weighted average of these ratings. In addition to demonstrating its personalization capabilities, Quinn et al. demonstrated how the model could be specialised to any application domain. Chieng et al. [1] has then adopted Quinn et al. work for the field of social networks. They proposed a user-center personalized trust model by designing a proof of concept of a small scale OSN named miniOSN. However, both Quinn's Model and Chieng's model do not provide a trust calculation mechanism that can be used. The work of Teo and Mahinderjit Singh [10] has further ranked the trust attributes by employing survey mechanism, data on all eight trust attributes was collected. Next, by using tool such as correlation coefficient, the relationship between these attributes were designed and the ranked values of each attributes was produced. The result is tabulated in Fig. 29.1. The correlation coefficient for faith is the highest among other trust attributes which is 0.92, followed by confidence and belief which is 0.91 and 0.90. The difference of a correlation value between reputation and credibility is merely

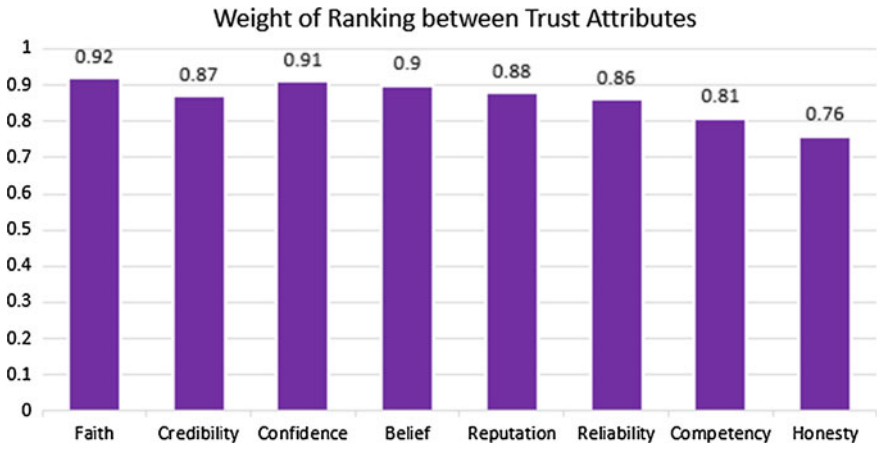


Fig. 29.1 Ranking of importance among all trust attributes

0.01 that is 0.88 and 0.87; followed by 0.86 and 0.81 which are reliability and competency. The lowest correlation value among these attributes are honesty which is 0.76. The value of correlation coefficient for all trust attributes with trust value is very close with each other.

From the result shown in Fig. 29.1, the authors concluded that faith is the most important in multi-faceted model of trust, confidence is ranked as the second most important.

29.2.2.7 Context Aware Recommender Model

Based on the research by Dutta et al. [16], a trust based recommender systems leveraged by context attributes, recommender system aims at solving the problem of information flooding for web applications is designed. The recommender system proposed is a trusted network based context aware recommender system that takes into the considerations trustworthiness of the recommending partners and context information such as time, location and company of a person along with the user and item. According to the author, the accuracy of the recommender output, enhanced when the most relevant contexts are selected and their weightages are appropriately taken incorporates with aspects of dynamic trust. In this case, trust is dynamic in natural and not a static parameter.

In this research; both Multi-facet trust model with ranked trust attributes and Action-based trust model will be adopted and enhanced further. The fundamental of multi-facet trust model with ranked features will be embedded into the enhanced algorithm.

29.3 Multi-facet Trust Management Model with Ranked Design

By adopting Chieng's trust model [1] with the influences from Quinn's multi-faceted model of trust, the author believed that this trust management model can be applied in OSN. The reason is due to the fact that users are abe to express their subjective views on trust. Therefore in miniOSN, trust is defined into abstract and concrete concepts, with eight trust attributes, namely, credibility, honesty, reliability, reputation, competency, belief, faith and confidence. In this trust management model, ratings to all eight trust attributes depending on the Fig. 29.1 in accordance.

The miniOSN inhibit the following characteristic:

(1) Subjective view of trust

In this model, we enable the user to express the subjective views of trust based on the eight trust attributes, user can give values to each attributes towards his every friends. Our aim is to provide flexibility for the user to rate. The user is able to edit

their trust level towards their friends from time to time, since trust is context dependence.

(2) Assignment of trust level and independent value

The calculated average rating values (x) for each of the friend of a particular user based on eight trust attributes are recorded in a log data table. The trust level and the independent value of a friend are depend on the set of range of values of x at Table 29.2 below. When the value of x dropped in between the value of 7.5–10, the trust level is considered as high and independent value is 4, if it is lower than 7.5 and higher than 4.5, the trust level is equal to medium, below 5.0 and above 2.5 is low and below than 2.5 is non-trusted together with independent value 1, respectively. The independent value will be used as a critical element in order to perform correlation analysis.

(3) MiniOSN 2.0 and Ranked Trust Attributes

From the result of *this correlation coefficient and* among eight of the trust attributes, faith have the most impact to the users when rating a friend, and followed arranged in an ascending order of ranking, which are confidential, belief, reputation, reliability, credibility, reliability, competency and honesty respectively. As in Chieng et al. [1]; Mahinderjit Singh et al. [9] research work, trust attributes in multi-faceted model as influenced by Quinn [15] considered the wattage for each attribute is equally important in which the weightage are set as the *default value. In minus 2.0*, appropriate weighting for each trait *are obtained through* correlation analysis *method. T*he weight for each trust attributes would be defined for different values as according to the ranking of attributes. Users in miniOSN 2.0 are able to view the trust value and the trust level of their friends. The trust value of each friend which is inserted is then divided with the ranked values as a whole. In miniOSN 2.0 users are given the option to be able to personalize the control of the accessibility of the posts and photos by setting the values according to eight trust attributes or the trust level of the post.

29.4 Proposed of Enhanced Action-Based Trust (EABT) Model

Action-based Trust algorithm [11] is chosen as it calculates trust value on the basis of the actions performed by the user which lead to users from being aware in sharing sensitive content in OSNs. Based on the proposed computational algorithm,

Table 29.2 Trust level scale with different range of trust value

Standard of trust been set	x, weighted average of trust attribute	Independent value
High	$7.5 \leq x \leq 10.0$	4
Medium	$5.0 \leq x < 7.5$	2
Low	$2.5 \leq x < 5.0$	3
Non-trusted	$0 \leq x < 2.5$	1

the value of trust for a user in OSN depends on every each action that he performs for example, like a photo, share, post a status, etc., then compute the value of trust for an user, this value referred as a trust factor by authors in [11]. Trust factor of a user may increase or decrease depends on the category of content the user interact with, which categories are classified as sensitive or not sensitive.

Gambhir et al. Used weight as the measurement for each of the actions the user performs in OSN. *Wa* represents weight for action, and *Wp* represents weight for post; Wc and Pc are weight for category and credibility of a post. *Wa, Wp, Wc* and *Pc* were taken into consideration as the parameters while computing the trust factor. Table 29.3 illustrates the idea of using different weight that simulating the algorithm in different test case scenarios as follows:

Figure 29.2 shows the Action-based Trust algorithm of computation of trust factor. Here, Pc determines whether the message which is to be posted is of right commitment or not. In other words, Post Credibility is the measure of the kind of message which is to be posted. It is used in the calculation of Trust factor only when the action performed is Post (If (Type(Wa)==POST)). It is incremented by a factor of 0.001 if the message being posted is categorized in a non-sensitive category (e.g., academics, music, etc.); and gets decremented by a factor of 0.009 if the message is categorized in a sensitive category (alcohol, violence, etc.). The existing algorithm is extended and enhanced further by integrating multi-factor trust attributes which has been ranked in Sect. 3.

Figure 29.3 shows the modified Action-based Trust algorithm.

Before computing the trust factor, seed values of T_f has been compute with the trust rating value given by a user and the weight of trust attributes, also the value is assumed to be up to three decimals place. If seed value is assumed to be integer, the range of the trust factor will get very large which will be very difficult to analyze. Seed value of T_f is calculate as the rating value of a user towards a friend which also another user within same social network site.

$$T_f = \frac{\left(\text{User's Rating Value}, \sum_{i=1}^{i=8} \textit{weight of Trust Traits}\right)}{8} \tag{29.1}$$

After T_f is calculated, this value will become the seed value of a user's friend and it also referred as Old (T_f) in the algorithm.

Table 29.3 Different weightage used in action-based trust algorithm

Weight for action, *Wa*		Weight for post, *Wp*		Weight for category, *Wc*	
Share	0.008	Photo	0.003	Sensitive	0.009
Like	0.006	Video	0.002	Non-Sensitive	0.001
Comment	0.007	Link	0.001		
Dislike	0.006	Message	0.003		
Tagging	0.005				
Post	0.008				

Algorithm CAL_TRUST_FACTOR (username, password, action, post)

Input: username, password, action, post
Output: Trust Factor of user.
Login from openid
While (true)
{
Calculate Weight for Action (W_a).
Calculate Weight for post (W_p).
If (Type(W_a)==POST) then
{
W_c=0
*Call **Matching_Process (Input, CAL_SEL)***
If (Flag == 0)
Calculate P_c = Old(P_c) + .009; //Every right commitment
Else
Calculate P_c = Old(P_c)-.009; //Every wrong commitment
}
Else
Calculate Weight for category (W_c)
Calculate E_s = W_p + W_c +W_a
If (Type(W_a)==POST) then
Calculate Trust factor (T_f) = old (T_f) + P_c + E_s
Else
Calculate Trust factor (T_f) = old (T_f) + E_s
}

Fig. 29.2 Exisiting action-based trust algorithm [11]

Algorithm CAL_TRUST_FACTOR (username, password, action, post)

Input: username, password, action, post
Output: Trust Factor of user.
While (true)
{
Calculate Weight for Action (W_a).
Calculate Weight for post (W_p).
Calculate Weight for catergory (W_c).
// = .001 assume all is non sensitive content
Calculate E_s = W_p + W_c +W_a
If (Type(W_a)==POST) then
Calculate Trust factor
(T_f) = old (T_f) + P_c + E_s //assume Pc = .001
Else
Calculate Trust factor (T_f) = old (T_f) + E_s
}

Fig. 29.3 EABT algorithm

29.5 Enhanced Action-Based Trust (EABT) Model (Miniosn 2.0)

The user will be required to register an account for the proposed Trust Prototype with Enhanced Action-based Trust (EABT) model; which is known as MiniOSN 2.0. Once the user has registered and signed in into miniOSN 2.0, he/she requires to confirm the friend request. After friendship being confirmed, the rating value user holds for a friend would be stored into the system database, the rating value of a friend could be varied from time to time based on the subjective views of user, which we declared as personalize and context dependent. The higher the value represents the more the feeling of trust express from user to friend.

In Log Data page which has shown in Fig. 29.4, the table displayed on the page showing the result of Trust Value, Seed Value (old(Tf)), Post Message and Post Photo for all the friends of a user, respectively. When the rating value of a friend was assigned by user, system will generate these values automatically based on the enhanced Action-based trust computation algorithm running background in the system.

The aim of this user-acceptance survey is evaluate user acceptance of the rating idea used within EABT model. A total of 28 participants who are actively using OSN were interviewed. Most of the candidates have IT knowledge, background and fall within the age range of 20–24 years old. Most participants stated that the rating feature in miniOSN is an ideal way to present trust between human. On the contrary, most of the participants are unwilling to set rating to be visible to others, this might due to the human physiological factor to avoid any hard feeling or unnecessary misunderstandings to people or friend in social media. The enhanced trust model helps participants to relate to privacy and confidentiality of oneself when

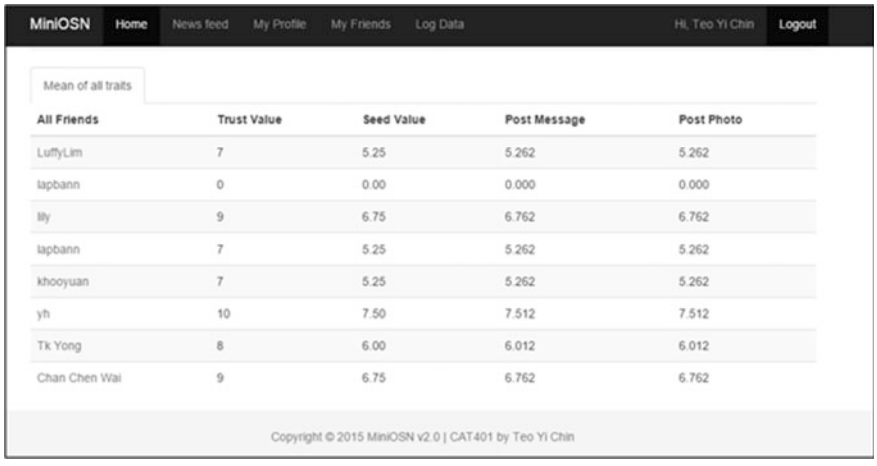


Fig. 29.4 MiniOSN 2.0's log data page

being personal information is being exposed in social media sites. Participants hope that the rating is confidential and it only used to guess the users' behavior characteristic. Through rating, user can choose to preserve confidentiality more effectively. On the other hand, few of the participants refuse to use it, as the rating is time consuming for those who tend to have a lot of "friend" in social networking sites. Similarly, some users are not specifically familiar with the calculation of the Trust Factor (seed values, Tf). Nevertheless, the whole concept is rather basic and simple to understand for the user. It is common to be compared to other online social networks, because trust values are part of the functionality in which online social network should take into account.

29.6 Conclusion and Future Work

OSNs are limited to simple privacy settings where users can control who can view their profile and interact with them or can include them in some community. Trust propagation does not manifest itself as a physical phenomenon in nature, but only exists on the mental and cognitive level. It is therefore difficult to assess whether computational models for trust propagation are adequate and reflect the way people reason about trust. Throughout the study, we have discussed that how the ranked multi-faceted model of trust based on eight trust attributes can be integrated with an action-based trust model. A new model known as EABT is then proposed and the user-acceptance is evaluated. As for future work, a trust calculation algorithm or mathematical model that is resistant to attacks such as Sybil attack, and is applicable to the Quinn's multi-faceted model of trust in this research could be implemented and evaluated. Such an algorithm should highlight because it could enhance the trust management service in OSN by providing accurate recommendations.

References

1. Chieng LB, Mahinderjit Singh M, Zaaba ZF, Hassan R (2015) Multi-facet TrustModel for online social network environment, Int J Netw Secur Appl (IJNSA) 7(1), Jan 2015; Gross R, Acquisti A, Information revelation and privacy in online social networks (The Facebook case). In: WPES '05 proceedings of the 2005 ACM workshop on privacy in the electronic society, pp 71–80. doi:10.1145/1102199.1102214
2. Johnson H, Lavesson N, Zhao H et al (2011) On the concept of trust in online social networks. In: Salgarelli L, Bianchi G, Blefari-Melazzi N (eds) Trustworthy internet. Milano, Springer Milan, pp 143–157
3. Ruohomaa S et al (2005) Trust management survey. In: Lecture notes in computer science, vol 3477. Springer, Berlin, pp 77–92
4. Sabater J, Sierra C (2003) Review on computational trust and reputation models. IIIA-CSIC, Campus UAB, Bellaterra, Barceloana, Spain, Sept 18

5. Hamdi S, Bouzeghoub A, Lopes Gancarski A, Ben Yahia S (2013) Trust inference computation for online social networks. In: 12th IEEE international conference on trust, security and privacy in computing and communications (TrustCom) 2013, pp 210–217
6. Cahill V et al (2003) Using trust for secure collaboration in uncertain environments. Pervasive Comput 2:52–61
7. Zacharia G (2000) Trust management through reputation mechanisms. Int J Appl Artif Intell 14:881–907
8. Marsh S (1994) Formalising trust as a computational concept. PhD thesis, University of Stirling, Department of Computer Science and Mathematics
9. Mahinderjit Singh M, Chieng LB, Hassan R, Zaaba ZF (2015) Friends personalization of trustworthiness for privacy preserverence in social networking, Lecture notes in engineering and computer science: proceedings of the world congress on engineering and computer science 2015, WCECS 2015, 21–23 Oct 2015, San Francisco, USA, pp 428–432
10. Teo CH, Mahinderjit Singh M (2015) Multi-facet trust factors ranking with correlation analysis technique: a case study on Online Social Network (OSN). In: 26th international-business-information-management-association, Madrid, vols I–VI, pp 1812–1822
11. Mohit G, Doja MN, Moinuddin (2014) Action-based trust computation algorithm for online social network. In: Proceedings of 2014 fourth international conference on advanced computing & communication technologies, Rohtak. N.p., 8–9 Feb
12. Mui L, Mohtashemi M, Halberstadt A (2002) A computational model of trust and reputation. In: 35th annual Hawaii international conference on system sciences (HICSS'02), vol 7. IEEE Computer Society
13. eBay (2002) eBay. http://www.eBay.com
14. Amazon (2002) Amazon Auctions. http://auctions.amazon.com
15. Quinn K et al (2009) An analysis of accuracy experiments carried out over of a multi-faceted model of trust. Int J Inf Secur Springer-Verlag 8(2):103–119
16. PallaDutta et al (2014) A novel approach to trust based recommender systems leveraged by context attributes. Int J Eng Technol (IJET) 6(3):1480–1486

Chapter 30
A Motorized Akpu Milling Machine Design for Self Employment and Wealth Generation in Nigeria

Gbasouzor Austin Ikechukwu and Enemuoh Chioma Lorreta

30.1 Introduction

Akpu popularly known as cassava originated from Latin America and was later introduced to Asia in the 17th century and to Africa in about 1558. In Nigeria, cassava is mostly grown on small farms, usually intercropped with Vegetables, plantation crops, yam, sweet potatoes, melon, maize etc. Cassava is propagated by 20–30 cm long cutting of the wood stem, spacing between plants is usually 1.5 m. Intercropping with bean, maize, and other annual crops is practiced in young cassava plantations. There are two common varieties of cassava, namely, the bitter and sweet varieties. The cyanide content differs as well as suitability for different growing and consumption conditions. Usually, higher cyanide is correlated to high yields. Nigeria is one of the world largest producer of Cassava tuber in the world, producing about 34 million tonnes of the world's 174.0 tonnes. Over the past 25 years significant market opportunities for cassava have opened up in the animal feed industry, initially in the EEC (European Economic Community) countries but more recently for the rapidly expanding animal feed industries of tropical developing countries. Cassava roots compete with other carbohydrate sources, especially maize and sorghum, on the basis of price, nutritional value, quality and availability. Cassava has several advantages compared with other carbohydrate sources, especially other root crops. It has a high productivity under marginal climatic conditions, which result in a low cost raw material. Root dry matter content is higher than

G.A. Ikechukwu (✉)
The Department of Mechanical Engineering, Chukwuemeka Odumegwu Ojukwu University, formerly Anambra State University, Uli P. M. B. 02, Nigeria
e-mail: unconditionaldivineventure@yahoo.com; ai.gbasouzor@coou.edu.ng

E.C. Lorreta
The Department of Computer Science, Chukwuemeka Odumegwu Ojukwu University, formerly Anambra State University, Uli P. M. B. 02, Nigeria
e-mail: e.lorreta@yahoo.com

S.-I. Ao et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-981-10-2717-8_30

other root crops at 35–40 %, giving optimum rates of 25:1 or better. Over 85 % consists of highly digestible starch. Cassava starch has excellent agglutinant properties which make it especially suitable for shrimps and fish feeds, replacing expensive artificial agglutinants. The potential disadvantages of cassava roots are their bulk and rapid perish ability, their low protein content and presence of cyanide in all root tissues. Through simple processing the disadvantages of bulk and perish ability can be overcome: A stable product is reached when moisture content falls below 14 %, natural drying is widely used to achieve this objective. Drying also permits the elimination of most of the cyanide from root tissues. The dried cassava product thus has only one disadvantage with respect to other carbohydrate feed sources: low protein content. This can be overcome through price competitiveness. For export markets, where transportation over thousands of kilometers is necessary, further processing to produce high density pellets is carried out to minimize transport costs.

30.1.1 Local Varieties

In the traditional bush-fallow system, some cassava plants are always left to grow with the fallow which is long enough to enable the cassava to flower and set seed. The natural out crossing habit of cassava leads to the production of numerous new hybrid combinations from self-sown seed from which farmers select and propagate desirable types. By this process, pools of new local varieties are continuously created which are adapted to the different agro-ecological zones of the country. As these selections are made on account of their excellent cooking qualities, low Hen (Hydro-cyanide) content and high yields, they are used as parents in breeding programmes mainly to improve pest and disease resistance.

The undermentioned are the local varieties:

i. Akpu ii. Dan wari iii. Oko Iyaw iv. Panya v. Akintola vi. Akon vii. Etunbe

30.1.2 Evolution of Traditional Graters

- The traditional method of grating cassava was by pounding it in a mortar with a pestle.
- Another Innovation is by grating it with a tin can with holes punched in the bottom since the cassava root is spongy and this leads to lots of hand injuries.
- This traditional technology improved by mounting the grating surface on a wooden table at a convenient height so the rubbing action is horizontal rather than in a downward slant when the grating surface is supported against the operators.

In Nigeria many of the cylindrical power milling machines developed to be used in villages are based on the existing design which has some unique design features intended to improve milling efficiency and output without necessarily increasing power requirement. There are however, many variations in design, power transmission, capacity and type of construction.

The cassava milling machine presented a great innovation in cassava processing since grating is central to traditional processing of cassava in Nigeria. Since then, several equipment manufacturers including Engineering firms, research institute, university departments, small-scale artisanal shops, blacksmiths and mechanics have developed and produced various types of cassava processing equipment. Over the past three decades there has been a gradual but steady increase in the adoption of cassava processing equipment in the cassava processing industry. The adoption of mechanized cassava processing appears to have escalated in recent years through assistance provided by non-governmental organization to the rural area.

The total areas under cassava cultivation in Nigeria, is about 0.3.60 million hectares. Their Botanical Name Manihot esculenta. Local Names in Nigeria include; Hausa: Rogo, Igbo: Akpu, Yoruba: Ege

30.1.3 Problem Statement

From the survey, it was discovered that most of the graters incorporated inside the cassava milling machine are usually corroding (reducing service life) due to the acidic nature of the cassava fluid and materials used for the fabrication. To ensure all Cassava products is free from any taste, odour, or infected by iron content of parts (food poisoning) which may affect the quality of their contents, hence need to modify the design and use appropriate materials for fabrication. The product tuber spoils after 2–3 days of harvesting, hence need for processing into safer stable products.

In view of the above mentioned problems and the overall importance of the cassava products the following objectives are required to address the shortcomings of the grater.

30.1.4 Objectives

- To design a prototype cassava milling machine
- To fabricate a prototype cassava milling machine
- Ensuring proper grinding performance thereby eliminating vibrating sieving since the cassava products can be re-grinded to finer particles after pressing and frying etcetera (ensuring no waste).
- To save time and cost of processing cassava products by the average Family.
- To promote healthy consumption of cassava products.
- To test-run and confirm working of the cassava milling.

30.1.5 Scope of Research

The scope of the research is to design and construct a viable motorized Akpu (cassava) milling machine which would be useful for home-use, retailers and small sale farmers.

- To carry out a review on cassava and cassava milling machines in existence.
- To obtain some data or information that will be required and that are suitable in the design and construction of cassava milling machine.
- To select suitable materials based on result of the analysis for the construction of the machine
- To prepare a neat and detailed working drawing for the construction process.
- To discuss the results of the performance test.
- To present the necessary information on the machine efficiency.

30.2 Economic Importance of Akpu Milling Machine

Cassava was introduced to Africa by the Portuguese more than 300 years ago close to the mouth of the Congo River by Portuguese explorers and traders from Brazil, and today is the primary carbohydrate source in sub-Sahara Africa. The plant grows as a bush or little tree. From there it was diffused by Africans, to many parts of sub-Saharan Africa over a period of two to three hundred years. In the course of its spreading across the continent, cassava has replaced traditional staples such as millet and yam, and has been successfully incorporated into many farming systems. It was initially adopted as a famine reserve crop as it provided a more reliable source of food during drought, locust attacks and the hungry season, the period before seasonal food crops are ready for harvesting. Approximately 75 % of Africa's cassava output is harvested in Nigeria, the Democratic Republic of Congo, Ghana, Tanzania and Mozambique. Cassava is grown by millions of poor African farmers, many of them women, often on marginal land. For these people, the crop is vital for both food security and income generation. Cassava is vegetative propagated through stem cuttings and produces well on poor soils. The tubers may be kept in the soil for extended time periods. This secures rural farmers a carbohydrate source in years with adverse growth conditions where other crops fail and famine would otherwise prevail. These features and high crop yield contribute to the importance of cassava in Africa, South East Asia, and South America [1]. Also cassava can be grown on poor soil with no investment in irrigation, fertilizers or pesticides. Currently, the African crop is threatened by African cassava mosaic virus, which can reduce yields by 75 %. Over the past decade, the virus has caused famines in several local regions of central Africa. The disease is prevalent because the African cassava varieties, having been introduced, have no natural defense against the virus. All cassava contains varying amounts of cyanide, a toxin that protects the plant from insects. Indigenous people have learned to avoid poisoning themselves

by spitting into batches of the ground tubers; the saliva introduces bacteria and fungi, which activate an 8 enzyme that breaks down the cyanide. Villagers accomplish the same thing by depositing freshly dug cassava tubers into a community pond; microorganisms in the water degrade the cyanide. It is speculated that starvation indirectly led to these culinary rituals; starving people may have found that, having once spit out the bitter cassava tubers, then trying them again, the bitter taste was lessened.

The diversity of the plant is remarkable. "There must be thousands of different varieties". In Brazil, every little village has its own varieties. There are bitter ones, sweet ones, even ones used as baby food.

Schaal et al. (1999) Burgeoning interest in the crop in recent times results from the realization of the potential of cassava as a food security and emergence crop which could generate employment for the rural poor and foreign exchange for the country. Since 1990 the Government, through the Ministry of Food and Agriculture, has demonstrated its determination and commitment to promote cassava for the alleviation of poverty particularly in rural households and communities. The main reasons for expansion of cassava are population growth, famines or seasonal hunger and market availability. In Nigeria, cassava is moving from a starvation-prevention crop to a cash crop for local urban consumption.

30.2.1 Handling and Processing

Handling and processing conditions often result in a very poor quality of the products. In addition to the high labour intensity and drudgery, the conditions during processing are generally unsanitary and unwholesome. During processing by women in rural areas, losses of some mineral and vitamin value do occur [2]. This can be avoided with better-designed equipment [3]. Agricultural improvements via technology and marketing can make a big difference in less developed countries.

30.2.2 Processing

Processing of cassava into various shelf-stable and semi-stable products is a widespread activity carried out by traditional cassava processors and small-scale commercial processing units. The traditional methods for processing cassava involve combinations of different unit processes including peeling, grating, dehydration, and dewatering, sifting, fermentation, milling and roasting. During processing, the cassava tuber is transformed from a highly perishable root crop into a convenient, easily marketable, shelf-stable product which meets consumer demand for a staple food. Processing may improve the palatability of the product and also reduce the level of cyanogenic glycosides in the tuber thereby detoxifying the product. Products fermented by some species of lactic acid bacteria such as

agbelima and Garr; may attain anti-microbial properties. Several problems are encountered during traditional processing which have created an urgent need for mechanization and upgrading of processing. The processing procedures are labour intensive and time consuming and mostly carried out manually, dust and foreign matter contaminations, losses due to rodents' birds and other domestic animals. Operations are not adequately mechanized because processors cannot afford equipment and do not have access to capital. Processing is often carried out under unhygienic conditions and some unsanitary practices such as improper effluent disposal during the dewatering of cassava mash have adverse effect on the environment. Some operations such as the roasting of gari on open fires present a risk to the health of the processors. But cassava is the "food of choice" even with an abundance of other options in urban settings. However, there is a lot of work needed in processing the tuber. To be edible, it must be peeled, washed, soaked, wet sieved, grated and bagged: no less than 14 steps. This work is done by hand by women of a village. Improvements in processing would greatly aid production. The operations involved in cassava processing depend on the end product desired. In general, the processing stages in cassava include:

i. Peeling ii. Washing iii. Grating iv. Chipping v. Dying vi. Dewatering/fermentation vii. Pulverization and sieving/sifting and viii. Frying/drying

When cassava products are used as a primary staple food, careful processing to remove these toxic constitutes is required to avoid chronic intoxication [4]. Incomplete processing may result in high cyanide exposure and give rise to severe diseases like tropical ataxic neuropathy and konzo, especially in population with poor nutritional status. Unfortunately, careful processing generally results to loss of proteins, vitamins, and minerals, i.e. in products with low nutritional value. In spite of the availability of efficient processing procedures, cyanide exposure from cassava diets prevails [5].

30.2.3 Industrial Uses Cassava Leaves and Livestock Feeds

Cassava leaves are edible and a more convenient food product than fresh roots. Cassava leaves are storable in dry form and since they have lower water content, they are less expensive to dry than the roots. If leaf harvesting is 15 properly scheduled, it does not have an adverse effect on cassava root yield [6, 7]. Cassava leaves have a nutritive value similar to other dark green leaves and are an extremely valuable source of vitamins A (carotene) and C, iron, calcium and protein [8]. The consumption of cassava leaves helps many Africans compensate for the lack of protein and some vitamins and minerals in the roots. Cassava leaves are prepared by leaching them in hot water, pounding them into pulp with a pestle and mortar before boiling in water along with groundnuts, fish and oil. This process eliminates cyanogens from the leaves, making them safe for human consumption. Because of the high-energy contents and low prices of cassava, livestock industries have since been using cassava chips in compounding animal feed production both locally and

internationally. The compound livestock feeds are developed for pigs, cattles and sheeps, goats and poultry.

- UTA Ibadan had succeeded in extracting oil from cassava seeds. The oil is yet to be developed to edible level, but it had been confirmed that it could be used for making soap and for some pharmaceutical products. VITANOI is also obtainable from it. That is, it is medicinal.
- Used as monosodium glutamate, an important flavouring agent in cooking.

Cassava flour is increasingly being used in partial substitution for wheat flour.

- Cassava can be also used for alcohol, syrups etc. Alcohol is in demand in both the food and beverage industry and in the pharmaceutical industry. However, only 5 % was processed into syrup for soft drinks and less than 1 % was used for refined flour or adhesives, so much of the value added production potential is neglected.

30.2.4 Role of Akpu (Cassava) Products

Cassava performs five main roles:

- famine reserve Crop
- rural food staple,
- cash crop for urban consumption,
- industrial raw material, and
- foreign exchange earner,

Also Nigeria is the most advanced of the African countries poised to diversify the use of cassava as a primary industrial raw material in addition to its the role as livestock feed.

Two factors were identified for Nigeria situation: the rapid adoption of improved cassava varieties and the development of small-scale processing technologies. Despite this development, the demand for cassava is mainly for food; and opportunities for commercial development remain largely undeveloped.

30.3 Design of Akpu (Cassava) Milling Machine

30.3.1 Parts Design and Material Selection

In the design of Akpu milling machine many things were considered when analyzing the system.

Manufacturing processes includes the processes involved in using various construction methods in producing the extracting machine. In manufacturing, the

principal common characteristic is that something physical is being produced or created i.e. output consists of goods or machine, which differ physically.

Manufacturing therefore requires some physical transformation or a change in utility of resources. The parts are different components that when assembled make up the unit in such processes care precision should be the top most priority when carrying out the construction. As far as the selection of material for the construction of machine component and parts is a vital aspect of design.

Various manufacturing processes were carried out during the fabrication, production and assembling of the components parts of this machine in order to be producing the required or particular goods.

30.3.2 Theoretical Design and Methodology

The material for the fabrication of the enhanced prototype of cassava milling machine are; the electric motor, mild steel plates and sheets, mild steel angle bars, hexagonal rod mild steel cylindrical tube.

A study was carried out, hardness and resistant ability of cassava tuber was investigated, the study involved the use of laboratory penetrometer and a data lodger. Hand-drilling machine was fitted with a 500 mm diameter blade. These devices were powered manually and by application of electricity. Forces were applied by allowing direct cutting of the tubers using rotating sharp blade. The penetrometer was used to test the strength of fresh tuber. The result obtained indicated that cassava tuber hardness increases with the reduction of moisture content. The penetration force of 60 N at 70 % moisture content wet basis was recorded.

30.4 Design Analysis of the Transmission Drives

In the design of the milling machine, the drives used were the belt drive and pulley transmission system

30.4.1 Determination of Rotor Diameter

Mass of rotor = 10 kg

Force to crush fresh cassava = 60 N

Recall that;

$$\mathrm{F} = \mathrm{Ma} = \frac{Mv}{t} = \frac{m\pi \mathrm{cND}}{60} \tag{30.1}$$

where $D = D_R$; the dimension of the rotor

$$D_R = \frac{F \times 60}{m\pi N_R} = \frac{105 \times 60}{10 \times 3.142 \times 1977}$$

$D_R = 0.101 \approx 0.1$ m = 100 mm

The diameter of the wooden rotor is 100 mm

30.4.2 Diameter of Motor Pulley

Rotation speed N of rotor = N_R = 1977

Diameter of rotor = 100 mm

Available rotational speed = 1450 rpm of motor Nm

$$\mathrm{DmNm} = D_R N_R \tag{30.2}$$

$$\frac{D_R N_R}{\mathrm{Nm}} = \frac{0.1 \times 1977}{1450} \quad \mathrm{Dm} = 0.135 \text{ m} \approx 135 \text{ mm}$$

30.4.3 Value of Electric Motor

$$P = \frac{\mathrm{FWD}}{2} \quad \frac{\mathrm{F}\pi \mathrm{NmDm}}{60} \tag{30.3}$$

where F = force to crush fresh cassava

Nm = rotational speed of motor

Dm = diameter of motor

Adding $^3/_4$ of the force to crush the cassava and also move the wooden shaft so that it overcomes the strength of the cassava to be milled; i.e. F = 105 N

$$P = \frac{105 \times 3.142 \times 1455 \times 0.273}{60}$$
$$= 21841\mathrm{w} \approx 2.93 \text{ hp}$$
$$\approx 3 \text{ hp}$$

Hence, an electric motor of 3 horse power, 1450 rpm, frequency of 50 Hz is needed for the fabrication of this machine.

Table 30.1 Dimensions of standard v-belt khurmi and gupta [9]

Type of belt	Power range in kw	Minimum pitch diameter of pulley D(mm)	Top width b(mm)	Thickness t(mm)
A	0.7–3.7	75	13	8
B	2–15	125	17	11
C	7.5–75	200	22	14
D	20–150	355	32	19
E	30–350	500	38	23

30.4.4 Value of Belt Based on the Power

A transmitted (2.2 KW) and according to the Indian standards (Is: 2494–1974), belt type B was selected from the Table 30.1.

30.4.5 Calculation of Belt Length, L

Khurmi and Gupta [9] developed equation for belt length as shown in equation below;

$$L = \frac{\pi}{2}(D_1 + D_2) + 2X + \frac{(D_1 - D_2)^2}{4X} \tag{30.4}$$

where L = length of belt (mm)
D_1 = wooden shaft diameter, (mm)
D_2 = electric motor diameter (mm)
x = center distance between pully (Fig. 30.1)
Apply Eq. 30.4 when,
$D_1 = 0.1$, $D_2 = 135$ and $X = 0.6654$

$$L = \frac{\pi}{2}(0.1 + 135) + 2(0.6654) + \frac{(0.273 + 0.2)^2}{4 \times 0.6654}$$
$$L = 0.36919 + 1.3308 + 0.08829$$
$$L = 1.78828\,\text{m} \approx 1800\,\text{mm}$$

A belt of A50 (12.5 × 1800) mm was selected.

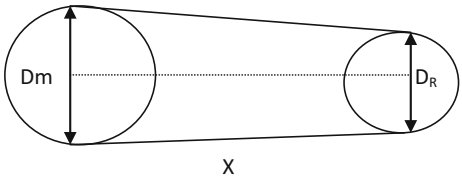


Fig. 30.1 Open belt drive

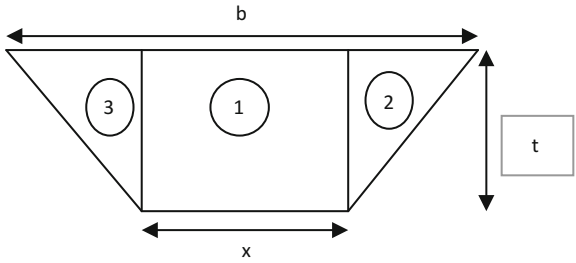


Fig. 30.2 Sectional area of belt cross

30.4.6 Design for Shaft

$$T = \frac{\pi}{16} \times \delta \times d^3 \tag{30.5}$$

$T = (T_1 - T_2)R$

30.4.7 Determination of T_1 and T_2

$$T_1 - Tm - Tc \tag{30.6}$$

$Tm = \delta_a$
Where T = torque on the iron shaft
δ = allowable shear stress
d = diameter of the shaft
T_1 = tension on tight of the belt
T_2 = tension on slack of the belt
Tm = centrifugal tension
Tm = maximum tension on belt (Fig. 30.2).

30.4.8 Determination of Belt Cross Sectional Area

From Table 30.1
Top width b = 17
t = 11
x = 13.4
$\tan 79 = \frac{11}{t_1}$
$t^1 = \frac{11}{\tan 79} = \frac{11}{5.1445} = 1.8$

Table 30.2 Density of belt material

Materials of belt	Mass density in Kg/m^3
Leather	1000
Canvass	1220
Rubber	1110
Balata	1170
Single woven	1250

$$\text{Area of belt} = \text{area of triangle 1} + \text{area of square 2} + \text{area of triangle 3}$$
$$= (1/2 \times 1.8 \times 11)\,2 + 13.4 \times 11 = 19.8 + 147.4 = 167.2 \text{ mm}^2 \quad (30.7)$$
$$= 19.8 + 147.4 = 167.2 \text{ mm}^2$$

The area of the belt = 167.2 mm^2
Maximum available, δ = 2.8 mpa = 2.8 N/mm^2
Apply Eq. 30.8
Tm = 2.8 × 167.2 = 468.16 N
Also, centrifugal tension Tc,

$$Tc = mv^2 \quad (30.8)$$

m = mass of belt per unit length

$$m = \varepsilon a \quad (30.9)$$

From Table 30.2, density of belt = 1140

$$m = \frac{1140 \times 167.2}{100000}$$

= 0.1906 kg
Also, linear speed of belt

$$V = \frac{\pi DN}{60} \quad (30.10)$$

D = Dm = 273
N = Nm = 1450 rpm

$$V = \frac{3.142 \times 0.273 \times 1450}{60} = 20.73 \text{ m/s}$$

Tc = mv^2 0.1906 × (20.73)^2 = 81.9 N
Apply Eq. 30.7
T_1 – Tm – Tc = 468.16 – 81.9 = 386.26 N
For V-belt drive, the tension is given by;

$$\frac{T_1 - T_c}{T_2 - T_c} = \varepsilon^{\varnothing \mu \text{cosec} \frac{x}{2}} \quad (30.11)$$

Table 30.3 Coefficient of friction of material

Belt material	Cast iron steel			Mod	Leader face
	Dry	Wet	Greasy		
Leather oak tamed	0.25	0.2	0.15	0.30	0.38
Leader chrome tamed	0.035	0.32	0.22	0.40	0.48
Canvass stitched	0.25	0.15	0.12	0.23	0.27
Rubber	0.30	0.18		0.32	0.40
Balata	0.32	0.2		0.35	0.40

μ = coefficient of friction between belt and pulley

ø = angle of wrap measures in (ratian)

X = groove angle = 60 (Gupta 2006)

From the table below, coefficient of friction between rubber and pulley (dry cast iron) = 0.3 (Table 30.3).

By considering the small pulley, the angle of wrap Ø was calculated using

$$\begin{aligned} \varnothing &= \left(180 - 2\,\mathrm{Sin}^{-1}\quad \frac{D_1 - D_2}{2x}\right)\frac{\pi}{180}\ \mathrm{rad} \\ &= \left(180 - 2\,\mathrm{Sin}^{-1}\quad \frac{0.035}{1.3305}\right)\frac{3.142}{180} \\ &= (180 - 2\times 1.50)\ \ \frac{3.142}{180} \\ &= (180 - 3.014)\ \ \frac{3.142}{180}\ =\ 3.089\ \mathrm{rad} \end{aligned} \tag{30.12}$$

From Eq. 30.13

$$\frac{386.26 - 81.9}{T_2 - 81.9} = e^{0.3\ \times\ 3.09\mathrm{cosec}\frac{60}{2}}$$

$$\frac{386.26 - 81.9}{T_2 - 81.9} = e^{0.20}$$

$$\frac{304.36}{T_2 - 81.9} = \frac{1.2177}{1}$$

$$1.21744T_{2-9}9.7 = 304.36$$

$$T_2 = \frac{304.36\ +\ 99.7}{1.21744}$$

$$T_2 = 332.28$$

30.4.9 Value of Powertransmitted by the Belt

$$\begin{aligned} Pb &= (T_1 - T_2)\mathrm{V} \\ &= (386.26 - 332.28)20.73 \\ &\quad (53.98)20.73 = \end{aligned} \tag{30.13}$$

Power per belt (Pb) = 1119.00 w

30.4.10 Number of Belt Used

The number of belts required to transmit 3hp power from electric motor was calculated using Khurmi and Gupta [9]

$$\mathrm{N} = \frac{Motor\ Power}{Power} \tag{30.14}$$

$\mathrm{N} = \frac{2238}{1119} = 2$
Number of belts to be used = 2
From Eq. 30.6

$$\begin{aligned} \tau &= (\mathrm{T}_1 - \mathrm{T}_2)\mathrm{R} \\ &= (386.26 - 332.28)\mathrm{R} = 10.8 \end{aligned}$$

Also, from Eq. 30.5

$$d = \frac{\sqrt[3]{\frac{T \times 16}{\pi\tau}}}{\sqrt[3]{\frac{10.8 \times 16}{3.142 \times 42 \times 10^6}}}$$

$= 0.0109 \approx 11$ mm ≈ 15 or 20 mm
The diameter of the iron shaft inside the wooden shaft = 25 mm

30.5 Result and Discussion

Engineers use CAD to create two- and three-dimensional drawings, such as those for automobile and airplane parts, floor plans, and maps and machine assembly. While it may be faster for an engineer to create an initial drawing by hand, it is much more efficient to change and adjust drawings by computer. In the design stage, drafting and computer graphics techniques are combined to produce models of different graters. I manipulate these models on video display screens until they

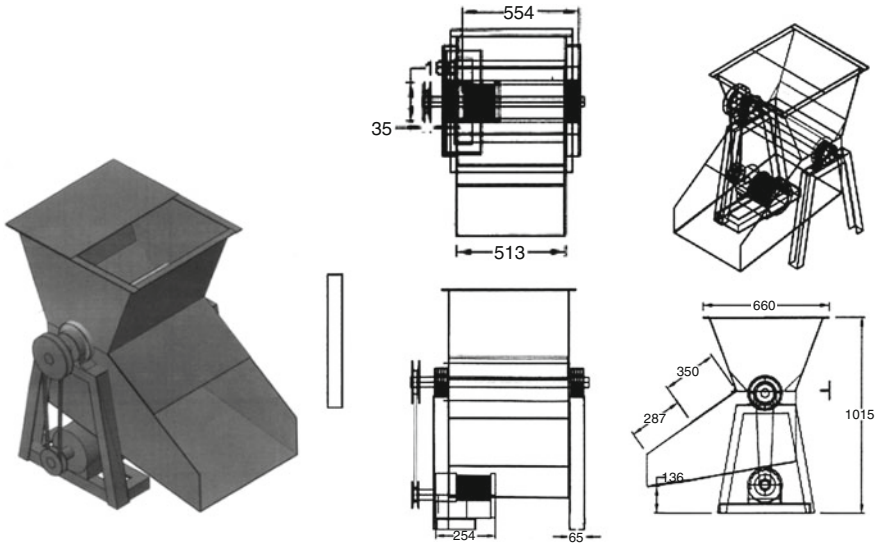


Fig. 30.3 The designed Akpu (Cassava) milling machine in Solid Drawing/Technical view

incorporate the best balance of features, including ease of production and cost. Using a computer to perform the six-step 'art-to-part'

- The first two steps in this process are the use of sketching software to capture the initial design ideas and to produce accurate technical views.
- The third step is rendering an accurate image of what the part will look like.
- Use of analytical software to ensure that the part is strong enough.
- Step five is the production of a prototype, or model.
- In the final step the CAM software controls the machine that produces the part.

During the design of the machine, the drafting software I used was AUTOCAD 2010 version. This was used to draw the orthogonal views, Isometric views and exploded drawings of the model cassava grater before the commencement of the development/construction of the machine (Fig. 30.3).

30.5.1 Performance Evaluation

There are several methods of testing for the efficiency of machines but with respect to this small-scale cassava grater, the following method for efficiency of the cassava grater was examined as follows:

i. Employing existing wooden grating drum grater that is prevalent for grating taking the weight of the cassava being grated and the time taken to grate this size. 50 g was considered for this method

ii. Tests were also carried out on constructed machine using 50 kg of cassava for five different batches. The time taken for each batch was accurately checked and recorded.

Each tuber of cassava was weighed and the weight of the whole batch of cassava obtained. The following measuring parameters were obtained and compared with the results of tests on the constructed machine.

30.5.2 Rate of Cassava Grating

This is the quotient of the weight (tones) of cassava extracted per hour.

For existing machine,

$$\eta r = \frac{\textit{weight of cassva grated}\ (\mathrm{kg})}{\textit{time taken}\ (\mathrm{Hrs})} \\ \frac{(19.56 + 16.65 + 18.77)3600\ (\mathrm{kg})}{(168 + 151 + 178)\ (\mathrm{Hrs})} \\ = 390.4\ \mathrm{kg/hr} \tag{30.15}$$

Therefore for test machine,

$$\eta r = \frac{\text{weight of grated cassva (kg)}}{\text{time taken (Hrs)}} \\ \frac{[51.6]3600\ (\mathrm{kg})}{[1169]\ (\mathrm{Hrs})} \\ = 158.9\ \mathrm{kg/hr}$$

30.5.3 Constraints to Further Development of Akpu (Cassava) Production

The major problems faced for the production and development of cassava are

- Lack of planting material (or effective distribution system), especially of improved varieties;
- Another problem is that cassava produces a large amount of cyanogenic glycosides, which our digestive enzymes break down to toxic cyanide. The plants produce this to defend against herbivores. This is why the cassava tubers must be so heavily processed, otherwise, ingestion leads to konzo (paralysis of the legs). Genetically engineered cassava has been created with greatly lowered cyanogenic glycosides.

30.6 Conclusion

The home use/small scale cassava milling machine was designed, fabricated and tested. It was found to be effective and efficient and could grate about 158.9 kg/h. This machine can be used at home-scale for domestic applications and it is affordable since the cost of production is low which will reduce during large scale production.

Based on the construction materials selection and quality of the research, the machine is durable and expected to last for long time.

Acknowledgements I wish to acknowledge the immerse support given to me by the Tertiary Trust Fund (TetFund), The Vice Chancellor in person of Prof. Okafor Fidelis Uzochukwu for their financial support, encouragement of staff to be attending conferences and presenting papers thereby boosting the image of the University through its output in research, thus, making our academic activities Robust and Excellent diverse field.

Recommendation

Also, I recommend that the machine should be produced on large scale for small- scaled use (commercialization)

The efficiency, design mechanism (in terms of grating unit), and speed at which the machine operates can be improved upon in the future.

The above can be done hand in hand with weight reduction while maintaining balance and reducing machine vibration.

References

1. Nweke FI, Spencer DSC, Lynam JK (2002) The Cassava transformation. Michigan State University Press, East Lansing, MI, Africa's Best Kept Secret
2. FDA (1994) Biotechnology of Food: Background information from the F1) A Nutition Today, vol 29, No. 4. Originally published as an E1) A Backgrounder Statement
3. Kolawole OP, Agbetoye LAS, Ogunlowo AS (2007) Cassava mash dewatering parameters. Int J Food Eng 3(1), Article 4.http://www.bepress.com/ijfe/vo13/issl/art4
4. Onabolu AO, Oluwole OSA, Rosling H, Bokanga M (2002) Processing factors affecting the level of residual cyanohydrins in gari. J Sci Food Agric 82:966–969
5. Oluwole OSA, Onabolu AO, Link H, Roshing H (2000) Persistence of tropical ataxic neuropathy in a Nigerian Community. J Neurol Neurosurg Psychiatry 69(1):96–101
6. Dahniya MT (1983) Evaluation of Cassava leaf, and root production in Sierra Leone. In: Proceedings of the International Society for Tropical Root Crops. Lima.
7. Lutaladio NB, Dahniya MT (1986) Cassava and the African food crisis. Tropical Root Crops International Institute of Tropical Agriculture (lITA), Ibadan, Nigeria, pp 24–26
8. Lathan MC (1979) Human nutrition in Tropical Africa. FAO, Rome
9. Khurmi RS, Gupta JK (2006) Theory of Machine revised edition. Eurasia publishing House (PVT) Ltd, New Delhi.
10. Alyanak L (1997) A modest grater means safer and quicker cassava in Uganda. News Highlights. News/1997/970508-e.htm, FAO. Rome
11. Asuming-Brempong S (1992) Economics of cassava production under structural Adjustment Programme in Ghana. In: Proceeding, fourth Triennial symposium of the International Society for tropical Root Crops—African Branch held in Kinshasa. Zaire, pp 261–268

12. Dadson JA, Kwadzo GTM, Baah KO (1994) Structural adjustment and the marketing of roots and Tubers in Ghana. Department of Agricultural economics, University of Ghana
13. Food and Agricultural Organization, FAO (2000) Cassava an essential part of diet. Championing the Cause of Cassava http://www.fao.org.INEWS/2000/000405-e.htm
14. Hahn SK, Mahungu NM, Otoo JA, Msabaha MAM, Lutaladio NB, Dahniya MT (2007) Engineering research to improve cassava processing technology. Int J Food Eng 3(6) Article 9
15. Grace MR (1977) Cassava processing FAO Plant Production and Protection series No 3. Food and Agriculture organization of the United Nations. Rome
16. Ikechukwu GA, Lorreta EC (2015) Development of a Motorized Akpu Milling Machine Design to Improve Poverty Eradication in Nigeria. In: Proceedings of The World Congress on Engineering and Computer Science 2015, 21–23 Oct 2015. Lecture Notes in Engineering and Computer Science, San Francisco, USA, pp 945–952
17. Onabolu A, Abbass A, Bokanga M (2003) New food products from cassava. IITA Publications http://www.cassavabiz.org/
18. Okigbo Bede N. (1984). Improved permanent production systems as alternative to shifting intermittent cultivation, page 1–100 in improved production systems as an alternative to shifting cultivation. FAO Soils Bulletin 53. Soils Resources Management and conservation services, land and water Development Division, Food and Agriculture Organization of the United Nations. Rome, Italy.
19. School BA, Olsen KM (1999) Evidence on the origin cassava: Phylogeography of manihot esculenta. Proc. Nat. Acad. Sci, USA 96:5586–5591

Chapter 31
Response Behavior Model for Process Deviations in Cyber-Physical Production Systems

Nadia Galaske, Daniel Strang and Reiner Anderl

31.1 Introduction

The opportunity of globalization for discrete manufacturing industries in low-wage countries presents a challenge for manufacturing companies in high-wage countries. For the preservation of jobs and maintaining their competitiveness, these countries research and develop new production facilities with highly efficient and automated production systems. New approaches aiming in a different direction of research are being developed. The German Federal Government is supporting the research of production systems using cyber-physical systems as part of the initiative *Industrie 4.0* [1]. These cyber-physical production systems are characterized by highly flexible production processes that enable the production of a high variety of products in small batches using the same production system without high costs. Products, components, resources and machines possess individual information about themselves and are part of a provided network of things and data (Internet of Things and Internet of Data). Using this network and communication interfaces, all participants can communicate with each other and develop the production process autonomously through an information exchange.

During these autonomous processes, deviations from normal state can occur. These can be induced by missing components, blocked manufacturing stations, missing workforces, or incorrect information of products, manufacturing stations,

N. Galaske (✉) · D. Strang · R. Anderl
Department of Computer Integrated Design, Technische Universität Darmstadt,
Otto-Berndt-Straße 2, 64287 Darmstadt, Germany
e-mail: galaske@dik.tu-darmstadt.de

D. Strang
e-mail: strang@dik.tu-darmstadt.de

R. Anderl
e-mail: anderl@dik.tu-darmstadt.de

S.-I. Ao et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-981-10-2717-8_31

resources, and workers. These deviations have to be detected, analyzed, and processed to define a response behavior and decide for the best possible response strategy.

Due to the novelty of cyber-physical production systems, possible deviations must be identified before applying the production system for mass production. Based on an analysis of the system and possible deviation scenarios, a deviation management system can be developed to support decision making process in cyber-physical production systems. For that purpose, an Ishikawa diagram with possible causes for process deviations is derived and a model for the response behavior of the cyber-physical production system is defined for each type of deviation. Based on these models, material flow simulations are developed to illustrate the behavior of cyber-physical production systems and the response to deviations [2]. These simulations can be used for the decision of the best possible response strategy and to detect weak points and challenges of the theoretical models of the responses to deviations.

31.2 State of the Art

31.2.1 Cyber-Physical Production Systems

Increasing competition due to the globalization and demands for more individualized product configurations lead to challenges for the economies of countries and discrete manufacturing companies. Cyber-physical systems offer an innovative solution to address the challenges in the production. These cyber-physical production systems are the foundation of the *Industrie 4.0*[1] initiative of German Federal Government [1].

Cyber-physical systems (CPS) are based on the two principles: *cyberizing the physical* and *physicalizing the cyber* [3]. This means that every physical object possesses a digital representation in the virtual world and every virtual object exists in the physical world. These systems can be defined by following characteristics [4–6]:

- CPS consist of sensors, actors, embedded systems, mechanical structures, and human-machine interfaces.
- CPS collect information using sensors, analyze these using worldwide services, and use actors to interact with the physical world.
- CPS are equipped with communication devices to connect with each other and with other entities in a global network.

The integration of cyber-physical systems in the production environment leads to a cyber-physical production systems (CPPS) as part of *Industrie 4.0* [1]. These cyber-physical production systems are characterized by highly flexible production

[1]More information about Industrie 4.0: www.plattform-i40.de.

processes, which are able to adapt according to the current market circumstances, and enable the production of a high variety of products in small batches. This approach brings many opportunities, as it enables the individualization of customer requirements, increases the resource productivity and efficiency, and optimizes the decision making process due to complete availability of information in real time [7]. Prominent industrial and research partners of the research and development initiatives like *Industrie 4.0* or the *Industrial Internet Consortium (IIC)*[2] show the significance of CPPS.

For the manufacturing context, cyber-physical systems can be applied in production resources, such as manufacturing stations, automation devices, single machines, and tools, as well as on individual components or products, resulting in smart products. These smart products are mechatronic products equipped with cyber-physical systems that possess information about their manufacturing operations and are able to communicate with each other using modern internet technologies [8]. The collaboration of smart products and smart production systems using internet technologies and context-awareness leads to a smart factory. A smart factory is a manufacturing solution with adaptive production processes, complex decision problems and a management for increasing complexity in rapidly changing conditions [9]. Products can be manufactured with variable manufacturing processes and varying manufacturing stations. The decisions for manufacturing processes and stations are based on an information exchange between components and manufacturing resources. Therefore the processes are defined as *communication-based cyber-physical production processes* [10].

However, some challenges associated with the implementation of cyber-physical production systems and processes, such as the need of unified standards and reference architectures, the safety and security of the production and communication systems, and the increasing complexity of production processes, e.g. due to Big Data [7], still need to be addressed.

31.2.2 Modeling and Simulation of Cyber-Physical Production Systems

For a better understanding of the behavior of a cyber-physical production system and its elements, it has to be represented using models. A model is defined as a simplified reproduction of a system, including its characteristics and processes [11]. Through the simplification in terms of abstraction in modeling techniques, the complexity of the observed problems in the modeled system can be reduced, thus making it easier to find a solution.

An approach to illustrate the relations of the participants in a cyber-physical assembly process and their influence on each other using UML class diagrams can

[2]More information about the IIC: www.iiconsortium.org.

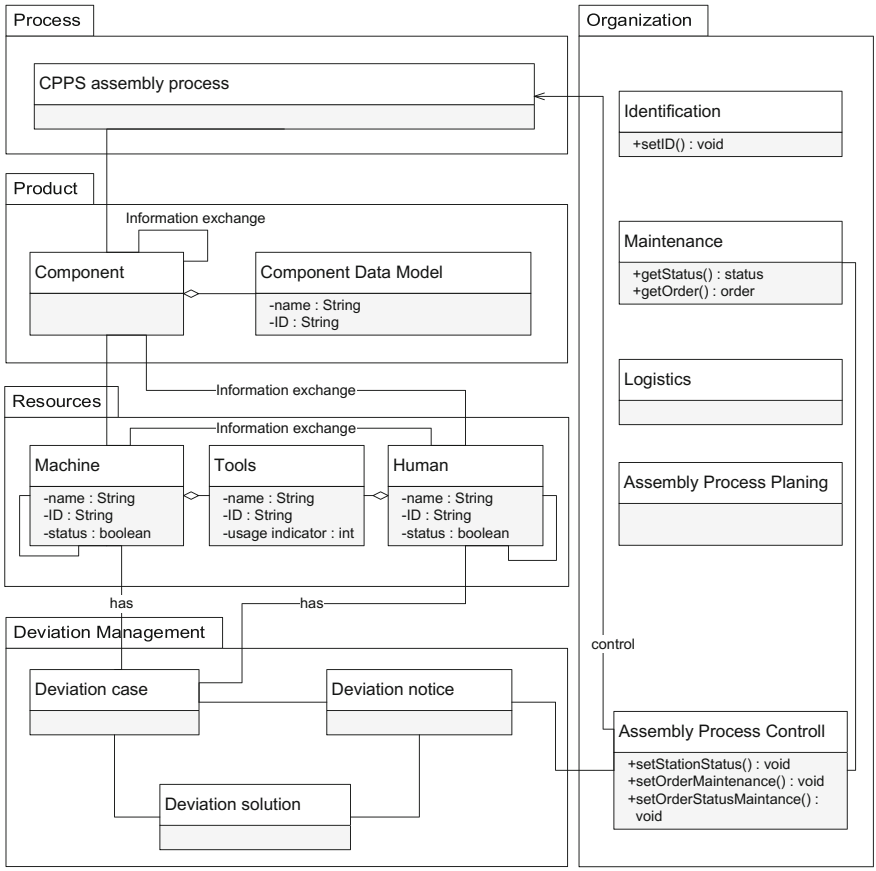


Fig. 31.1 Meta model of the cyber-physical assembly process

be found in [10, 12]. This is possible with the provided UML standards of the Object Management Group [13], which are the foundation of the modelling techniques described in the literature, e.g. Weilkiens [14] or Miles and Hamilton [15]. Figure 31.1 shows the packages and the classes of the class diagram in a simplified way.

The class diagram is structured into five packages: process, products, resources, deviation management, and organization. This architecture is chosen to combine participants and other elements into groups that have similar tasks in the system.

The key package for this chapter is the deviation management. It identifies and classifies the occurred deviation case, identifies a suitable solution and response, and creates a deviation notice for the process control. With a confirmation from the process control, a response of the system can be executed and considered in the production process.

Classes in the package of the product are the component and a component data model. The component data model is an individual, digital representation of each component. Therefore, it is possible to represent all product data and the actual data of the manufactured components [5]. With this information, planned and actual data of each component can be compared [16], for example to diagnose the fulfillment rate of quality requirements or to influence the product development process. Components can therefore influence their own production process and act as a key enabler for smart production planning and process control [17]. The component data model offers a possibility to store data of each process event. Thus, information about process deviations can be stored as part of the data model.

A simulation is used to represent a system with its elements and its dynamic processes. It provides the possibility to examine the behavior of a system which does not yet exist and transfer the result into the system development [11]. Simulation methods can be classified according to the time interval, in which state change occurs. In the field of manufacturing, discrete-event simulations (DES) offer suitable methods for simulating production processes and sequences, as well as material flow [18].

31.3 Process Deviations in Cyber-Physical Production Systems

31.3.1 Process Deviations, Failures, and Disturbances

A standard or default situation describes a sequence of events that is the result of previously planned actions. This sequence of events is subjected to certain principles and is connected in a causal relationship. In a standard operating state, the production process runs as planned and no intervention is needed.

During the production however, process deviations can occur, which lead to a non-executable process sequence. Such events are called deviation state. In this case, problems arise in the execution of manufacturing operations causing the initial plan of the production process to become unfeasible. If the deviation between the current and planned state is significantly large, the process can be disrupted, causing a delay in the order processing [19]. In order to prevent the problem escalation, the production system has to react according to deviation strategies.

Process deviations can be caused by failures or disturbances in production processes. Failures are defined as the non-functional state of a system element, while disturbances are perceived as temporary events that appear unexpectedly and cause an interruption or delay in the task execution. When disturbances take place, the process drifts from its optimum course [20, 21]. Failures and disturbances lead to inconsistency between the planned and the actual production process, in which process deviations are observed. Failures and disturbances on production system elements can be caused for example by missing workforce, quality problems, speed losses, as well as planned and unplanned downtime [22, 23].

31.3.2 Identifying and Categorizing Process Deviations

Cyber-physical production systems have the ability to identify and analyze process deviations as well as to decide for the best response strategy. Due to the intelligence carried in the component data model, a manufacturing component in cyber-physical production systems possesses information about its manufacturing and assembly operations. In case of a process deviation, the manufacturing component must be able to react accordingly and to make the decision between possible alternative processes based on the stored and available information.

Before modelling process deviations in cyber-physical production systems, it has to be defined, what kind of deviation can occur in a production system. Meyer et al. [24] provides an overview of approaches for determining and categorizing failures and disturbances causing process deviation in production processes.

In this section, an Ishikawa diagram is used to identify potential factors causing a process deviation. The Ishikawa diagram illustrated in Fig. 31.2 contains possible causes of process deviations in cyber-physical production systems. The source of process deviation is divided into the five categories (referring to [25]):

- Material (Component),
- Machine (Manufacturing stations),
- Method (CPPS),
- Man Power (Worker), and
- Milieu (Environment).

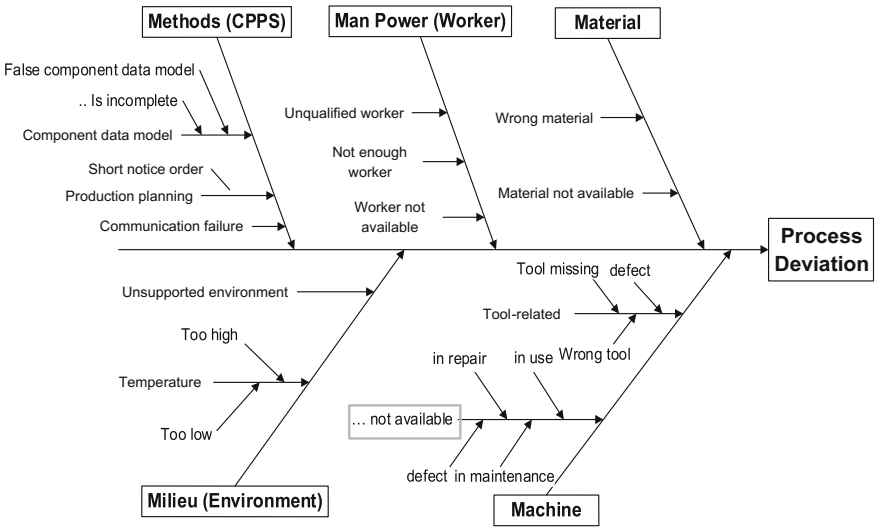


Fig. 31.2 Ishikawa diagram for determining causes of process deviations in cyber-physical production systems

In this paper, a theoretical concept of cyber-physical production systems is analyzed. The focus is to examine the behavior of production system's elements. Therefore, the Ishikawa diagram is used to find the cause of a problem in a system which does not yet exists, so it can be analyzed and modeled. To analyze the plausibility and interdependencies of different potential causes, deviation sources for each category can be analyzed and evaluated using a consistency matrix. In this way, similar potential causes for process deviation can be put together in a use case scenario for a detailed analysis [19]. In the next chapter, the process deviation scenario caused by the unavailability of the assigned/default manufacturing station is selected for the modeling and simulation analysis due to its representativeness.

31.4 Modeling and Simulation

31.4.1 Modeling Response Behavior

The increasing complexity and dynamics of cyber-physical production systems cause a high susceptibility to deviations resulting from failures and disturbances in the production processes. In order to be able to manage process deviations in a cyber-physical production system and determine the suitable response strategy, the modeling and simulation of process deviations is required.

The goal is to develop a logical model [26] for the description of the system and the system's behavior during a process deviation scenario. The model aims to show how a cyber-physical production system behaves in certain situations under certain circumstances. The modeling of the cyber-physical production system consists of the development of a dynamic simulation model, the execution of the simulation, and the analysis of the simulation results. The behavior of a cyber-physical production system will be demonstrated using only a section of the production line with a number of manufacturing station, as shown in Fig. 31.3.

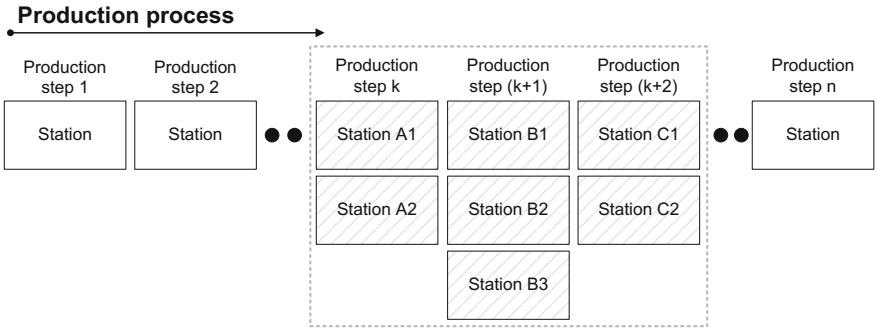


Fig. 31.3 System boundary of the cyber physical production systems with exemplary number of parallel stations

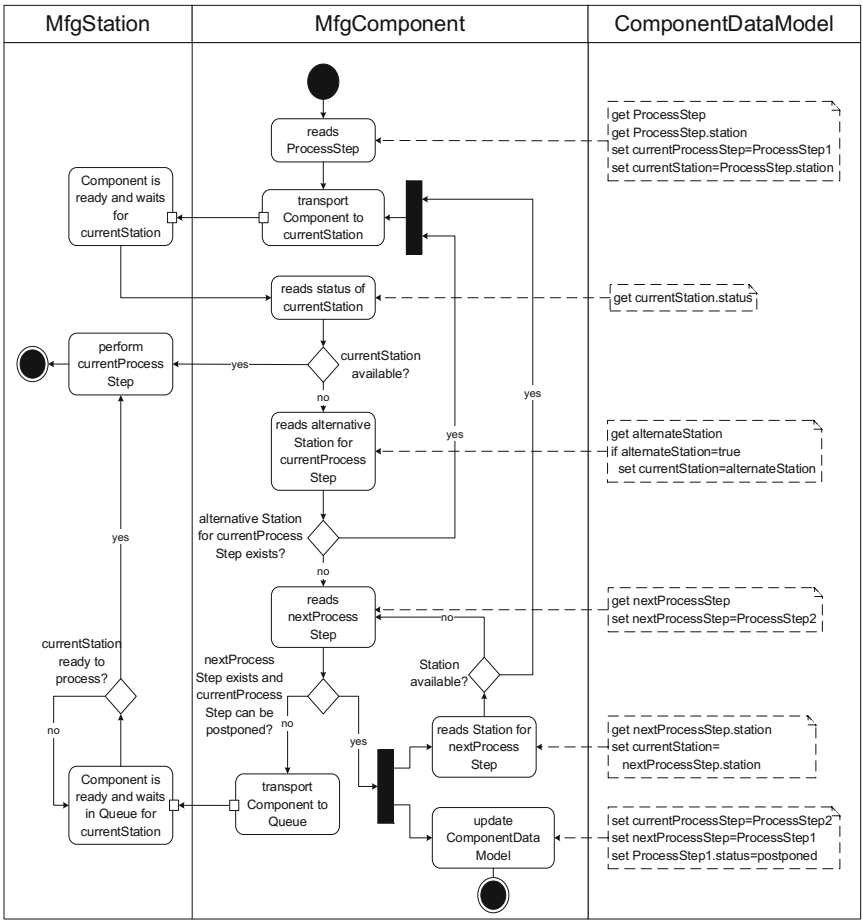


Fig. 31.4 Decision process model

The modeling of the chosen scenario is done using UML activity diagrams. The standard scenario is described as follows: a manufacturing component in a cyber-physical production line arrives at a manufacturing station according to the information stored in its component data model. The station is available and the manufacturing process can begin.

In the deviation scenario, the required manufacturing station is not available. Thus, the production process is interrupted and the system has to determine the suitable response strategy for the component as quickly as possible. Figure 31.4 visualizes the decision process model for determining the suitable response for the process deviation caused by unavailability of the required manufacturing station. Each action is associated with the corresponding system element represented as swim lanes.

In the first step, the component reads the information stored on its component data model. As specified by the component data model, the current process step is determined and the component is transported to the required station according to this process step. Using the cyber-physical system approach, the component is able to communicate with other components, as well as with the manufacturing station. The component sends a query to the station, whether the manufacturing operation can be started. When the component detects that the manufacturing station, on which the current manufacturing operation should take place, is not available, the component tries to identify alternative or parallel stations for the current process step using its component data model. If an alternative station exists, the component is transported to this station and requests for the manufacturing process to be performed.

If no alternative station exists, the component reads the next process step from its component data model and sends a query, whether the process sequence can be altered. If this is the case, the current process step is postponed and the next process step is set as the new current process step. The component data model is overwritten with the updated process sequence. The manufacturing station for the new process step is identified and the manufacturing process can be carried on.

If no parallel station is available or no subsequent process step exists, the component is added to the queue of the current manufacturing station and waits for the current station to be ready to perform the current manufacturing process step.

31.4.2 Implementation

In this chapter, the modeled process deviation scenario is implemented using a discrete-event simulation software. Figure 31.5 shows the design of the simulation, with the manufacturing process steps marked in boxes and the manufacturing stations underlined. In this simulation, the production line consists of three subsequent process steps, in which the first process step consists of two manufacturing

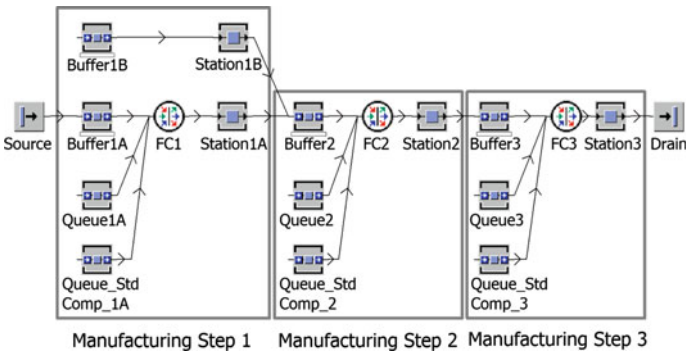


Fig. 31.5 Simulation design

stations that can work parallel. In conformity with the chosen scenario in the previous chapter, the process deviation occurs because *Station1A* as default station of the first manufacturing step is not available.

For the simulation, the behavior of two different types of manufacturing components are observed and analyzed:

1. Standard components
2. Smart components

Standard components represent normal or default objects in traditional manufacturing and assembly processes. They are neither equipped with sensors, actors, nor a communication interface and do not have the intelligence provided by the component data model. The production process of standard components runs according to the planned process sequences. If a deviation happens, the standard components cannot react autonomously and thus must wait until the problem is solved.

In the simulation design shown in Fig. 31.5, standard components can only be manufactured on the default station (*Station1A*). In case of a process deviation, the standard components have to wait in the corresponding queue (*Queue_StdComp_#*) until they can be processed again.

Smart components, on the other hand, are intelligent manufacturing and assembly objects equipped with cyber-physical systems (sensors, actors, and a communication interface) and a component data model. Using the information stored in the component data model, smart components can communicate with each other as well as with the manufacturing stations. The processing sequence for each component can also be varied to a certain extent. Thus, the originally planned process sequence can be adjusted in case of a deviation caused by failures or disturbances in the production process.

In the simulation design shown in Fig. 31.5, smart components can be manufactured either on *Station1A* or *Station1B*. The process sequence between the first and the second manufacturing step can also be varied. If the process deviation takes place, a smart component can choose to do one of the following three actions:

1. Search for an alternative station of the same manufacturing step (*Station1B*),
2. Search for an available station of the next manufacturing step (*Station2*), and then go back to the first manufacturing step (either *Station1A* or *Station1B*) after it has been processed, or
3. Wait in the *Queue1A* until *Station1A* is available

31.4.3 Results and Discussion

Using the decision logics stored as algorithms for each of the component in the simulation model, it is possible to run the simulation and verify the behavior of both

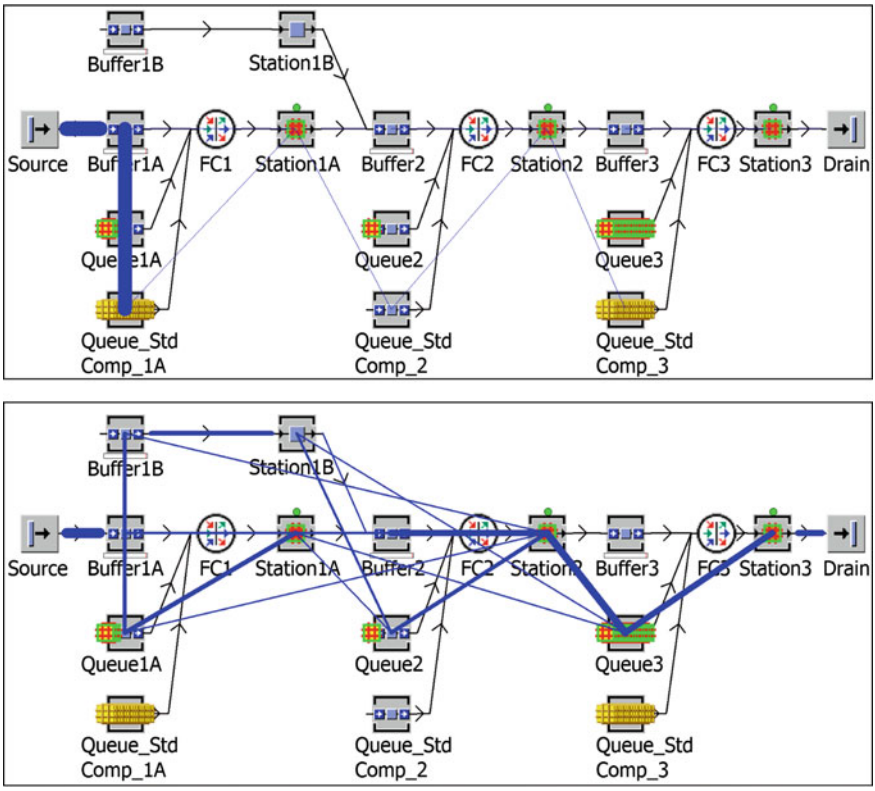


Fig. 31.6 Sankey diagram for the material flow of standard components (*above*) and smart components (*below*)

types of the components in the simulation model using Sankey diagrams, as shown in Fig. 31.6. Due to the intelligence of the smart components and the possibility to alter the production sequences according to the availability of manufacturing stations, the bottleneck in the production process flow can be minimized. Smart components also show higher throughput per hour than standard components.

In this simulation model, only a section of the production line is analyzed. Nevertheless, the methods and the results of this simulation study can be transferred for the modeling and simulation of a bigger production system, as another study of the authors suggests [19].

For a holistic observation of process deviations in cyber-physical production systems, the interdependencies between different potential causes of deviations have to be researched. This allows the analysis of the escalation scenario with several deviation causes occurring at the same time and the research for a proper response strategy for this case.

31.5 Conclusion

In this paper, process deviations in cyber-physical production systems are defined, analyzed, and modeled using UML. Using basic modeling elements of UML, it is possible to visualize the decision process for smart components in a cyber-physical production line. The modeled concept is implemented in a simulation model using discrete-event simulation software. With the simulation, the differences in behavior of standard and smart components during a process deviation in a cyber-physical production system can be illustrated and observed.

The concept for modeling and simulation of process deviations in cyber-physical production systems presented in this paper serves as a recommendation for further research of the behavior of cyber-physical production systems. This is the foundation for developing concepts to ensure the robustness for the real application of cyber-physical production systems.

Future works need to address the responsiveness of the management of process deviation and the response behavior. For a full contemplation of a cyber-physical deviation management, the communication between the simulation software and the decision response system plays an important role in order to enable a near real-time response. Therefore, research in this particular area is needed.

For a holistic approach in managing process deviations and the escalation training, several scenarios where different process deviations occur in random sequences should be modeled and simulated. In this case, the dependencies and consequences of each cause of process deviations can be considered and the robustness of the cyber-physical production system can be tested.

References

1. BMBF: Zukunftsbild Industrie 4.0. http://www.bmbf.de/pubRD/Zukunftsbild_Industrie_40.pdf
2. Galaske N, Strang D, Anderl R (2015) Process deviations in cyber-physical production systems. Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering and Computer Science 2015, WCECS 2015, 21–23 Oct 2015, San Francisco, USA, pp 1035–1040
3. Lee EA (2010) CPS foundations. In: Design automation conference (ACM), 737–742 (2010)
4. acatech: Cyber-physical systems. Driving force for innovation in mobility, health, energy and production. Munich (2011)
5. Anderl R, Strang D, Picard A, Christ A (2014) Integriertes Bauteildatenmodell für Industrie 4.0. Informationsträger für cyber-physische Produktionssysteme. ZWF 109, pp 64–69
6. Bauernhansl T, ten Hompel M, Vogel-Heuser B (eds) (2014) Industrie 4.0 in Produktion, Automatisierung und Logistik. Anwendung, Technologien und Migration. Springer Vieweg, Wiesbaden
7. Kagermann H, Wahlster W, Helbig J (2013) Recommendations for implementing the strategic initiative Industrie 4.0. Securing the future of German manufacturing industry (2013)
8. Anderl R, Picard A, Albrecht K (2013) Smart Engineering for Smart Products. In: Abramovici M, Stark R (eds) Smart Product Engineering. Proceedings of the 23rd CIRP

Design Conference, Bochum, Germany, March 11th–13th, 2013, pp. 1–10. Springer, Berlin, Heidelberg (2013)
9. Radziwon A, Bilberg A, Bogers M, Madsen ES (2014) The Smart Factory. Exploring adaptive and flexible manufacturing solutions. Procedia Engineering 69:1184–1190
10. Strang D (2016) Kommunikationsgesteuerte cyber-physische Montagemodelle. Shaker, Aachen
11. VDI 3633-1: Simulation of systems in materials handling, logistics and production. Fundamentals. VDI, Düsseldorf (2010)
12. Strang D, Anderl R (2014) Assembly Process driven Component Data Model in Cyber-Physical Production Systems. Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering and Computer Science 2014, WCECS 2014, 22–24 Oct 2014, San Francisco, USA, pp 947–952 (2014)
13. Object Management Group (OMG): Unified Modeling Language (OMG UML), Superstructure. Version 2.4.1 (2011)
14. Weilkiens T (2007) Systems engineering with SysML/UML. Modeling, analysis, design. Morgan Kaufmann, Burlington, Mass
15. Miles R, Hamilton K (2006) Learning UML 2.0. O'Reilly, Sebastopol, CA
16. Picard A (2015) Integriertes Werkstückinformationsmodell zur Ausprägung werkstückindividueller Fertigungszustände. Shaker, Aachen
17. Picard A, Anderl R (2014) Integrated component data model for smart production planning. In: Schützer K (ed) Proceedings of the 19th International Seminar on High Technology. Piracicaba, Sao Paulo, Brazil
18. Mattern F, Mehl H (1989) Diskrete Simulation - Prinzipien und Probleme der Effizienzsteigerung durch Parallelisierung. Informatik Spektrum 12:198–210
19. Galaske N, Anderl R (2016) Disruption management for resilient processes in cyber-physical production systems. Procedia CIRP 50:442–447
20. REFA: Methodenlehre der Betriebsorganisation. Hanser, München (1991)
21. Heil M (1995) Entstörung betrieblicher Abläufe. Deutscher Universitätsverlag, Wiesbaden
22. Ingemansson A, Bolmsjö GS (2004) Improved efficiency with production disturbance reduction in manufacturing systems based on discrete-event simulation. J Manuf Technol Manage 15:267–279
23. Knüppel K, Meyer G, Nyhuis P (2014) A universal approach to categorize failures in production. waset.org (eds) Int J Mech Aerosp Ind Mechatron Eng 8:24–28
24. Meyer G, Knüppel K, Busch J, Jakob M, Nyhuis P (2013) Effizientes Störgrößenmanagement. Ansatz zur Kategorisierung von Störgrößen in der Produktion. Prod Manag 18:49–52
25. Kamiske GF, Brauer J-P (2008) Qualitätsmanagement von A bis Z. Erläuterungen moderner Begriffe des Qualitätsmanagements. Hanser, München
26. Hoover SV, Perry RF (1989) Simulation. A problem-solving approach, Addison-Wesley, Reading, Mass

Chapter 32
Complete Bid Configuration. Supporting the Process from the Perspective of the General Contractor

Tomasz Błaszczyk and Paweł Błaszczyk

32.1 Introduction

The problem of the relevant contract type and contractors' selection is a key task of each project owner. However, many scientific papers examine them theoretically or in a relation to the specific conditions of the project, it seems that there is still no comprehensive system modeling contracting process at different levels of WBS in general, supporting project's owner and the owners of individual tasks in a project in making optimal decision from the point of view of the whole process, as well as their particular benefits and risks. Before formulating the concept of the model, we assumed that it must meet the criteria of a mutli model, as described in the paper of Budzinski' and Becker [1]. In particular, it should take into account the multi-stage nature of the decision-making process, multi-criteria nature of the problem, a number of decision-makers and experts, scalability of the decision problem flexibility of decision variants, and linguistics of the date should be taken into account. The Project Management Body of Knowledge (PMBoK Guide) PMI [2] (current and older editions) delivered a complex, systematic approach to project management processes where the project procurement and contracts management issues are considered but not comprehensively described in terms of input and output data with an indication of the effective optimization methods. In our paper [3] we suggested to develop this approach by introducing a bid configuration process. In the literature, there are many well-known approaches and opinions on the ways to select contractors and conclude contracts for individual projects scopes. Ranging from the general guidance to choosing the form of the contract (as described in Turner and Simister

T. Błaszczyk (✉)
University of Economics in Katowice, Katowice, Poland
e-mail: tomasz.blaszczyk@ue.katowice.pl

P. Błaszczyk
University of Silesia, Katowice, Poland
e-mail: pawel.blaszczyk@us.edu.pl

S.-I. Ao et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-981-10-2717-8_32

[4]), by indication of the specific selection methods (such as Holt [5]) until ready proposals of systems supporting sub-contracting in certain types of projects (e.g. in the paper of Arslan et al. [6]). Almost all the authors emphasize that the selection of the project contractor (and, of course, possible sub-contractors) is a problem of multicriterial nature. In the modern approach to procurement management in projects it is noticeable to move away from the total transfer of risk to the contractors for the partnership approach, taking into account the risk premium for both parties and trade-offs allowing more flexible development opportunities for the assessed criteria. Lack of this flexibility means that the rigid need rigid to adapt to the customer requirements caused the final increase in project costs. Of course, such solutions are theoretically and practically effective (e.g. Bubshait [7], Lahdenperä [8], Boukendour and Hughes [9] or Missbauer and Hauber [10]), but not always allowed in all legal environments (see Bochenek [11]). At the core of our proposed methodology lies the concept of critical chain by Goldratt [12] supported with the idea of the cost buffer with incentive fund described by Błaszczyk and Nowak [13]. This approach was formalized by Błaszczyk et al. [14] with introducing the matrix of factors influencing standard projects' evaluation criteria of duration and cost, and by developing work effort estimations for each projects' activity by Błaszczyk and Błaszczyk in [15] for deterministic case, and Błaszczyk et al. in [16] for fuzzy estimations. Models supporting the decision-maker in the role of general contractor was described by Błaszczyk and Błaszczyk in [17] for the most general case and in [18] for specific expression of fixed price and cost-plus contracts and more broadly in [19]. The general outline for the DSS realizing this approach procedures was submitted by Błaszczyk and Błaszczyk [20]. This paper is a revised and extended version of [21], presented on the World Congress on Engineering and Computer Science 2015.

32.2 Objectives

The first step in our research was to gain knowledge of methods, tools and procedures used by project managers in making procurement decisions and also about their needs in this area. To gain this knowledge we conducted the survey using online survey tools (such a SurveyMonkey). The invitation to participate in this research was sent to the members of the Polish branch of PMI, IPMA Association of Poland, as well as project managers working in selected corporations and former students of the project management studies. We received 123 fully filled forms For the purpose of the survey we formulated 26 questions concerning, among others, the attitude of the respondent towards the decision-making process of purchasing in his organization such as the usage of the subcontractors, service providers, methods and criteria for their evaluation and selection (and also the criteria for evaluation and selection the best bid), used types of contracts, formal and legal restrictions about making procurement decisions, awareness of methods and tools to help make purchasing decisions in project management, awareness and the need for using such methods and tools for supporting decisions. The conducted survey confirmed that a large part

of the people involved in the implementation of projects, including project managers, do not know the formal methods supporting procurement decision making in particular the multicriteria methods. The most often used method to select the bids/bidders is to rank offers by using several criteria. As the most commonly used for other purposes than the choice/compare offers methods paired comparison was pointed. Most people declare that they have some knowledge of one of the formal methods but they do not apply it in practice. This confirms our earlier observations indicating that these methods are seen as unsuitable because of their complicated and time-consuming use or because they do not fit to the realities of the business. Therefore these method are seen as useless in real-life practice. Hence, methods and tools to support decision-makers in making procurement decisions should take into account the need for their understanding of the users (employees) and the possibility of uncomplicated applications, preferably supported by a computer application. By adopting such an assumption to one of the world's most popular project management methodologies—PMI's PMBoK Guide, we proposed the additional process that enables planning and controlling of activities in relation to the activities associated with obtaining contracts and optimization offers. Most of the respondents (43.09 %) said that during the implementation of the project they always used the services from suppliers/subcontractors. In 49 cases, each time, such a decision was considered individually, while 18 respondents indicated that the whole scope of the project was implemented by themselves. Most of the respondents (63 %) indicated that in their organization there was a list of qualified suppliers. Among the organizations that have such kind of a list (registry) provided, that they are also looking for other suppliers (not listed), while in 20 % of cases, orders may only be granted to companies which were previously verified and placed on that list. Most often the mentioned criteria used for the offers selection are the price/performance cost, speed of execution, experience, specialization vendor warranty. Among the leading organizations the list of the most common criteria applied to companies evaluation are (in order): warranty, confidentiality of technical/commercial information, specialization, speed of execution, its technology manufacturing. When deciding on selecting the contractor most respondents are guided by the financial health of the contractor (51 %) and the number of executed orders (42 % of respondents). Sometimes informal reviews (49 % of respondents) and the prestige of the con-tractor (42 % of respondents) are taken into account. While selecting a particular offer usually respondents take into account: price, deadline and quality. Sometimes respondents take into account an assessment of the supplier and the total cost of the use. Around 77 % respondents agreed with the statement that the cheapest offer is not necessarily the best. Most people pointed out that the used type of contract depends on the specific project. From the different type of contracts the most respondents pointed out the fixed price contract and on the third position the cost plus type contracts. The majority of the respondents (67 %) use in their organizations penalties against contractors/suppliers. From all the organizations that use this type of clauses in contracts most of them apply penalties at a flat rate. A quarter of respondents said that their organizations id not use this type of clauses. In our research we also try to discover how people understand the term "quality" of the offer. Around 58 % of respon-

dents answered this question (it was an optional question). Answers mostly referred to the assessment of conformity with the specifications of the contract offer. In as many cases the respondents pointed that the use of a system of multi-evaluation does not indicate, however, what the criteria are, or blame usually be taken into account, explaining this fact "an individual approach to each offer"/"any project", "expert opinions", "indicators elected by the board", "specific guidelines", etc., In addition to questions about the applied evaluation criteria, respondents were asked to indicate others not in form, but they use in the practice of evaluation criteria. No indication was accompanied by as much as 73 % of completed questionnaires. Much of the respondents indicated timing parameters included in slightly different expressions than the output in question (execution time, speed of execution, etc.). In terms of the issue of cost it was indicated in addition the price of delivery, payment terms, and various measures of cost of use and maintenance. Respondents declaring themselves as employees of the IT industry paid attention to the (expected) flexibility bidders opinions in the environment, their resources and their availability. Others, not proposed by the authors of the survey concerned the issue of references, experience, technologies used and the potential for cooperation, or even partnership. The survey also noted a significant number of positive answers (88 %) to the question "Do you agree with the statement:" I am ready to increase contract price in the settlement after completing his task in certain special cases "?" Most respondents (29.85 %) permits such an action when it is necessary to take an emergency action or in a situation where the cost of services is lower or the project to reach the other benefits (26.87 % of respondents). Almost 18 % of respondents answered that such a solution does not allow them the right, while 20 % of respondents are limited by procedures.

Decision support system supporting procurement decisions in project planning has in its intended form a new e-service available to users of the global network, involving the use of an electronic agent system for matching potential business partners defining the needs of the project undertaken and offering tools for multi-criteria assessment of potential bidders to select the most favorable offer or a decision of a production of product/service. Almost all respondents use in the process of purchasing at least one tool for supporting the management of projects such as web browsers, spreadsheets, internet price comparison websites or internal company tools. A few responders use in the making purchasing decisions the recommenders and online auction sites. The main objective of the project is to support the decision in projects through:

- Creating a tool to support the selection of the main contractor specific offer in the conditions of the plurality of evaluation criteria.
- Creating a tool for supporting the decision of an own production or buying a service or product
- Providing channels of communication between the principal contractor, subcontractor, and the project owner to the project
- Creating of a forum for contractors expanding area of business entities from the local market to the national/international

- Collecting data about the preferences of potential contractors in relation to the selection of patterns and schemes and analysis of potential offers.

32.3 Users

In general, all users in the proposed system can be divided into two main classes: internal and external users. The external users class includes all those to whom the proposed service is transferred as a optimization tool, while the internal users class includes administrators and all analysts who use the information from the database system and prepare the appropriate analysis and reports. The system is to provide an innovative service consisting in the main outline of the possibility of multiple-comparison, the selection of contractors, selection of offers, Make-or-Buy analysis and the completion of the transaction. The system in its assumptions should be the general system, without taking into account the particular specifications of any particular industry. The potential external users of the system will in fact be project owners, main contractors and their subcontractors. Due to the characteristics of potential customers in the system there will be defined following roles: project owner, main contractor and sub-contractor. Customers service can perform several roles simultaneously in the system. For example, the main contractor of the project may wish to delegate the implementation of some tasks to other subcontractors, thereby form a new project and will also be the owner of the project. At the same time a subcontractor in one project can act as the project owner or a contractor in different projects. The system provides the possibility for users lists of verified contractors. The owners of the project are representatives of stakeholders in the implementation of a specific project. In view of these users group, it is assumed that they tend to choose a few (dozen) potentially interesting offerings and technically uncomplicated ones (not needing specialized expertise, knowledge of the methodology or decision making tools) to compare these offers, assuming that the criterion of assessment is not only the price of a given model but also other defined by its performance. This group of users can also expect certain facilities for communicating with a potential principal contractor (the sender offers) in order to carry out the same transaction. They also have full access to the project's WBS and all its parameters. Main contractors interested in the implementation of a specific project can send offers to clients, owners and general contractors. They have no effect on the defined project, but during the implementation of the project they have the opportunity to provide a make-or-buy analysis and make a decision about self-realization of tasks or to outsource several tasks. This way they can create a new sub-project output. Subcontractors interested in running a particular task can create and send offers that are being analyzed and evaluated by the main contractors. Internal users of the system, besides administrators responsible for the proper functioning of the system, data analysts are mainly concerned with qualitative and quantitative analysis of a set of database accumulated by the system. They play a significant role in the process of improving the operation by analytical and optimization algorithms and external analytical services.

32.4 Constraints and Assumptions

Identification and analysis of the functional requirements of the system was carried out on the basis of the scenario method with the use of system structures and behavior modeling. Due to a large number of system functions, the behavioral model of the system is presented in a hierarchical use case diagrams, according to the top-down method. In this article, we present only the most important functional requirements of the designed system. The use of case diagrams has been prepared in accordance with the UML 2.4 specification. In order to define the limits of the product and preparing the model of the first level, first we identified all the actors involved in the operation of the system. There are basic users defined in the previous section and subsystems and tools that interact with the host system. The first group of actors we includes: Customers directly offered by the e-services, namely:

- Project Owners (PO)—institutional client interested in the implementation of the defined project WBS.
- Main contractors (PC)—institutional clients interested in the duties of the main contractor.
- Sub-contractors (SC)—Institutional clients bidding on individual tasks in the project.
- Data Analysts (AD)—experts in qualitative and quantitative analysis of a set of database accumulated by the system.
- System Administrators (ADM), supervising the maintenance of business continuity and security systems and implementing basic administration and office work based on the results of the system (e.g. invoicing customers).

The group of subsystems and tools consists:

- e-mail system (PRS)—a mail system of client or analyst participating in the process of registration and correspondence.
- analytical and reporting system (SyAR)—Satellite subsystem performing analyzes of substantive content databases and how they search for statistical processing and report generation market analysis.
- the analytical engine optimization (SiAO)—Satellite subsystem performing formal analysis of bids, offering customers preferences analysis tools and decision support (building ranking of bids).

32.5 Functional Requirements

For such defined actors the first level use cases diagram defined for these users of the system includes:

- Viewing information about the system and implemented solutions—available to every customer, with the information about the principles of the operation, descriptions of used tools, manuals, tutorials, training materials, promotional, etc.
- Browse and search for listings—the service available to every customer to be able to view available jobs and generating sales of database queries to search listings meet both predefined and asked the customer criteria.
- Preferences analysis, make-or-buy analysis, evaluation of bidders—the group of use cases are available for registered users, allows you to specify criteria for evaluation of bidders, analysis of their significance and determine the ranking of bids, and also an analysis of make-or-buy.
- Account Management—a service available to the registered customer, giving the opportunity to:

  – Create, review, edit, and print projects
  – Import and export projects to/from the most popular programs including project management tools such as MS Project
  – The ability to integrate tasks with MS Outlook calendar
  – Create, review, edit, and print inquiries
  – Create, review, edit, and print listings (offers portfolio)
  – Create, review, edit, and print a list of qualified contractors
  – Contacts between the general contractor and sub-contractor and between the project owner and general contractor

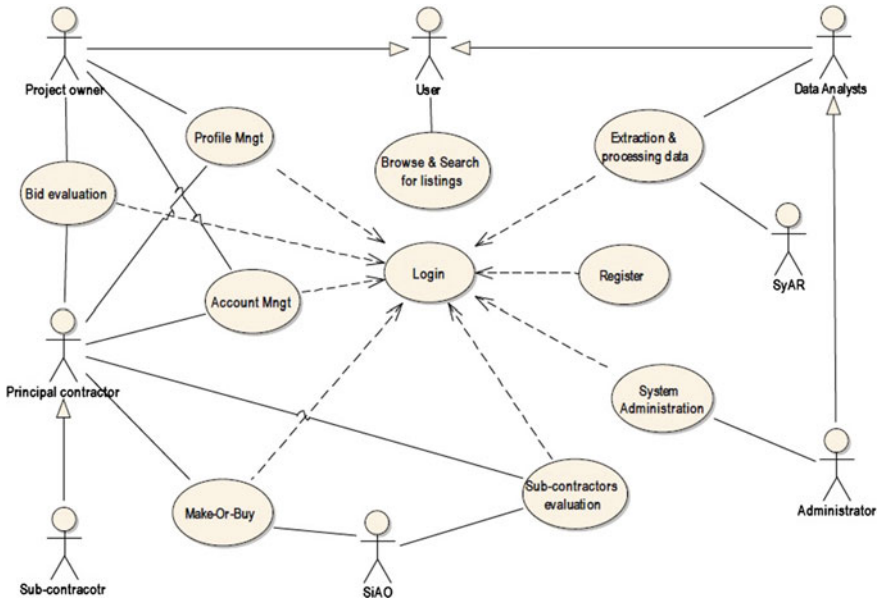


Fig. 32.1 General use case diagram of the system

  - Review the history of contacts with other service users

- Extraction and processing of the data—a service available to DA users allowing for access to the database and offering relevant tools to generate complex queries and basic statistical analysis.
- System Administration—the group of the use cases of the appropriate system administrator (ADM). Describes services related to the maintenance, data administration, preparation of reports for the administration department.
- Login and registration—use cases relevant to all service users allowing for the registration and access to individual accounts (in which additional services are implemented).
- Profile Management—use cases relevant to all registered users allowing users to edit and modify profile including personal data, corporate, managing access to verified bidders lists, browsing history, payments etc.

Identify the main actors and use cases possible to determine the relationship between them (in particular association) and to prepare a use case diagram of the first level, which is the general model of the system Fig. 32.1.

32.6 Non-functional Requirements

32.6.1 Ergonomic Requirements

From the point of view of ergonomics of the use of the system should be intuitive and have functional UI and menus. The product should be easy to use even for people with the minimal knowledge of information technology, or in the field of decision-making. They should use vocabulary and symbolism fully understandable for to the average Internet user. All technical aspects should remain hidden from the user. Verification of compliance with these requirements should be the maximum permitted time in which the user begins to move smoothly through the site. This time should not exceed a few minutes. The user must have the ability to display and expose these elements in which he is mostly interested, and hide those that are currently unnecessary (e.g. Free to modify the scope of information shown in the listing tables). The criterion for this requirement is the functionality of the software for all the users allowing for the for appropriate modification of the display of individual functional elements (menus, table formatting, etc.)

32.6.2 Performance Requirements

A very important group of non-functional requirements are performance requirements. They concern the critical issues that affect the evaluation of the system by

users and the willingness to use the site in the future. These requirements will mostly refer to the maximum time access to the site and implementing basic and advanced analytical operations on the system. Depending on the type of users we can distinguish the main performance requirements. The basic performance requirement is an implementation of all operations related to the service information and a database searching standard time similar to operations carried out in the most popular websites. Another requirement concerns the duration of the performance of the analytical operations carried out by the system. The make-or-buy analysis, evaluation of bidders and building preferences must be carried out in or near the time the regular site calls the database. Operations reporting and data analysis carried out by the system SyAR for the needs of users, depending on the complexity of the analysis should be preceded by adequate information about the estimated waiting time. Results of the analysis and reports should be available within the user accounts and email the information about their availability. These reports should be available to users in no more than 15 minutes of an order of the report. The requirements for the precision and accuracy of the calculations should be met 100 %. Any calculation errors are unacceptable and the accuracy of the statistical calculations and optimization algorithms should be considered as a sufficient level of interpretation to two decimal places. The requirements apply statistical analysis of their compliance with the methodology of multi-descriptive statistics and statistical analysis. The analysis must be made on the basis of the above basic methods. by statistical departments. The Reliability to the system requirements are extremely rigorous. Due to the realization of paid services and the specificity of their benefits, the online system should be able to work 24 h a day, 365 days a year. Maintenance work must be performed during the service operating at lower intensity (in the early hours of the morning) and disable access to services cannot be longer than 4 h. Any system failures—extremely undesirable—must be removed immediately, within fifteen minutes after the occurrence through the current upload of backup.

32.6.3 Requirements for the Working Environment

The proposed system will operate in a cloud environment (e.g. Amazon AWS). Due to the required performance and safety of the system it should be implemented in Java EE in accordance with the template MVC and using Spring framework. The system should be modular. The analytical module responsible for carrying out all the analyzes provided for in the decision-making system and a web module for communication with the user should be separate and independent units. For performance reasons and ease of scaling, the database engine should be considered NoSQL database (e.g. Redis). Communication with the database should be done with the use of the buffer, the size of which shall be determined by the administrator. The de-signed system should take into account the specific nature of the server on which it is installed. You should also decide on the type of the cloud environment (IAAS/PAAS). The detailed audit of the server security and functions must be made. The possibility

of continuous access to the site 24 h a day, 7 days a week must be ensured. Creating regular backups of both service applications and the database is required. These copies will be stored both on a separate server and data storage media deposited in the service owner premises.

32.6.4 Safety Requirements

Access to all functionality of the service, including the accessibility for every user the public area of the site must be recorded in the logs. There should be stored such information as user data including IP, date, time, time spent on each page, executed commands. Access to the paid part of the site will be possible after completion of the registration process and user authentication. The data in this particularly sensitive personal data must be adequately protected. Stored passwords must be encrypted and access to the database strictly limited and monitored. Each incoming data from the user to the site must be monitored and archived. Then, before using the data, they must be subjected to a process of filtration. In particular, it is important to ensure the safety of the service through the use of appropriate software security that is implemented at the level of the program code. Lack of adequate protection could lead to a break because of both applications and the attack on users. Firstly, there should be implemented safeguards against the risks associated with the injection and execution of malicious code. In particular there should be implemented the protection against SQL injection and Cross Site Scripting, Cross Site Request Forgeries, injection system commands, injection newline, transfer malicious code to a function interpreting and performing Java code modification path or file name processing hostile scripts from a foreign server, a modification of the names of dynamically created variables or functions. The second type of threat that must be taken into account during implementing the service are the threats arising from the manipulation of parameters. In particular, the system must be safe from attacks involving the manipulation of demand chain, form fields, headers, and HTTP request COOKIE values. Each data must necessarily go through the process of filtration. The parameters and the degree of data filtering restrictions must be washed ascertainable by the administrator. The next type of security is protection against unauthorized access to resources and the disclosure of confidential information. In particular, mechanisms must be used to prevent leakage of information, including blocking error messages and system messages blocking access to debug commands, not to place comments in the HTML code. The problem of information leakage can be seen also in another aspect. It concerns the disclosure of hidden or confidential data, files or documents, which themselves may be a subject to attack or they may contribute to its execution. The problem of unauthorized access to resources, which should remain confidential, mainly due to the fact that they are placed under the document root of the web server. This way you can access them directly by typing the exact URL in a browser window. Elements of the operating system that must be protected are: hidden files, backups and temporary files, source files. The application must also be protected by

the so-called Google Hacking. The internet search engine Google is now one of the largest available on the market. In its indexed resources there are millions of links to Web sites around the world. In addition, this tool has a number of options that allow for comprehensive and advanced search. It can be, and often is, used by crakers. Thanks to a crafted queries attackers can get a lot of useful information on the site. In addition, it must prevent the disclosure of source files. The disclosure of source files and other resources stored locally on the server is made possible by a technique called directory traversal. These attacks are prone to all scripts using template files or in some other way of referring to files on the server. This attack runs through the manipulation of the chain demands. Appropriate mechanisms for authentication and securing of the user session management must be used. The system must be secured to a dictionary attack, strength, power inverted attacks and attacks involving the interception of the session, the presumption session. Adequate procedures to verify the password in terms of its strength and blocking access to the account after crossing the threshold amount of incorrect logins must be implemented.

32.6.5 Legal Requirements

Determination of the legal requirements for the system requires legal consultation in many aspects of the operation of the service. They cover both the data protection rules, as well as issues related to cost the accounting principles services offered by the system, the principles of the free use of the system, the transmission of the information between the interested parties, contract, etc. In this respect, it is necessary to make a professional legal analysis.

32.7 The Process of Contract Aquisition

As we described in Sect. 32.3 users of this service can have one of three roles: the project owner, main contractor or sub-contractor. Once defined user has an opportunity to play different roles in several projects. The whole project or its parts (subprojects, tasks, sub-tasks, activities) could be defined by any user, whichever role it plays. It means that if any part of WBS, precedence or conditional relations has not been entered to the system by the project owner, who searches for the general contractor, it could be defined by potential bidders. This also applies to the auxiliary and preparatory processes, which may not be known to the ordering party. According to PMBoK the input data to the process should contain at least:

- Templates for the contract documents
- Scope statement
- Project Schedule
- Quality requirements

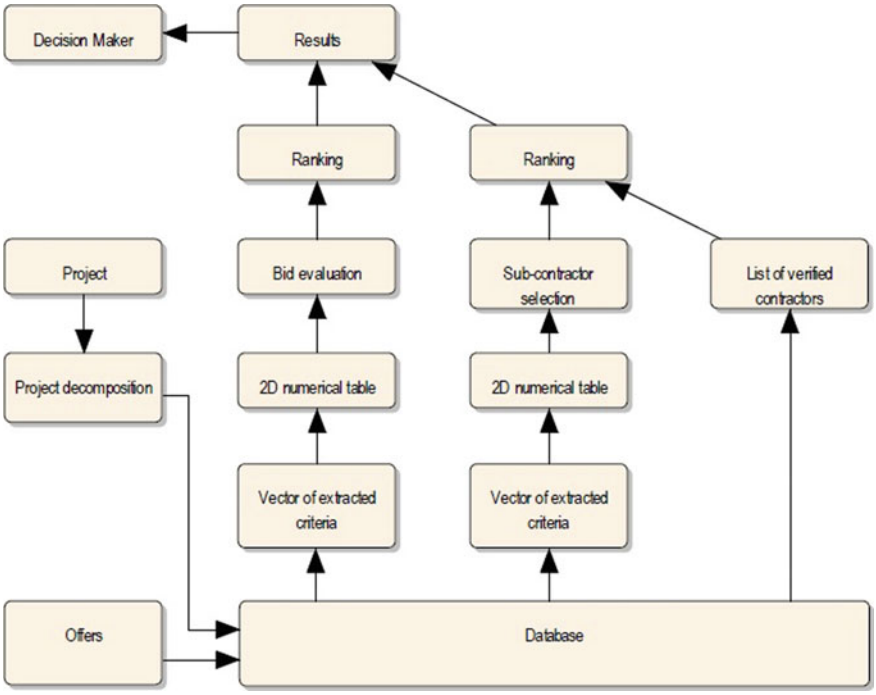


Fig. 32.2 The schema of data and methods used for decision making

- Time limitations
- Enterprises' environmental factors

The owner of the project defines the project after logging. The decomposition of the project into tasks must be done on the basis of the input. One must define the WBS, specification of works and Network Diagram. One of the functionality of the system is the ability to import project from popular project management support tools including e.g. MS Project. For such a user-defined project user can perform the Make-or-Buy analysis. As a result of this analysis, the user shall decide which of the tasks in the project will perform alone, and which will be outsourced. If the user is going to have some work performed for a sub-contractor he can use the service for the quoting process and define essential terms of the contract and then send an inquiry. Such an inquiry can apply to request a quotation, a request for a proposal and a request for a bid. Inquiries can be sent both to contractors who are on the list of verified contractors as well as to all the others in the database service. If a potential sub-contractor is not on the list of pre-qualified performers, the user has the possibility to carry out the process of qualification of subcontractors. The process flows as shown at the Fig. 32.2.

A positively verified subcontractor may enter the list of verified performers. The potential subcontractor receives the inquiry and prepares a response. He introduces

its parameters to the service. In addition to the introduction of appropriate offer parameters such as price or deadline, he also has the ability to attach additional files (e.g. in pdf format). Before bidding the contractor shall conduct its own analysis of Make-or-Buy, as a result of which he may decide to execute that task himself yourself or order it for another sub-contractor. Then, he has the role of the project owner or main contractor. The user defines the system as a new project under the terms received from the main contractor. The main contractor after receiving bids from potential sub-contractors shall analyze the bids and can carry out the qualification procedure for sub-contractors. During this analysis, the main contractor can determine the offer as satisfactory or unsatisfactory. At any moment, the main contractor has the ability to opt out of the subcontractor offer and decide to perform the task on his own. As we indicated in [3], outputs of the completed process should be covered by:

- A list of the client's qualified contractors (if provided)
- Contract acquisition
- Make-or-buy decisions
- Preliminary (conditional) agreement with potential (sub-)contractors

However, successful contract acquisition depends on the upper level decision-maker, this DSS supports this decision making process. If the contract documentation contains templates of agreement between ordering and executing parties and ranges of evaluated parameters (costs, deadlines,...) were a subject of negotiation, the finally agreed contract may introduce the mechanism of successful fees fund to motivate con-tractors and to optimize project parameters and risk allocation. Specified approach and models for different types of contracts were introduced in [19].

32.8 Conclusion

In this paper we presented an outline of the DSS dedicated to supporting the procurement and contract management processes for different types of projects. It uses a formalized record of critical chain concept in multi-criteria extension using the influence matrixes. Its versatility allows it to be used in web-version and individual users can optimize both the selection of contractors and the configuration tenders prepared by themselves, which should allow to obtain better performance of the entire project, including all levels of WBS. In the content we pointed references to the detailed methodological solutions contained in our previous works. A further stage of the system preparation will comprise its programming and testing.

References

1. Budziński R, Becker J (2015) Transformation of knowledge sources in decision support system. J Aut Mobile Robot Intell Syst 9(2):28–36

2. (2012) A guide to the project management body of knowledge, 5th edn. Project management Institute, USA
3. Błaszczyk P, Błaszczyk T (2015) On the support of the general contractor's decisions in the processes of event project management. In: Camillo A (ed) Global Enterprise Management, Volume I: New Perspectives on Challenges and Future Developments. (Chapter 11), Palgrave Macmillan, pp 189–204
4. Turner JR, Simister SJ (2001) Project contract management and a theory of organization. Int J Project Manage 19:457–464
5. Holt GD (1998) Which contractor selection methodology? Int J Project Manage 16:153–164
6. Arslan G, Kivrak S, Birgonul MT, Dikmen I (2008) Improving sub-contractor selection process in construction projects: web-based sub-contractor evaluation system (WEBSES). Autom Constr 17:480–488
7. Bubshait AA (2003) Incentive/disincentive contracts and its effects on industrial projects. Int J Project Manage 21:63–70
8. Lahdenperä P (2015) The beauty of incentivized capability-and-fee competition based target-cost contracting. Procedia Econ Financ 21:609–616
9. Boukendour S, Hughes W (2014) Collaborative incentive contracts: stimulating competitive behaviour without competition. Constr Manage Econ 32:279–289
10. Missbauer H, Hauber W (2006) Bid calculation for construction projects: regulations and incentive effects of unit price contracts. Eur J Oper Res 171:1005–1019
11. Bochenek J (2014) The contractor selection criteria in open and restricted procedures in public sector in selected EU countries. Procedia Eng 85:69–74
12. Goldratt E (1997) Critical chain. North River Press
13. Błaszczyk T, Nowak B (2008) Project costs estimation on the basis of critical chain approach (in Polish) In: Trzaskalik T (ed) Modelowanie Preferencji a Ryzyko '08, Akademia Ekonomiczna w Katowicach
14. Błaszczyk P, Błaszczyk T, Kania MB (2011) The bi-criterial approach to project cost and schedule buffers sizing in Lecture Notes in Economics and Mathematical Systems. New state of MCDM in the 21st century. Springer, pp 105–114
15. Błaszczyk T, Błaszczyk P (2012) New formal approach to project critical buffe. In: Kwong WS, Ng SH, Jiao R, Xie M (eds) Proceedings of the IEEE "IEEM", Hong Kong, pp 1070–1074
16. Błaszczyk P, Błaszczyk T, Kania MB (2013) Project Scheduling with Fuzzy Cost and Schedule Buffers in Lecture Notes in Economics and Mathematical Systems. In: Kim HK, Ao S-I, Rieger BB (eds) IAENG Transactions on Engineering Technologies. Lecture Notes in Electrical Engineering Volume 170, Springer, pp 375–388
17. Błaszczyk T, Błaszczyk P (2013) Selection of sub-contractors of the project while minimizing settlements of contractual penalties and success fees. In: Laosirihongthong T, Jiao R, Xie M, Sirovetnukul R (eds) Proceedings of the IEEE "IEEM". IEEM Society on Industrial Engineering and Engineering Management, Bangkok
18. Błaszczyk P, Błaszczyk T (2013) Project subcontractors selection in fixed price and cost-plus contracts in Lecture Notes in Engineering and Computer Science: The World Congress on Engineering and Computer Science, WCECS 2013, 23–25 Oct 2013, San Francisco, USA, pp 1131–1136
19. Błaszczyk P, Błaszczyk T (2014) Models supporting project subcontractors selection for main types of contracts. In: Kim HK, Ao S-I, Amouzegar MA (eds) Transactions on engineering technologies. Springer, pp 781–796
20. Błaszczyk T, Błaszczyk P (2015) "Contracting decisions in project management" An outline of the dedicated decision support system. In: Gen M, Kim KJ, Huang X, Hiroshi Y (eds) Industrial engineering, management science and applications 2015. Lecture Notes in Electrical Engineering Volume 349, Springer, pp 347–356
21. Błaszczyk T, Błaszczyk T (2015) The Description of a New Model of the Decision Support System for the Projects' Bid Configuration with Multiple Sub-contractors Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering and Computer Science 2015, WCECS 2015, 21–23 Oct 2015, San Francisco, USA, pp 1024–1029

Chapter 33
HIV/AIDS Humanitarian Supply Chain Management: The Case of Zimbabwe

Talent Tatenda Chingono and Charles Mbohwa

33.1 Introduction

The Supply of Medicines and related commodities is the key to any successful health system, this insures the availability of essential commodities and equipment to enable testing, treatment, care and support. This supply chain is governed by the government and partner NGOs [3]. Most NGOs distribute the commodities to the countries' district and provincial hospitals and clinics. Patients would then collect the drugs from these hospitals and clinics. NatPharm, a government owned parastatal, is a major enabler of this as it caters and deals primarily with the general public seeking attention at general/government facilities [15].

Literature has shown that relief humanitarian commodities are always in short supply. HIV/AIDS in Zimbabwe and its associated or related circumstances such as food shortages require humanitarian intervention. In fact, 9 out of every 10 disasters that NGOs responds to are events that may have a major impact on areas of the world, yet do not get the media attention that other natural or ma-made disasters attract [1]. The crisis in Zimbabwe is a great example. This Southern African country has seen a 32 % increase in the number of individuals who do not have enough to eat, and 2.2 million Zimbabweans representing 25 % of the rural population needed help between January and March 2014 to avoid hunger/hidden hunger. This problem is a result of drought, which has led to a poor harvest and high food prices. The country is also in the region with the highest HIV/AIDS prevalence and fatality rate, this further complicates the situation.

T.T. Chingono (✉) · C. Mbohwa
Department of Quality and Operations Management, University of Johannesburg, Bunting Road Conbas, Auckland Park, South Africa
e-mail: ttchingono@uj.ac.za

C. Mbohwa
e-mail: cmbohwa@uj.ac.za

S.-I. Ao et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-981-10-2717-8_33

The major objective of this study was to determine both the specific nature of the humanitarian supply chain. Analyse and identify deficiencies and problems of a humanitarian logistics and supply chain management nature that exist in Zimbabwe.

33.2 Existing Literature Relevant to the Study

Auxiliary information uncovered that scholars have mostly dwelled on coming up with strategies aimed at introducing philanthropic logistics and supply chain management to Aid organisations. These basically wanted to make their operations quicker and cheaper. These researchers include Clark and Culkin [2], Davidson [4], Näslund and Williamson [12], Thomas [17], Van Wassenhove [24], Kleindorfer and Van Wassenhove [8], Thomas and Van Wassenhove [20], Tomasini and Van Wassenhove [21], Thompson [18]. Others recommended supply chain Analytics for philanthropic logistics conversion [16] also likewise keeps tabs on the significant part about philanthropic logistics, Networks to Africa done backing to the execution of the thousand years advancement objectives. These need additionally concentrated once describing those real occasions of the philanthropic act, starting with begin to complete. They likewise summarized their accomplishments [5–8, 10, 14, 17, 18, 23, 25].

33.3 Methodology and Research Design

The research survey was conducted in Zimbabwe. Most NGOs and other Organizations with their Zimbabwean head offices for in Harare were surveyed. Participation was voluntary, allowing all willing organizations to participate in the survey [11]. This minimized the element of bias in the sample of respondents. The sampling procedure by its nature was assumed to be random. Research involved administering questionnaires, conducting interviews and meetings with the relevant managerial, supply chain or logistics staff at head offices of willing organisations to identify and document weak points in the logistics and supply chain of HIV/AIDS commodities. 150 Questionnaires were distributed and 105 responses were recorded. Secondary data was also gathered so as to identify strengths and existing capacity in the HIV and AIDS commodity management, and determine improvements required.

33.4 Results and Discussion

Supply chains were mostly managed and sustained through close partnerships with suppliers. Electronic Data Interchange (EDI), Subcontracting and the use of external suppliers were also quite popular. Very few organisations used the Vertical integration method of supply chain management.

The success of organisation in managing their supply chain was also captured. Most organisations indicated that they were often successful, 56 % made this indication. This also tends to point out that in some cases, but fewer, their supply chain was not successful in providing the much needed aid to beneficiaries in good time.

Most organisations were satisfied with their methods of supply chain management but in planning for the future, most organisations intend to start implementing third party logistics and vertical integration, this was the case for mostly for those that were not satisfied. The main advantage of the vertical integration is the increased control, to increase or reduce production, at will, so there is more supply chain coordination.

From the 105 responses recorded, more than 75 had a logistics department present and around 15 did not. Logistics is that part of the supply chain process that plans, implements, and controls the efficient, effective forward and reverse flow and storage of goods, services, and related information between the point of origin and the point of consumption in order to meet customers' requirements. It is pivotal for Humanitarian organizations to have such a department which ensure goods reach their intended destination in good time. Most organizations also had a clear logistics plan. It also has advantage of quick analysis, and can incorporate the complex tradeoffs.

The major benefit that was recorded was that of flexibility and being able to predict future requirements. This helps ensure sustainability of the supply chain. Other notable benefits included better quality information and more accurate costing.

Most participants indicated that they faced a serious problem when it comes to skills shortage, vendor support and integration with customers system, but these were rather rare when compared to responses of those facing a little problem mostly in Hidden cost.

Most organisations plan to introduce the standard package of e-business as everyone is equal on this platform, when you are interacting with your clients or vendors using e-business, it is difficult for them to tell how big your business is. With a professionally developed and maintained website, any small organization can look as impressive on the Internet as the large corporations do. It is a level playing field that helps create opportunity for small organizations.

Important measures that must be promoted and implemented include closer cooperation between companies, improved information provision, better infrastructure and more funding and support amongst others.

Generally forecasting and quantification for national HIV/AIDS commodities is mostly done by the AIDS and TB Unit with assistance from Natpharm. It is normally driven by funds available for procurement and final quantities to order. This is mostly not accurate as there is no systematic process for forecasting national commodity requirements.

Government is encouraged to establish a well-defined and sustainable systematic process for forecasting and quantifying national commodity requirements. Better and more accurate forecasts will only happen when important logistics data on supplies distributed to victims is collected and reported on a routine bases.

According to the survey, the facilities resupply quantities are determined digitally through computer. Currently, most donors are using their own procurement

mechanisms to bring HIV and AIDS commodities into the country. These are not well coordinated thus they impose the risks of duplication and gaps in commodities availability. It is of great importance that procurement is coordinated centrally through information sharing among different NGOs, government and donor agencies. This can be done through encouraging information sharing via the already existing Procurement and Logistics Subcommittee or via the National ARVs programme. Procurement of commodities should also be coupled with other functional members of the logistics system, including forecasting and quantification, inventory management and calculating stock status, within the pipeline at all times. Procurement was mostly driven by the availability of funds hence more funds should be made available to organisations such that risks and possibilities of stock outs and defaulting are reduced [16].

Most Organisations indicated that they did not have adequate storage space and space for future expansion. Most organisations had 24 h security guards and lock and key. In some key areas there are security cameras. Unannounced audits of High value goods especially of HIV/AIDS Commodities, normally quarterly were also noted. In store, high value goods are caged in a section of storage and in transit, truck cabins are locked.

Good handling of the goods limits loses and breaking of the stock. Good TC management in the room. Losses and theft are minimised due to the handling procedures implemented. Regular stock take for fault moving items and FIFO for stock management. Regular stock taking to reduce stock theft. Stored under recommended condition. Storeroom maintained in good tender care. In most facilities there was good security for stored and in-transit goods, some even have cameras and Bio-metric systems. To reduce expiries, the IT system is configured to pick items whose expiry is nearest. Zimbabwe seems to be doing well in terms of security and storage, but one could also note that it was over emphasised at most International organisations. The security was more like that of a military camp, with armed guards and strict procedures restricting access to responsible personnel who could assist in the survey. One could deem this an unnecessary expense. These funds should also be channelled to the humanitarian cause.

Most organisations relay and use their own facility managed transport. This also enables them to have more control of their supply chain. Most organisations do not have adequate carriers to cover a nation-wide campaign or relief programme with the average number being 5. The box plot also explains further carrier management in surveyed organisation. Most respondents had between 3 and 8 carriers. Half of the participants had less than 5 (Figs. 33.1 and 33.2).

Relief agencies most used their own fleet and it will also mostly be a truck load. No participants indicated that they use rail or ocean transport. Rail as tends to be slow and ships cannot be used as the country is landlocked.

A large number of participants indicated that most supply chains were complex and long as it need at least ten contact point from the source of aid to the victim. Reducing these would help their operations to be faster and smoother (Fig. 33.3).

Most organisations that participated in the survey agreed that Reliability was most important to improve as this will ensure that victims receives their much

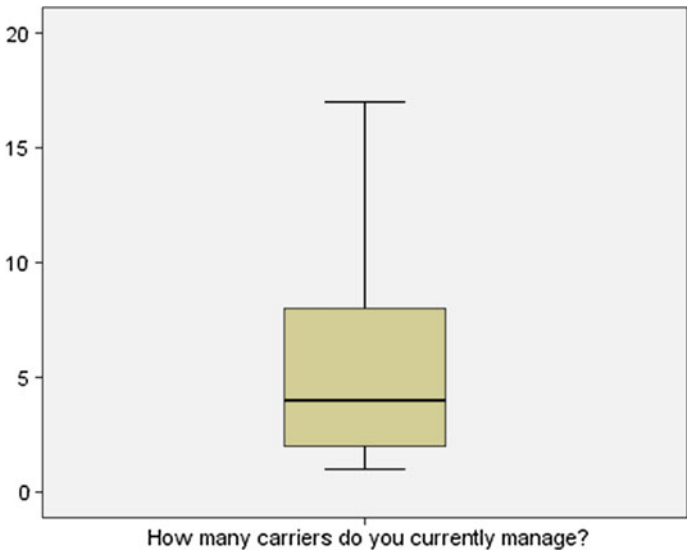


Fig. 33.1 Box plot of carriers managed

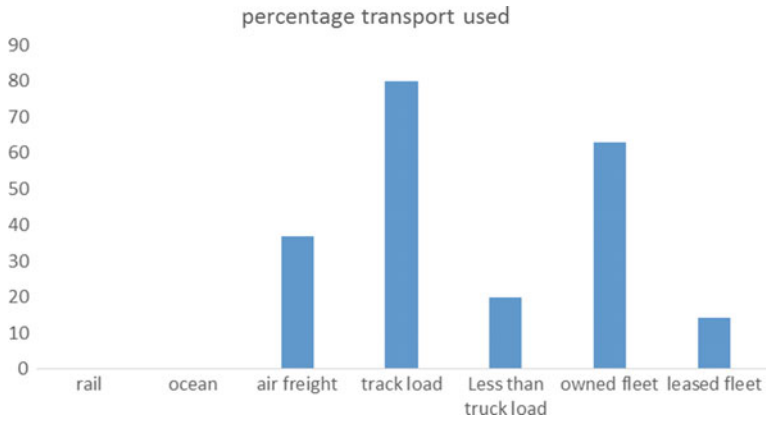


Fig. 33.2 Type of transport used

needed drugs and commodities on time and every time. This also ensures that no patient defaults and becomes resistant to treatment (Fig. 33.4).

A moderate number of respondents indicated that availability is also important to improve. More than 26 supported this point, they insist that drugs should be readily available for them to distribute in time before patients default (Fig. 33.5).

Transit was also of substantial importance to some organisations. This might be through immigrations or on the road. These indicated that some immigration laws need to be modified in order for commodities to move faster. Others needed more

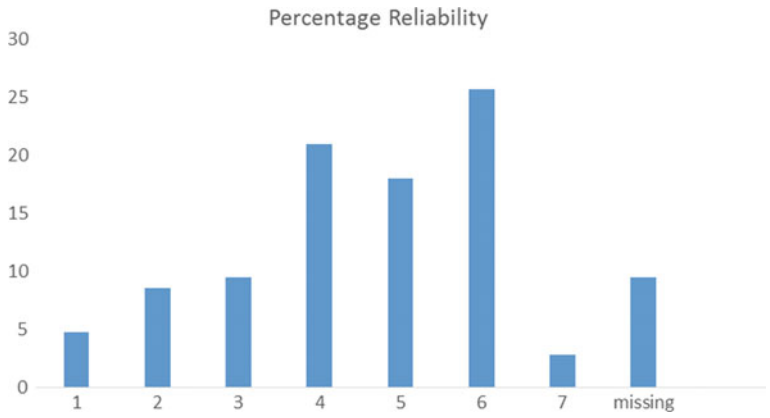


Fig. 33.3 Importance of improving reliability

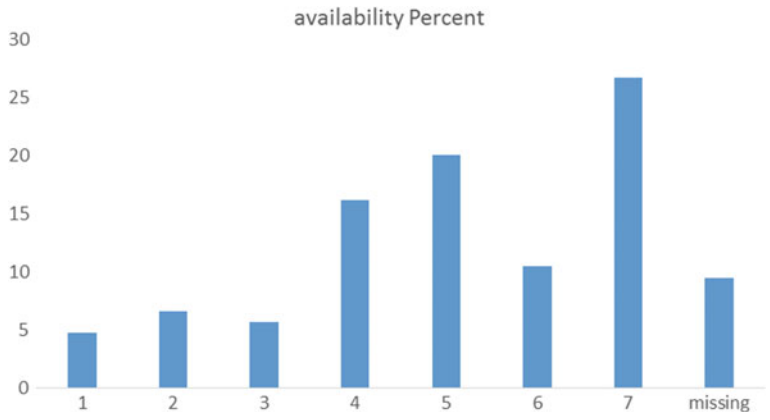


Fig. 33.4 Importance of improving availability

carriages and some suggested that some major roads need to be maintained and repaired (Fig. 33.6).

Most organisations indicated that whenever they had received commodities from donors/source they always delivered in time. Thus they indicated that this was the least important to improve (Fig. 33.7).

More than three quarters valued the importance of improving expense above 5. This might raise some questions as one might argue that most if not all of their commodities are from donor organisations. This might be expense in relation to their day to day operational overheads and also in labour remuneration (Fig. 33.8).

Customer service/victim also must be improved greatly if the fight against the epidemic is to be won. In some instances some donors were considered to be the customer by most NGOs and they tried to meet stringent regulations in Order for them to continue receiving aid to distribute (Fig. 33.9).

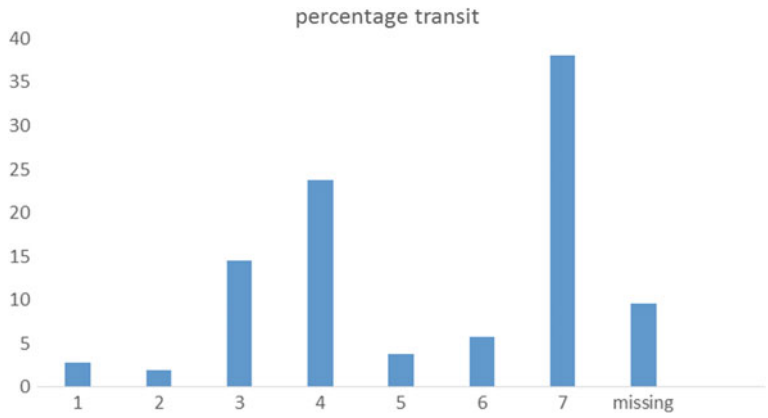


Fig. 33.5 Importance of transit

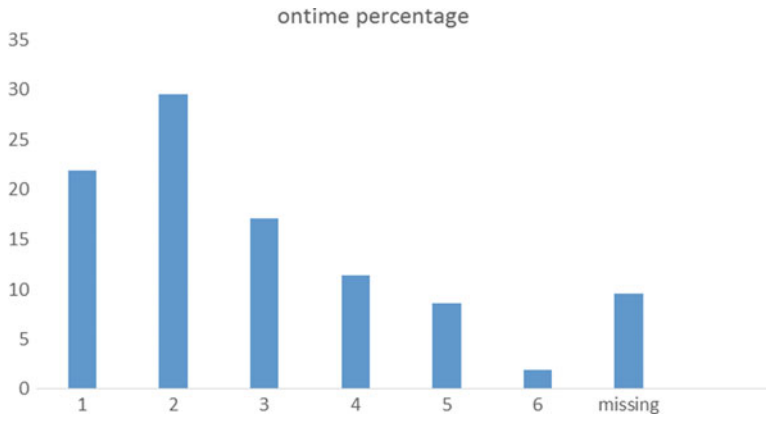


Fig. 33.6 Importance of on-time delivery

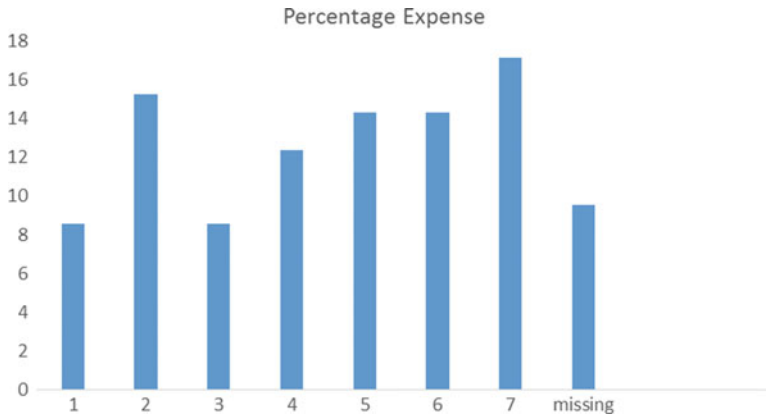


Fig. 33.7 Importance of expense

Fig. 33.8 Importance of customer service

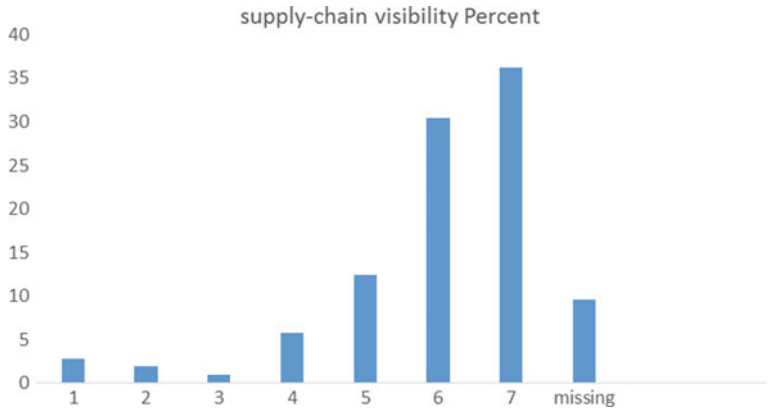


Fig. 33.9 Importance of visibility

Visibility of the supply chain also needs to be greatly improved such that managers can fully know and understand where goods are and where there are needed mostly. The goal is mainly on strengthening and improving the supply chain by providing real time data on commodities location to stakeholders, including victim/customer.

Organisations are embarking on new efforts to remain relevant and competitive in the market and find it necessary to manage their knowledge capital. Knowledge management practices are concerned with the creation, capture, handling and transfer (sharing) of knowledge held by individuals within the organisation. These activities provide extra value to the organisation in addition to its core functions, and would require involvement from an organisational wide perspective to truly deliver universal benefits.

Identified Gaps:

- NGOs are more passion driven with lack of the "know how" and a need for hand-holding by more experienced organizations
- A need for capacity building for upcoming and new NGOs
- Financial challenges, with a need to build credibility to potential funders
- There is no credible NGO platform (e.g. forum)
- The tendency of working in silo and competing instead of complementing each other's work
- Underestimation of the importance of research, and documentation

Identified opportunities

- Experienced NPOs to mentor new and upcoming organizations
- Establishment of a regulated body without government dependence

33.5 Conclusions and Recommendations

The success of humanitarian operations is hard to measure. Economic success is the standard performance measure in the pro-profit world. For non-profit organizations this evaluation is more complex, considering difficult-to-formulate elements such as unmet need fulfilled and more tractable ones like cash flow. Keeping complete track, control and accountability of the humanitarian programs and their outcomes is challenged by the high urgency and pace of this type of operations, and time for analyzing and recording is usually tight [3].

Increased humanitarian aid is necessary, but so too is a move to a more proactive emergency approach based on a recognition of the severity of the crisis in all its forms. Donor agencies and governments and must try and ensure that the provision of humanitarian aid remains distinct from political processes. Their policies towards Zimbabwe must not be implemented at the expense of the humanitarian imperative to ensure that Zimbabweans have unhindered access to the assistance they need to survive. The government of Zimbabwe must give humanitarian agencies the space and support they need to function independently [3].

References

1. Beamon BM, Balcik B (2008) Performance measurement in humanitarian relief chains. Int J Public Sector Manage 21(1):4–25
2. Clark A, Culkin B (2007) A network transshipment model for planning humanitarian relief operations after a natural disaster. Presented at EURO XXII—22nd European Conference on Operational Research, Prague
3. Chingono TT, Mbohwa C (2015) Management of the humanitarian supply chain in Zimbabwe. In: Proceedings of the world congress on engineering and computer science 2015,

WCECS 2015. Lecture notes in engineering and computer science, 21–23 Oct 2015. San Francisco, USA, pp 1030-1034
4. Davidson LS (2006) Key performance indicators in humanitarian logistics. Engineering Systems Division GBG
5. Global AIDS Responds Progress Report (2012) Zimbabwe country report, Reporting period: Jan 2010–Dec 2011
6. Gray R, Oloruntoba R (2006) Humanitarian aid: an agile supply chain? Supply Chain Manage: Int J 11(2):115–120
7. Kerina D, Stray-Pedersen B, Muller F (2013) HIV/AIDS: the Zimbabwean situation and trends. Am J Clin Med Res 1(1):15–22
8. Kleindorfer PR, Van Wassenhove LN (2004) Managing risk in global supply chains in Strategies for building successful global businesses, Chapter 12. In: Gatignon H, Kimberley JR (eds) Cambridge University Press, Cambridge, Massachusetts, pp 288–305
9. Kovács G, Spens KM (2007) Humanitarian logistics in disaster relief operations. Emerald Group Publishing Limited
10. Mentzer JT, DeWitt W, Keebler JS, Min S, Nix NW, Smith CD, Zacharia ZG (2001) Defining supply chain management. J Bus Logist 22:2
11. Näslund D (2002) Logistics needs qualitative research—especially action research. Int J Phys Distrib Logist Manage 32(5):321–338
12. Näslund D, Williamson SA (2008) Supply chain integration: barriers and driving forces in an action research-based inventory intervention. Supply Chain Forum: An Int J 9(1):70–80
13. Noguera M, Alt D, Hirschhorn L, Maponga C, Osewe P, Sam-Abbenyi A (2003) Zimbabwe: antiretroviral therapy program—issues and opportunities for initiation and expansion. DELIVER/John Snow, Inc., for the U.S. Agency for International Development, Arlington, VA
14. Nyaguthie (2008) The important role of humanitarian logistics. Oxfam-GB, Pretoria, South Africa
15. Rickard J (2006) Humanitarian logistics: musing aloud. Monday Dev 24(20):6–7
16. Takang E, Dragana V, Celestine K, Jabulani N (2006) Management of HIV and AIDS commodities in Zimbabwe—A capacity assessment of NatPharm and Ministry of Health and Child Welfare
17. Thomas A, Kopczak L (2005) From logistics to supply chain management: the path forward in the humanitarian sector. Technical Report, Fritz Institute, San Francisco, CA, USA
18. Thompson (2008) Supply chain analytics for humanitarian logistics transformation. In: Humanitarian logistics: Network for Africa Rockefeller Foundation Bellagio Center Conference, Bellagio, Lake Como, Italy
19. Tomasini R, Van Wassenhove LN (2004) Genetically modified food donations and the cost of neutrality. Logistics response to the 2002 Southern Africa food crisis. INSEAD Case 03/2004-5169
20. Tomasini R, Van Wassenhove LN (2009) From preparedness to partnerships: case study research on humanitarian logistics. Int Trans Oper Res 16:549–559
21. Tomasini R, Van Wassenhove L (2009) Humanitarian logistics. CPI Antony Rowe, Chippenham and Eastbourne
22. UNDP (2010) Human Development Report-Real wealth of Nations: pathways to human development
23. UNIDO (2011) Pharmaceutical sector profile; Zimbabwe. Global UNIDO Project: Strengthening the local production of essential generic drugs in least developed and developing countries
24. Van Wassenhove LN (2006) Blacket memorial lecture. Humanitarian aid logistics; Supply chain management in high gear. J Oper Res Soc 57(5):475–489
25. Zimbabwe National HIV and AIDS strategic Plan 2011-15 (2011)

Chapter 34
Techno-Econometric Analysis of Membrane Technology for Biogas Upgrading

Samson O. Masebinu, Esther Akinlabi, Edison Muzenda, Charles Mbohwa and Akinwale O. Aboyade

34.1 Introduction

To reduce greenhouse gas (GHG) emissions and secure a sustainable supply of energy globally, the utilization of the earth's non-renewable resources must be reduced, while more sustainable and efficient alternative energy sources are developed. From an array of environmentally friendly energy alternatives, synthetic gas, natural gas, hydrogen and biofuels have been identified as the most promising substitutes for vehicular fuel in the nearest future [1]. Of the aforementioned fuels, biofuels which include, all liquid and gaseous fuels; ethanol, methanol, biodiesel, hydrogen and biogas, derived from processing of biomass, is considered the most environmentally friendly energy source but of importance to this study is biogas. Biogas comprises mainly CH_4, CO_2, smaller traces of acidic gases and impurities such as H_2S, N_2, water vapour and volatile organic gases (VOCs) [2]. For biogas to be suitable as vehicular fuel, it must be upgraded by removal of CO_2 and other trace

S.O. Masebinu (✉) · E. Akinlabi · C. Mbohwa
Department of Mechanical Engineering Science, University of Johannesburg, Auckland Park, South Africa
e-mail: somasebinu@uj.ac.za

E. Akinlabi
e-mail: etakinlabi@uj.ac.za

C. Mbohwa
e-mail: cmbohwa@uj.ac.za

E. Muzenda
Department of Chemical Engineering, University of Johannesburg, Auckland Park, South Africa
e-mail: emuzenda@uj.ac.za

A.O. Aboyade
Energy Expert, Innovation Hub, Pretoria, South Africa
e-mail: aaboyade@uj.ac.za

S.-I. Ao et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-981-10-2717-8_34

impurities. After upgrading, biogas is generally referred to as biomethane which can be used for all applications designed for natural gas.

The interest in biomethane stem out of the fact that locally produced fuel alleviate, to a large extent, the reliance on imported fuel, ensure stability in fuel supply and prices, enhance national security and develop the economy. Biomethane fueled vehicle emits approximately 20–25 % less CO_2 as compared to gasoline fueled vehicle. This is due to the fact that the carbon to hydrogen ratio of CH_4 is approximately 52 % lower than the ratio in gasoline. Though energy density which is an important factor for automotive fuel is much higher in liquid fuel than for gases, however, this disadvantage can be partially offset by compression of biomethane (CBG). To implement an economically viable biomethane project, some key factors must be considered. They are market for the product; availability of organic waste & collection strategy; location of upgrading plant; upgrading technology; transportation of product/distribution network; financial model & financing; permit; and regulatory framework. Of the aforementioned factors, transportation of product/distribution network, financial model, permit and regulatory framework are the major barrier in the deployment of biomethane as vehicular fuel in South Africa. Despite the challenges been faced in the deployment of biomethane, the South African government is keen to exploit the possibility of biomethane as vehicular fuel to reduce her carbon footprint and provide an alternative fuel source for its transportation sector.

34.1.1 Membrane Technology for Biogas Upgrading

A membrane separation process is used to split a feed mixture containing two or more species through a semi-permeable barrier where the species that moves through the barrier is called permeate and the other that does not move through is called retentate [3]. Various membrane materials are able to separate CO_2 and CH_4 and both polymeric as well as inorganic materials can be used. However, in industrial scale gas separation, only polymeric membrane materials are applied due to their low manufacturing cost when compared to inorganic materials. In order to apply membranes on an industrial scale, large membrane area are normally required [4]. Hollow fiber membrane is commonly used in industry due its high packing density, low investment and operating cost. However, pre-treatment process is always required when used because it is very susceptible to fouling and difficult to clean [5]. Critical operating parameters that affects the quality of upgraded biogas and CH_4 recovery when using membrane are the operating pressure, feed flow rate, and composition of the feed. Membrane configuration has been reported to be of single stage without recycle, multistep single stage and multistage configurations. The extent to which a feed mixture can be separated in the single stage is limited. Thus, in order to achieve a higher degree of separation, cascades of membrane modules with recycle are often used. Multistage configuration of two stages or more has been reported to yield higher CH_4 recovery and product purity though with higher investment cost due to additional membrane area and compression cost [6].

34.1.2 Economics of Biogas Upgrading

The economic assessment performance of any given configuration of separation processes varies and depends very much on the assumptions used in the assessment. Economic considerations include information on total investment cost, annual variable operating and maintenance cost, annual cost of CH_4 lost in the plant and annual capital related cost. The GPC is the total cost incurred to produce a cubic meter of biomethane. The GPC is influenced by the scale of the plant, technology adopted, location and operating process conditions. According to SWEA data, the investment cost of membrane installation for biogas plant of 100 m^3/h of biomethane is in the range of \$8,150 to \$8,500/(m^3 biomethane/h). For the same capacity of the installation with water scrubbing equipment, the price is \$11,250/(m^3 biomethane/h) and \$11,600/(m^3 biomethane/h) for biogas plant with PSA. As the volume of produced biomethane increases to 500 m^3/h, the investment cost reduces to about \$3,900/(m^3 biomethane/h) [2]. GPC is reported to decrease as the volume of feed biogas increases but generally, GPC is roughly in the range of \$0.1 to \$0.7/m^3 of biomethane.

34.1.3 Conversion of Vehicles

The conversion kits consist of fuel storage cylinders and bracket, fuel lines, regulator, a fuel-air mixer, pressure reducer and a switch that allows the driver to alternate between gasoline and CBG manually. The conversion cost ranges between \$2,700 to \$5,500 for 4–8-cylinder engine in medium size car and vans. While the conversion cost for heavy duty truck ranges between \$5,300 and \$10,600 [7]. In the international market, the cost of light duty OEM NGVs is higher than gasoline vehicle in the range of \$1,900 to \$4,500 depending on the national tax regime for new vehicle while price increase for medium duty commercial vehicle ranges from \$6,500 to \$9,000 depending on the type of vehicle and its application [8]. For heavy duty vehicle, the price has been reported to be higher by 20–25 % the cost of its diesel engine equivalent [8].

34.1.4 Life Cycle Analysis

The life cycle analysis (LCA) of cost, energy demand and GHG emissions are important component in assessing deployment of any vehicle fuel. LCA of vehicle fuel include their extraction, processing, transport, utilization and emissions. A well-to-well (WTW) analysis describe a complete cycle for vehicle fuel. The GHG savings achieved in the production and utilization of biomethane varies considerably but generally, it depends on digested substrate, substrate transport

distance, chosen digestion technique, production capacity, upgrading technique and end use equipment efficiency. Biomethane produced from municipal waste and animal manure has been reported to achieve GHG savings approximately 50 % and 80 % respectively when compared to conventional fossil fuel [9]. Using biomethane as fuel for vehicle, Brightman and Parker [10] reported a lifecycle CO_2 reduction of 49–63 %.

34.1.5 Research Objective

The usual target of an enrichment process using membrane is to produce a retentate stream essentially stripped of CO_2 and with a low flowrate but highly CO_2 rich permeate stream. This article presents a simulation of two possible membrane configurations of an existing biogas upgrading plant that uses hollow fiber membrane module for CH_4 enrichment with descriptive detail of the cleaning process. The simulation was conducted to investigate the performance of the configurations on product purity, product recovery and compressor power requirement. We also assessed the economic viability of upgrading biogas to biomethane and its utilization as transport fuel.

34.2 Methodology

The biogas composition used in the upgrading process were data from a closed landfill at Sebenza in the Ekurhuleni Municipality of South Africa [11]. The biogas composition is as follows; CO_2 38.9 %; CH_4 55 %; O_2 0.4 %; N_2 5 %; H_2S 0.002 % and H_2O 0.66 %. The feed is at atmospheric pressure and operating temperature of 27 °C.

34.2.1 Process Simulation and Data Collection

The simulation of the upgrading process was done using ChemCAD, a steady state process simulator, from Chemstation. The PSA unit, a dynamic and unsteady state process, was represented with a component separator unit operation and was constrained to plant result for H_2O and H_2S removal.

The process operating conditions were data collected from the plant. The membrane intrinsic properties (selectivity and permeability), configurations and module design data were not shared due to intellectual property infringement and confidentiality concerns. Intrinsic properties were obtained from literature. The module design data were verified by the plant manager and was accepted to be inconformity with industrial standard.

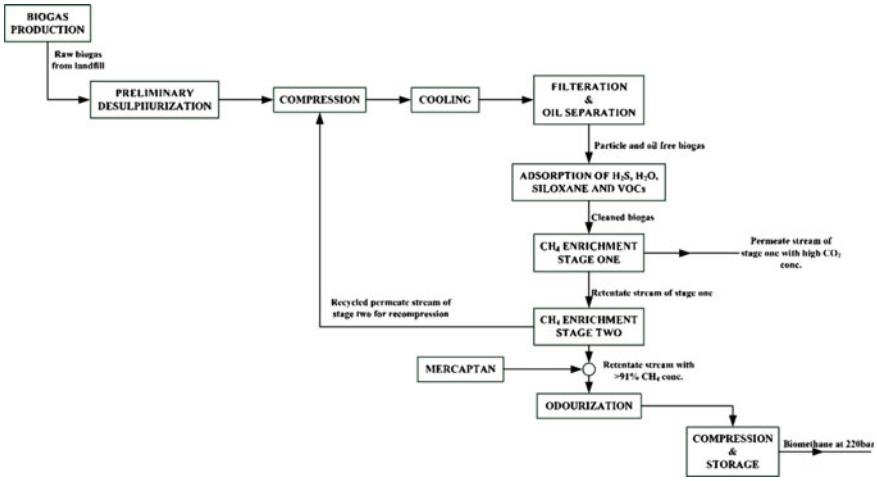


Fig. 34.1 Block flow diagram describing the process

A built-in membrane model for gas permeation available in ChemCAD 6.5.5 was used for the CH_4 enrichment. Almeesoft Engineering gas permeation software was also interface with ChemCAD during sensitivity analysis to study the effect of varying the process conditions on CH_4 recovery and product purity. The important process parameter specified for the simulation of the membrane are given in Ref. [12] and represents base condition for the simulation. Two membrane configurations were considered for this study; single stage without recycle (SSWR) and double stage with permeate recycle (DSPR) as shown in Fig. 34.1.

34.2.2 Process Economics

Hao et al. [13] procedure for calculating natural gas processing cost per cubic meter was modified in [11] to include the cleaning cost for biogas upgrading process. The cleaning cost is largely dependent on the composition and volume of the biogas to be treated [14].

Equipment sizing procedures and other cost such as equipment erection, piping, instrumentation, electrical, building, utilities and site development cost were also factored into calculating the fixed cost following guidelines in Refs. [15, 16]. The profitability of the process was determined using the net present value (NPV), internal rate of return (IRR), and the benefit cost ratio (BCR). A project life span (t) of 20 years was taken as the basis for calculation. Since the real value of money changes over time, discounting factor based on the prime lending rate (r) of 9 %.

34.2.3 Cost Analysis of Biomethane as Vehicular Fuel

The biomethane fuel consumption of a MBT in this work was calculated based on gasoline energy equivalent required to cover the same distance under the same operating condition. To estimate the price of biomethane in per litre of gasoline equivalent, Eq. 34.1 was used.

$$\frac{P_m}{M\%}/L = P_{GE} \tag{34.1}$$

where, P_m = price of CH_4 (CNG) at pump; M% = CH_4 concentration in biomethane (product purity); L = the energy content of 1 Nm^3 of 100 % CH_4 in gasoline; P_{GE} = price of biomethane in per liter of gasoline energy equivalent (GEE). An average distance of 64,000 km per year travelled by gasoline MBT with fuel consumption of 6.62 km/l was adopted as a model. From the field survey carried out in South Africa, the cost for converting a Toyota Quantum 16 passenger seat MBT to bi-fuel ranges from R20,000 to R60,000.

34.3 Result and Discussion

34.3.1 Simulation Validation

Figure 34.1 shows block flow diagram of the plant. The process flow diagram is presented in [11, 12]. The plant had 91.16 % CH_4 recovery with 87.2 % CH_4 purity. SSWR yielded 87.47 % CH_4 recovery with 77.89 % CH_4 purity while DSPR yielded 89.01 % CH_4 recovery with 84.59 % CH_4 purity. The simulated DSPR process suggests a good approximation to the plant data with a percentage error of 2.64 %. The error in the comparison could be attributed to higher CO_2/CH_4 selectivity of the plant membrane and feed pressure drop within the membrane module.

34.3.2 Methane Recovery and Product Purity

34.3.2.1 Effect of %CO_2 in Feed

Figures 34.2a, 34.3a and 34.4a show the effect of increased CO_2 content in feed on CH_4 recovery, purity of product and membrane area required for separation on both configurations respectively. 89.8 % CH_4 recovery was achieved with about 10 % CO_2 in the feed stream in the SSWR configuration requiring 25.7 m^2 of membrane area to achieve 89.85 % CH_4 in the product. In DSPR configuration, CH_4 recovery was 90.48 % with a product purity of 90.69 % requiring 51.75 m^2 membrane area for 10 % CO_2 in the feed. As the CO_2 concentration increased, CH_4 recovery reduces

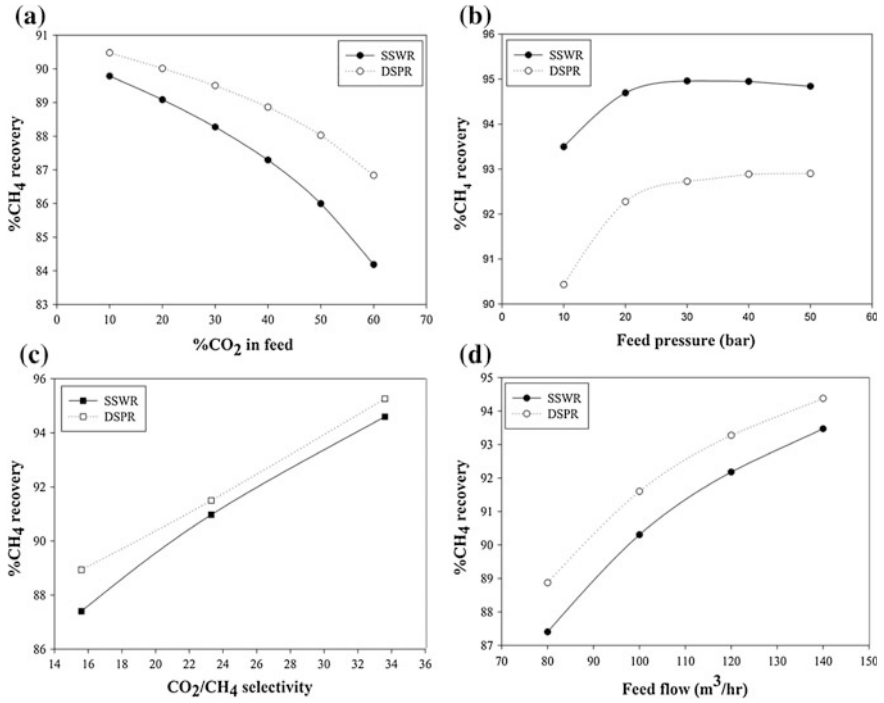


Fig. 34.2 **a** Effect of $\%CO_2$ in feed, **b** Effect of feed pressure, **c** Effect of membrane selectivity, **d** Effect of feed flow rate

as well as purity of product but with an increase in membrane area. With 60 % CO_2 in the feed, CH_4 recovery has reduced to 84.19 % with a 70.54 % CH_4 in the product while for DSPR, 86.8 % CH_4 recovery was achieved with 82 % CH_4 in the product. Membrane area increased to 27 m^2 for SSWR while DSPR required 53.58 m^2. DSPR achieved higher CH_4 recovery and purity over SSWR because more membrane area was available for separation. Also the recycling of permeate of the second stage membrane separation ensures that less CH_4 is lost in DSPR configuration.

34.3.2.2 Effect of Feed Pressure

Figures 34.2b, 34.3b and 34.4b show the effect of feed pressure on CH_4 recovery for both configurations, product purity and membrane area required respectively. The process condition and membrane shell data remains the same as earlier described except for the feed pressure been varied. It was observed as shown in Fig. 34.2b that increasing the feed pressure increased CH_4 recovery until a limit is reached where further increase in the pressure does not improve the CH_4 recovery rate. As the feed pressure increases from 10 to 30 bar, the CH_4 recovery increases rapidly from 93.5 to 94.96 % for SSWR and thereafter the increase in pressure

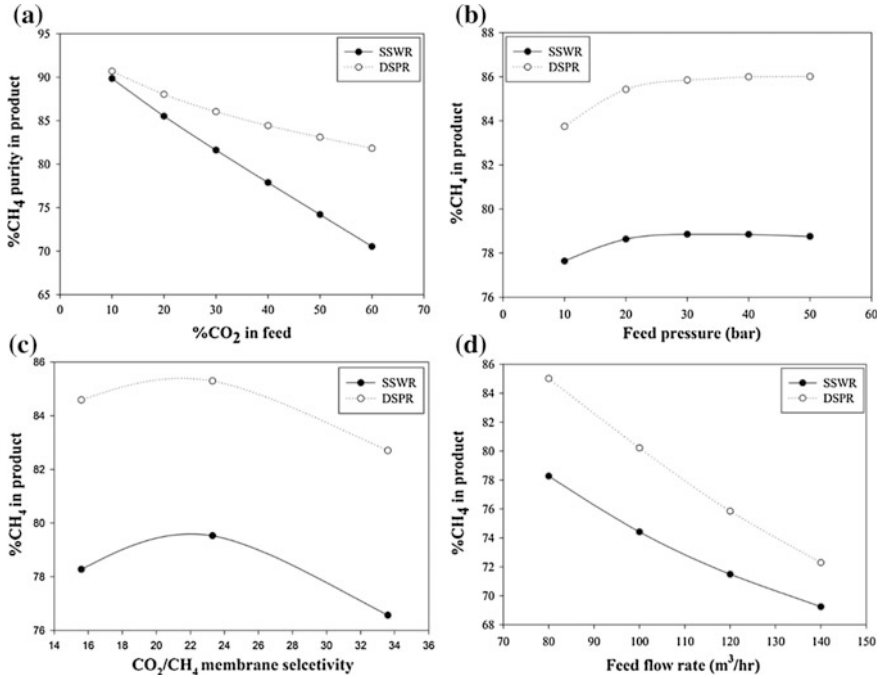


Fig. 34.3 Effect of varying process parameters on product purity **a** Effect of $\%CO_2$ in feed; **b** Effect of feed pressure, **c** Effect of membrane selectivity, **d** Effect of feed flow

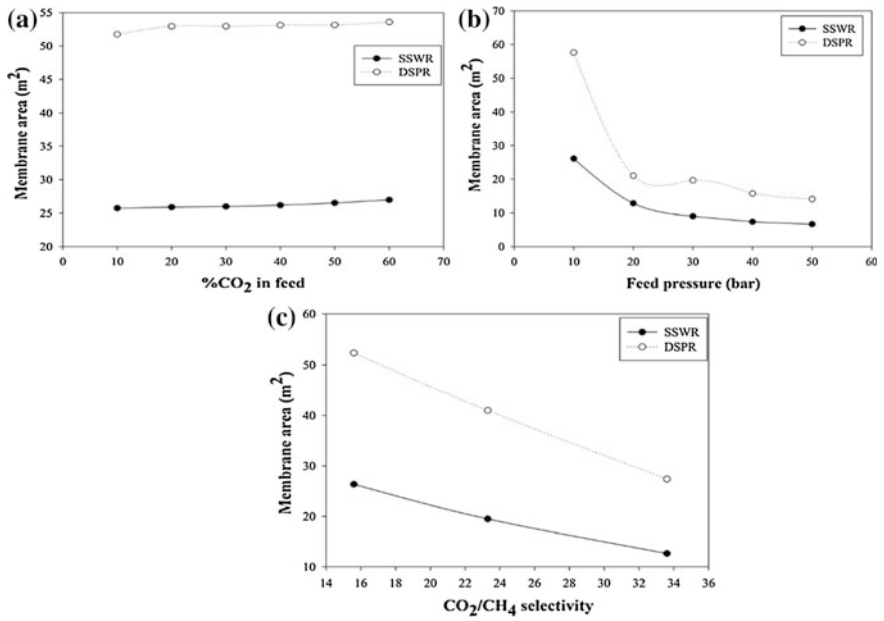


Fig. 34.4 Effect of variation of process variables on membrane area **a** Effect of $\%CO_2$ in feed, **b** Effect of feed pressure, **c** Effect of membrane selectivity

causes a gradual reduction in CH_4 recovery to 94.84 % at 50 bar. DSPR also achieved an increase in CH_4 recovery from 90.4 to 92.7 % as feed pressure increased from 10 to 30 bar. Further increase in pressure only increased the CH_4 recovery by 0.2 %. CH_4 recovery was higher for SSWR when compared to DSPR because after the first stage separation, the feed pressure to the second stage which is the driving force reduces and causes a reduction in separation efficiency of the second stage as compared to the first stage. But despite a lower CH_4 recovery, the product purity achieved in the DSPR, 83.75–86.01 %, when compared to SSWR, 77.6–78.8 %, as shown in Fig. 34.2b, was much higher. At lower pressure, the high packing density advantage of hollow fiber membrane becomes increasingly important due to the high membrane area available hence higher purity can still be achieved. The increase in CH_4 recovery and purity of product can also be attributed to the fact that the increased pressure creates a greater driving force across the membrane surface area while the almost constant CH_4 recovery after its peak, in this case 30 bar is due to the membrane module design.

Improving the overall structural design will increase the membrane performance with variation in process condition [17]. This suggests that achieving high CH_4 recovery does not only depend on improving the process condition but also ensuring that an appropriate equipment design is implemented to accommodate changes in process condition. Also a higher CH_4 recovery was achieved and the required membrane area decreased as the feed pressure increases as shown in Fig. 34.4b. As the feed pressure increases from 10 to 30 bar, the membrane area for SSWR decreases from 26 to 9 m^2 after which increasing feed pressure to 50 bar only causes a reduction to 6.7 m^2. DSPR uses higher membrane area for separation which ensures high product purity is achieved as feed pressure increases until it reaches its peak. The behaviour of DSPR is also similar to SSWR, as feed pressure increase to 30 bar, there as a significant reduction in membrane area required from 57.6 to 19.6 m^2, after which an increase in pressure resulted in fluctuation in membrane area.

This fluctuation suggests that the module has reached its pressure limit with respect to module design. It may be economically justified in some circumstances to increase the compression pressure of the feed stream to reduce membrane area cost and improve recovery but the structural design of the membrane casing to safely accommodate such pressure must be considered. Hence a feed pressure of 18 bar will be considered as an optimal and safe operating pressure for the DSPR based on the module design data provided.

34.3.2.3 Effect of Membrane Selectivity

Membrane intrinsic properties have high influence on CH_4 recovery which increases with an increase in the selectivity of the membrane. Figures 34.2c, 34.3c and 34.4c shows the effect of selectivity on CH_4 recovery, product purity and membrane area required respectively. An increase in the membrane selectivity from 15.8 to 33.3, while other process conditions were held constant, resulted in a spontaneous increase in the CH_4 recovery from 87.4 to 94.6 % for SSWR while DSPR increased from 88.9

to 95.3 % as shown in Fig. 34.2c. The product purity for both configurations increased as selectivity increased from 15.6 to 23.3, SSWR increased from 78.3 to 79.5 % while DSPR increased from 84.6 to 85.3 %. A reduction was observed as the selectivity increases to 33.6 with SSWR CH_4 purity in product reducing to 76.6 % and DSPR to 82.7 as shown in Fig. 34.3c. The reduction could be attributed to selectivity properties of the membrane towards CO_2. If less CO_2 permeate through the membrane, the CO_2 concentration in the product stream increases hence the product purity decreases. Also it has been reported in literature that highly permeable polymers have low selectivity and vice versa [12]. Hence a general trade off relationship between gas permeability and selectivity is required.

Aside increase in CH_4 recovery, increasing the selectivity reduces the required membrane area for separation as shown in Fig. 34.4c. With 15.6 CO_2/CH_4 selectivity, membrane area required to achieve 87.4 % CH_4 recovery was 26.38 m^2 for SSWR and at 33.3 selectivity, 12.63 m^2 was required to achieve 94.6 % CH_4 recovery. For DSPR, 52.40 m^2 was required to achieve 88.9 % CH_4 recovery with 84.6 % product purity and at 33.6 selectivity, 27.69 m^2 was required to achieve 95.3 % CH_4 recovery. The decrease in membrane area is due to the intrinsic property of the membrane to permeate more CO_2 over the initial contact area of the membrane and the feed gas.

34.3.2.4 Feed Flow Rate

Figures 34.2d and 34.3d shows the effect of increasing feed flow rate on CH_4 recovery and product purity. The plant been studied has the capacity to produce between 20 and 100 m^3/h biomethane. The base feed flow rate been used for this study was 80 m^3/h. Increasing the feed flow rate, CH_4 recovery increased from 87.4 and 88.9 % at 80 m^3/h to 93.5 and 94.4 % at 140 m^3/h feed flow for both SSWR and DSPR respectively as shown in Fig. 34.2d. The product flow rate peaked at 103 m^3/h for SSWR and 100.5 m^3/h for DSPR. This shows that if the plant is operating at full capacity, more CH_4 will be recovered in the process and CH_4 lost will be reduced. The product purity reduces as the feed flow rate increases as shown in Fig. 34.3d. CH_4 in product reduces from 78.27 to 69.24 % for SSWR while DSPR had a reduction from 85.02 to 72.29 %. It was also observed that as the feed flow rate increases from 80 to 140 m^3/h, there was a slight increase in the membrane area. The membrane area increased from 26.38 to 28.4 m^2 for SSWR and from 53.2 to 54.6 m^2 for DSPR. The reduction in product purity as feed flow increases was due to insufficient membrane area for separation hence more CO_2 is not permeated and its presence in the retentate reduces the purity of the product.

34.3.3 Compressor Power Requirement

The effect of feed composition, feed pressure and feed flow rate on compression power required for both configurations has been investigated. Figure 34.5a–c shows

the effect of increased $\%CO_2$ in feed, feed pressure and feed flow rate on compressor power required for separation respectively. As CO_2 increases in the feed stream from 10 to 60 %, the compression power required to facilitate the transportation of CO_2 through the membrane surface increases from 8.89 to 34.62 hp for SSWR and 17.05–53.39 hp for DSPR as shown in Fig. 34.5a. The required power per pressure increase is the minimum driving force to ensure that CO_2 permeate through the membrane surface area available and the target CH_4 recovery was achieved. Increasing feed pressure, 10–50 bar, increased the required compressor power from 22.33 to 38.82 hp for SSWR and 39.78–69.68 hp for DSPR as shown in Fig. 34.6b. As earlier discussed, increasing feed pressure reduces the required membrane area but the limit of the feed pressure depends on the membrane module design pressure capacity. Increasing feed flow rate increased the compressor power required for separation as show in Fig. 34.5c. As the feed flow rate increases from 80 to 140 m^3/h, the compressor power required for SSWR increased slightly from 22.18 to 25.22 hp because there was no recycling of the permeate stream. $\%CO_2$ in the recycled stream of DSPR increased from 60.62 to 79.18 % which increased the compressor power from 37.73 to 46.26 hp. As earlier discussed, increasing CO_2 in the suction of the compressor increases the required power for membrane separation to achieve the desired CH_4 recovery.

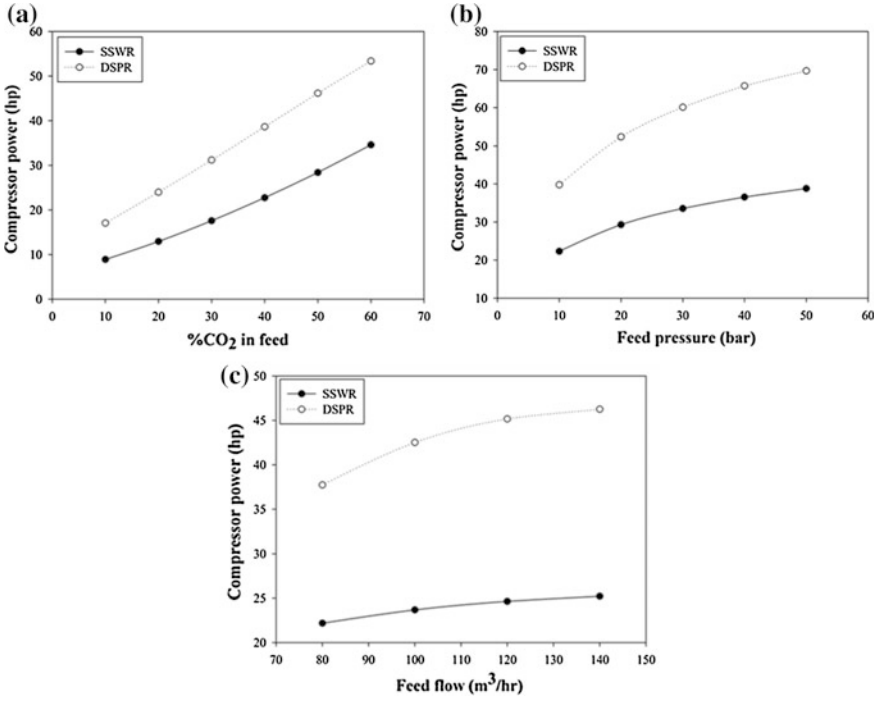


Fig. 34.5 Effect of variation of process parameters on compressor power **a** Effect of $\%CO_2$ in feed, **b** Effect of feed pressure, **c** Effect of feed flow

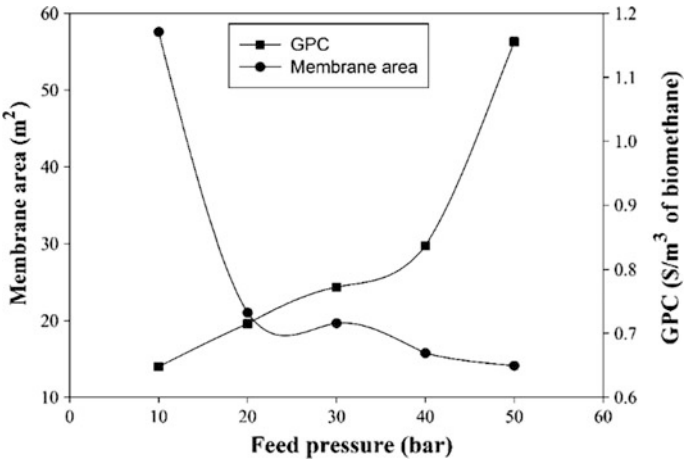


Fig. 34.6 Effect of feed pressure on GPC

The result of the parametric sensitivity analysis is consistent with previously published simulated work of Deng and Hagg [18], and Chmielewski and Urbaniak [2] on membrane performance in methane enrichment process.

34.3.4 Effect of Process Conditions on GPC

After consideration of both configurations, DSPR configuration yielded better technical result. We investigated the effect of variation in process conditions on the GPC for economic viability of the process.

34.3.4.1 Effect of Feed Composition

It was observed that an increase in CO_2 content in the feed from 10 to 30 % increased the GPC gradually from \$0.73 to \$0.79/m^3 of biomethane. The observation was assumed to be due to gradual but steady increase of the 2-staged compressor power from 12.68 to 23.11 kW. With increasing CO_2 in the feed, the fraction of lost CH_4 was reduced. This reduction should have reduced the GPC but the compression power was increasing in the progression of the increasing CO_2 in the feed; hence the slightly increase in GPC. An increase in CO_2 from 30 to 60 %, significantly increased the GPC from \$0.79 to \$1.39/m^3 of biomethane as the 2-stage compressor power increased by 72 % as compared to 30 % of CO_2 in the feed. The increase in the compressor power was linked to an increase in CO_2 permeating through the second stage of the membrane that increased from 32.1 to 70.2 % in the recycled stream.

34.3.4.2 Effect of Feed Pressure

The effect of feed pressure on GPC is as shown in Fig. 34.6. Increasing feed pressure from 10 to 50 bar increased the required compressor power. As the compressor power increased from 29.66 to 51.96 kW, the GPC increased from \$0.65 to \$1.16/m^3 of biomethane. An increase in feed pressure increases the CO_2 permeating through the first stage membrane, thereby a highly CH_4 concentrated stream is fed to the second stage membrane. Between 10 bar and 30 bar, the GPC increased from \$0.65 to \$0.77 of biomethane. The sharp increase in GPC from \$0.77 to \$1.16/m^3 of biomethane as feed pressure increases from 30 to 50 bar was as a result of the low rate of reduction in the membrane area which only reduces from 19.65 to 14.09 m^2 as compared to the rate of area reduction from 57.6 to 19.65 m^2 when feed pressure increased from 10 to 30 bar. A reduction in membrane area and CH_4 lost in the purge stream lowers the impact of the compressor power on GPC when compared to the GPC for increasing CO_2 in the feed. However, the optimal operating pressure for a membrane module will be determined by the pressure limit of the module design since generally increasing pressure increase CH_4 recovery up to a limit beyond which the module may structurally fail and reduction in membrane area won't be proportional to the increase in feed pressure. Figure 34.6 shows the intersection point at 20 bar of feed pressure to membrane area which is the maximum pressure limit of the module. Beyond the intersection point, increasing feed pressure increased the GPC with an insignificant reduction in membrane area when compared to the rate of membrane area reduction between 10 and 20 bar. Therefore, limiting the feed pressure below 20 bar will reduce the membrane area required for separation and limit the sudden increase in GPC.

34.3.4.3 Effect of Feed Flow Rate

Maximising the plant production capacity to produce 100 m^3/h biomethane, reduces the GPC. At feed flow of 80 m^3/h the GPC is \$0.79/m^3 of biomethane with a product flow rate of 45.99 m^3/h. As the feed flow rate increases to 140 m^3/h, the GPC reduces by 100 % to \$0.39/m^3 with a product flow rate of 100.55 m^3/h. The reduction in GPC was due to a reduction in the CH_4 lost in the upgrading process which reduces by 30 % in the purge stream, and a 119 % increase of the product flow rate. A slight increase in the 2-stage compression power from 28.14 to 34.50 kW was observed which was due to the increase in the feed flow rate.

34.3.5 Cost of Upgrading

To optimize the process, the feed pressure and membrane area were the two factors considered. Generally, increasing the membrane area makes more economic sense since it is cheaper than increasing the feed pressure which increases the

Table 34.1 Simulation result of the optimal process

Parameters	Simulation result
Feed flow rate	140 m^3/h
Feed pressure at 1st stage and 2nd stage	14, 11 bar
Temperature and pressure at permeate side	30 °C, 1.08 bar
Membrane area required	89.26 m^2
Upgraded biogas flow rate	81.26 m^3/h
CH_4 purity	91 % vol.
CH_4 recovery	96 %
CO_2 recovery in purge	98 % vol.
Compression duty	141 kW

compression power and also requires the membrane module to be designed to accommodate such pressure without failure. Also, the cost of electricity in South Africa has been increasing. Therefore, it will be economical to reduce any recurrent expenditure such as utilities in the optimization process. The membrane area was increased to accommodate the increasing feed flow rate and achieve the desired product. The feed pressure was slightly increased to 14 bar from 10 bar to increase the driving force of the feed stream across the increased membrane area. Table 34.1 shows the optimized process condition for simulated process configuration. The feed composition was as defined earlier, membrane CO_2/CH_4 selectivity of 15.6 was maintained while target CH_4 purity was 91 %.

The process increased the product purity by 4.4 % and recovery by 5.4 % with a product flow rate of 81.26 m^3/h as against 46 m^3/h of the base case. A membrane area of 89.26 m^2 was used to achieve 91 % CH_4 concentration in the product stream and 98 % CO_2 concentration in the purge stream. Table 34.2 shows the summarized economic impact of the optimized process. In the simulated process, the GPC was approximately \$0.42/Nm^3 (R4.48/Nm^3) of biomethane. The plant in this study dispense CBG on production site, hence cost of transporting CBG was not incurred.

Table 34.2 Economics of the process

	USD
Total plant investment cost	1,352,452.21
Fixed cost	930,434.05
Start-up cost	12,627.18
Annual variable operating and maintenance cost	126,271.75
Utility cost	5,928.77
Annual cost of CH_4 lost in permeate	9,829.70
Annual biogas lost	1,429.98
Gas processing cost/Nm^3	0.4172
Annual capital related cost	267,965.01

Table 34.3 Profitability of using biomethane as vehicular fuel to end user

Comparison	Gasoline	CNG	CBG
Litre/year	9,770	9,770	9,770
Annual fuel cost	R138,153.35	R97,606.22	R90,897.95
Annual savings		R40,547.13	R47,255.40
Investment payback (years)		1.5	1.3
Percentage saving		29 %	34 %

34.3.6 Profitability of Upgrading and Utilization

The pump price of CNG (92 % CH_4) in Johannesburg was R9.90/litre while 95 grade octane gasoline was R14.4/litre as at 15th June, 2014 [19]. The dispensing price for the biomethane based on 91 % CH_4 concentration was R9.30/liter GEE. At the end of the 20th year, the discounted cumulative NPV will be R15,240,343. Based on the prevailing prime lending rate of 9 %, the project will break even in the 5th year after plant start-up. The IRR is 22.41 %. An interest rate greater than IRR will make the investment not viable within the project life. The net discounted benefit of the project is R29,767,996.63 while the net discounted cost is R14,527,653.22. The ratio of the benefit to cost (BCR) is 2.05. The overall investment is profitable as indicated by the NPV, IRR and BCR if the biomethane is been sold on site and waste is provided at no cost as assumed in this analysis.

To the end user, a base case of a gasoline MBT was used to determine the profitability of using biomethane as vehicular fuel. Parameters for annual distance, distance travelled per liter of fuel, and conversion cost has been earlier defined. A MBT covering a distance of 64,680 km/annum at 6.62 km/l will require 9,770 l of fuel. The cost effect for the volume of fuel for using gasoline, CNG and CBG is as calculated in Table 34.3.

For using CNG over gasoline, the end user will save 29 % of annual fuel cost which is approximately R40,547 per annum with payback period for converting the taxi into bi-fuel of approximately one and half years. Using CBG over gasoline, the end user saves 34 % of annual fuel cost which is approximately R47,255 with a payback period of 1 year and 3 months. The dispensing cost of CNG includes all taxes and cost of transportation to gas station which makes it slightly expensive than CBG. This calculation indicates that using CBG as vehicular fuel for taxis is a profitable venture when sold directly from the production site.

34.4 Conclusion

Technical operational study and economic evaluation of a biogas upgrading plant has been presented. The performance of DSPR yielded a better result but with higher membrane area and compression power. Economic studies investigating the

effect of three process conditions on gas processing cost (GPC) for DSPR was conducted. An increase in CO_2 and feed pressure increased the GPC. At optimized condition, the calculated GPC is \$0.46/m^3 of biomethane. The NPV, IRR and BCR for producing biomethane were R15,240,343, 22.41 % and 2.05 respectively with a break-even in the 5th year. Using CBG over gasoline, the end user saves 34 % of with a payback period of less than 2 years for vehicle retrofitting. The fuel prices differential between gasoline and CBG as well as the overall savings, makes CBG attractive to the producer and end user.

Acknowledgements The authors wish to acknowledge the University of Johannesburg for providing funding for this project, the industrial partners, for granting access to their plant.

References

1. Nigam PS, Singh A (2010) Production of liquid biofuels from renewable resources. Prog Energy Combust Sci 37(1):52–68
2. Chmielewski AG, Urbaniak A, Wawryniuk K (2013) Membrane enrichment of biogas from two-stage pilot plant using agricultural waste as a substrate. Biomass Bioenergy 58:219–228
3. Bagajewicz M, Ryden C, Baumgarner H (2009) Use of membrane in gas conditioning
4. Kase Y (2008) Gas separation by polymide membranes. In: Norman NL et al (eds) Advanced membrane technology and applications. Wiley, Canada, pp 581–598
5. Mulder M (1996) Basic principle of membrane technology, 2 edn. Kluwer Academic Publisher, Twente, Netherlands, p. 577
6. Spillman R (2003) Economics of gas separation membrane processes. In: Noble RD, Stern SA (eds) Membrane separation technology: principles and application. Elsevier, Beverly, USA, pp 589–668
7. Clarke S, DeBruyn J (2012) Vehicle conversion to natural gas or biogas. In: Factsheet AM (ed) Ministry of agriculture, food and rural affairs. Ontario, Canada, p 8
8. Sanne M, Seisler JM (2008) How to implement a biomethane project: Decision maker's guide. Europe Natural Gas Vehicle Association, Sweden, p 64
9. EUCAR, CONCAWE, JRC (2007) Well-to-wheel analysis of future automotive fossil fuels and powertrains in the European context. European Commission, Joint Research Council, p 88
10. Brightman T, Parker T, Matthews C (2011) LowCVP biomethane for transport (HGV cost modelling), in biomethane for transport-HGV cost modelling. Transport and Travel Research Ltd., United Kingdom, p 48
11. Masebinu SO, Muzenda E, Aboyade A (2015) Economic analysis of biogas upgrading and utilization as vehicular fuel in South Africa. In: Proceedings of the world congress on engineering and computer science 2015, WCECS. Lecture notes in engineering and computer science. IAENG, San Francisco, USA
12. Masebinu SO (2015) Parametric study and economic evaluation of a simulated biogas upgrading plant. In: Chemical engineering. University of Johannesburg, Johannesburg, South Africa, p 130
13. Hao J, Rice PA, Stern SA (2008) Upgrading low-quality natural gas with H_2S- and CO_2-selective polymer membranes Part II. Process design, economics and selectivity study of membrane stages with recycle stream. J Membr Sci 320:108–122
14. Hao J, Rice PA, Stern SA (2002) Upgrading low-quality natural gas with H_2S- and CO_2-selective polymer membranes Part I. Process design and economics of membrane stages without recycles streams. J Membr Sci 209:177–206

15. Berstad E (2012) Testing and optimization of PVAm/PVA blend membranes for biogas upgrading. In Chemical engineering and biotechnology. Norwegian University of Science & Technology, Trondhiem, p 112
16. Sinnott RK (2005) Chemical engineering design, 4 edn. Coulson & Richardson's chemical engineering series. Coulson & Richardson's, New York, USA, p 1055
17. Makaruk A, Miltner M, Harasek M (2010) Membrane biogas upgrading processes for the production of natural gas substitute. Sep Purif Technol 74(1):83–92
18. Deng LY, Hagg MB (2010) Techno-economic evaluation of biogas upgrading process using CO_2 facilitated transport membrane. Int J Greenhouse Gas Control 4(4):638–646
19. Department of Energy. Fuel prices June. [electronic] 04 June 2014 [cited 17 June 2014]; http://www.energy.gov.za/files/esources/petroleum_fuelprices.html

Chapter 35
Validation of a Novel Method for Ethyl Lactate Separation Using Langmuir Hinshelwood Model

Okon Edidiong, Shehu Habiba and Gobina Edward

Nomenclature

LA	Lactic acid
E	Ethanol
W	Water
EL	Ethyl lactate
T	Temperature (K)
$n_{LA,0}$	Initial molar amount of lactic acid (mol)
k_i	Adsorption coefficient
k	Rate constant (mol g^{-1} min^{-1})
K_{eq}	Reaction equilibrium constant
C_i	Concentration of i at the surface of the catalyst (mol/L)
t	Time (min)
r	Reaction rate (mol g^{-1} min^{-1})
C_s	The concentration of vacant site on catalyst surface
S	Vacant site on catalyst surface
LH	Langmuir Hinshelwood (LH)
ER	Eley-Rideal
PH	Pseudo-homogeneous

O. Edidiong · S. Habiba · G. Edward (✉)
Centre for Process Integration and Membrane Technology (CPIMT),
School of Engineering, The Robert Gordon University, Aberdeen AB10 7GJ, UK
e-mail: e.gobina@rgu.ac.uk

O. Edidiong
e-mail: e.p.okon@rgu.ac.uk

S. Habiba
e-mail: h.shehu@rgu.ac.uk

S.-I. Ao et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-981-10-2717-8_35

35.1 Introduction

Ethyl lactate is a biodegradable and non-toxic material with excellent solvent properties which could potentially replace halogenated and toxic solvents for a broad range of consumer and industrial uses. In the food industry, approximately 85 % of the estimated 50,000 tons/year of ethyl lactate manufactured worldwide is being used for food product, while the remaining percentage is used in other industries such as textile, cosmetic, leather and pharmaceutical industries [1]. The esterification reactions involving lactic acid and alcohols are not only used in the manufacture of the corresponding esters, but are also used in lactic acid purification process [2]. The industrial manufacture of esters by esterification of acid with alcohol was first performed in a continuous stirred tank reactor (CSTR) and later in a catalytic distillation column in the presence of cation-exchange resins [3]. According to Altiokka and Citak [4], several studies have shown a lot of interest on the investigation of the esterification reactions between carboxylic acids and alcohols catalyzed with acids. However, both heterogeneous and homogeneous catalysts have been used for this reaction. While a cation-exchange resin can serve as heterogeneous catalyst, mineral acids such as sulphuric and hydrochloric acids are known as homogeneous catalysts [4]. Based on the heterogeneous and homogeneous approaches, the esterification reactions in the presence of ion-exchange resins can be explained using several kinetic models including Langmuir Hinshelwood (LH), Eley-Rideal (ER) and pseudo-homogeneous (PH) models [5]. Although the PH model does not consider the sorption effect on the resin by various components in the reactant mixture, the ER and LH models consider the sorption effects in their reactions kinetics [5]. The ER and LH model takes place when a reaction occurs between an adsorbed and the non-adsorbed reactant species. Although cation-exchange resin catalysts possesses some outstanding properties such as excellent porous medium and surface acid site, this catalysts does not only work as a catalyst in esterification reaction, but also affect the equilibrium conversion due to their swelling nature [6]. Resin catalysts have been successfully applied in the liquid-phase acid catalysed reactions especially in the reactions involving water and lower alcohols [7]. The most popular cation-exchange resins for ester production include amberlyst 35, dowex 50Wx, amberlyst 15 and amberlyst 36 [8].

Although the traditional method of solving equilibrium problems in esterification reactions involve the addition of an excess amount of alcohol to the reaction system [9], using a membrane can result in higher conversion by shifting the chemical equilibrium towards the formation of the product by in situ removal of water from the reaction mixture [10]. According to Sharma et al. [11], the kinetics of the esterification reaction products can be analysed using two methods; either by titration or using GC with a choice of detector and carrier gas. Using the GC-MS, the product can be analysed by matching the retention time of the reaction product to the retention time of the commercial ethyl lactate. Sharma et al. [11] applied both methods in the analysis of the esterification product (propyl nonanoate) using flame

ionisation detector (FID) as the detector and nitrogen as the carrier gas. Zhang et al. [3] studied the kinetics of esterification of lactic acid with ethanol catalysed by different cation-exchange resin. In their study, the order of the resin catalytic activity was D002 < D001 < Amberlyst-15 < NKC < 002. Sanz et al. [12] studied the esterification of lactic acid with ethanol and methanol catalysed by cation-exchange resin. They found out that amberlyst 15 is the most suitable catalyst for the esterification process. In the present study, the esterification reaction of lactic acid and ethanol catalysed by amberlyst 15, amberlyst 16, amberlyst 36 and dowex 50W8x catalysts was carried out at 60 °C to determine the reaction kinetic using the retention time and the peak area. The adsorbed components on the surface of the catalysts were described and validated using Langmuir Hinshelwood model.

35.2 Experimental

35.2.1 Materials

An aqueous lactic acid 99.9 wt% and ethanol 99.9 wt% solutions were purchased from sigma-Aldrich, UK and were used as received without further purification. The catalysts used in the experiments were commercial solid cation-exchange resins also purchased from Sigma-Aldrich, UK. Helium gas (99.98 % purity) was used as carrier gas and was obtained from BOC, UK.

35.2.2 Catalyst Cleaning

Prior to the esterification process, the fresh commercial cation-exchange resin was weighed into a 50 mL beaker and was rinsed with 2 mL of deionized water and 10 mL of ethanol. The catalysts were reweighed and oven dried (carbolite oven) at 65 °C for 24 h to remove any impurities and moisture completely before the esterification process.

35.2.3 Batch Process Esterification

After the catalyst cleaning process, 30 mL of lactic acid with 5 g of the resins was charged into the reactor and heated to 60 °C. After attaining the desired temperature, 50 mL of ethanol which had been previously heated separately was added to the mixture. The stirring and heating of the reaction mixture was achieved using a magnetic hot plate with a stirrer. The stirrer speed was adjusted to a frequency of about 400–800 rpm. The water from the reaction product was removed by connecting two vacuum pumps to the openings of the reactor. The inlet water flows

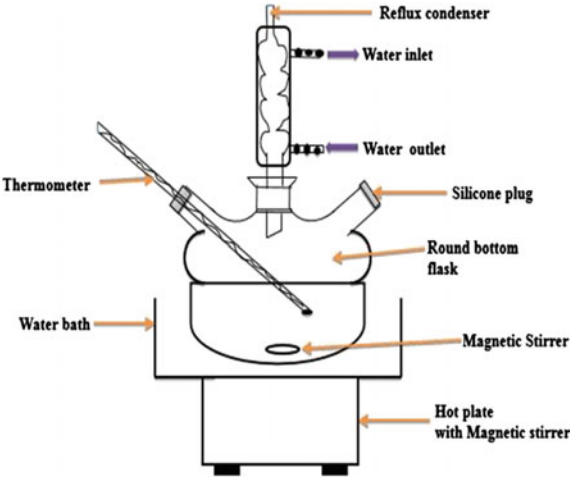


Fig. 35.1 Batch process experimental setup

through the pipe to flush the system while the out let water displace the waste water from the reaction system. About 1 mL of the reaction product was withdrawn from the reactor simultaneously for the qualitative analysis with the GC-MS. Figure 35.1 shows the batch process esterification setup.

35.2.4 GC-MS

Agilent technologies 7890B autosampler Gas chromatography (GC) system coupled with Agilent technologies 5977A mass spectrometry detector (MSD) was used to determine the kinetics of the esterification reaction product. A microliter syringe (Hamilton HM80300 microliter TM syringe) was used for sample preparation before it was inserted into the GC vial for injection. A capillary column with the column dimensions of 30 m × 250 μm × 0.25 μm was used for the analysis and at the rate of 10 °C/min, at 63.063 kPa. The Helium gas temperature was set at 40 °C with the flow rate of 1.2 mL/min and equilibration time of 0.25 min while the inlet pressure was 100 PSI. The oven temperature was program at 40 °C with the holding time of 2 min at maximum operating temperature of 25 °C. The sample scanning rate was 9.773 min. The solvent was analyzed using the split ratio of 50.0. The sample vial was carefully cleaned prior to each analysis. The sample was carefully inserted into the sample rage embedded in the injector port of the GC-MS where the carrier gas transfers the solvent to the GC column. The NIST GC software program was used for data collection. Figure 35.2 shows the GC-MS instrument that was used for the esterification analysis.

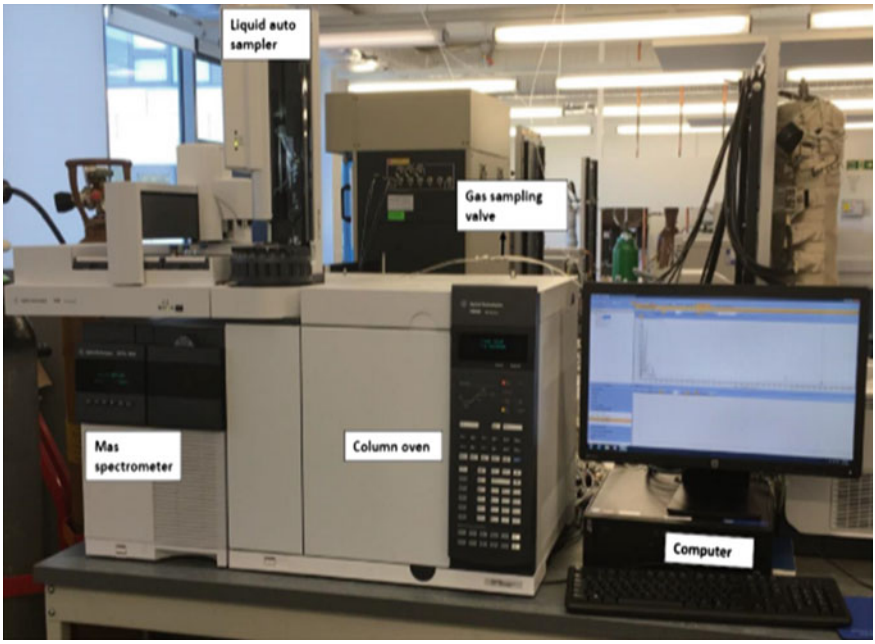


Fig. 35.2 Agilent technologies 7890B autosampler Gas chromatography (GC) system coupled with Agilent technologies 5977A mass spectrometry detector (MSD) at the Centre for Process Integration and Membrane Technology (CPIMT), RGU

35.2.5 Catalyst Characterization

Fourier transform infrared coupled with attenuated total reflection (Nicolet iS10 FTIR-ATR) was used for the structural analysis of the resins. This method was used in order to determine the phenomenon of the component with the strongest absorption strength on the surface of the resin catalysts. Prior to the analysis of the sample, a background spectrum of air (without any solvent) was obtained. The scanning pattern was observed within the range of 4000–400 cm^{-1} with the resolution of 4 cm^{-1}. Figure 35.3 shows the pictorial view of the FTIR-ATR. The resin catalysts morphological characteristics were also determine using SEM/EDAX in other to examine the thermal stability and mechanism strength of resin catalyst before and after the esterification process.

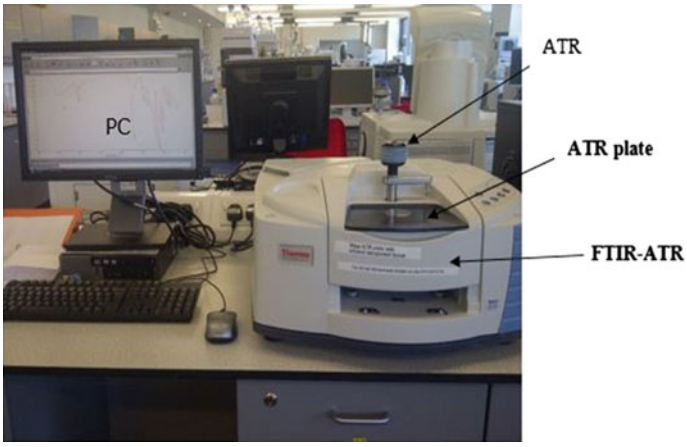


Fig. 35.3 Pictorial view of the Thermo Scientific Fourier Transform infrared coupled with attenuated total reflection (Nicolet iS10), School of pharmacy Life Science, RGU

35.3 Results and Discussion

35.3.1 GC-MS Results

The chromatogram/mass spectra results for the commercial ethyl lactate and the esterification product were compared in order to investigate the reaction kinetics using the retention time and their peak area. Figures 35.4 and 35.5 presents the chromatogram results of the commercial ethyl lactate and the esterification product in the catalysed with amberlyst 36 resin catalyst at 60 °C. From the chromatogram results, the mass spectra of the reaction products were generated from the highest peak in the spectra using NIST software program of the GC-MS system. From the mass spectra of the reaction product, ion 45 exhibited the structure of ethyl lactate solvent as shown in Fig. 35.6. This was in accordance with the library spectra of the commercial ethyl lactate solvent which also confirm that the produced esterification reaction product can be used for industrial purposes. Although the esterification product results were in accordance with the commercial ethyl lactate, From Table 35.1, it was observed that the esterification product catalysed by amberlyst 36 showed the appearance of peak at the retention time of 1.444 min and the peak area of 107,063,789 m^2 whereas the commercial ethyl lactate eluted at the retention time of 2.127 min and the peak area of 13,152,320 m^2. The faster elution time was attributed to the presence of catalysts in the solvent thereby fastening the rate of the reaction. It was found that the solvent elution time increases with respect to the peak area. A similar result was also obtained for amberlyst 15, 16 and dowex 50W8x at 60 °C.

Figure 35.7 shows a plot of the solvent concentration (mol/dm^3) against agitation time (s) at the esterification temperature of 60 °C. From Fig. 35.7, it was found that the solvent concentration increases with increase in agitation time (s).

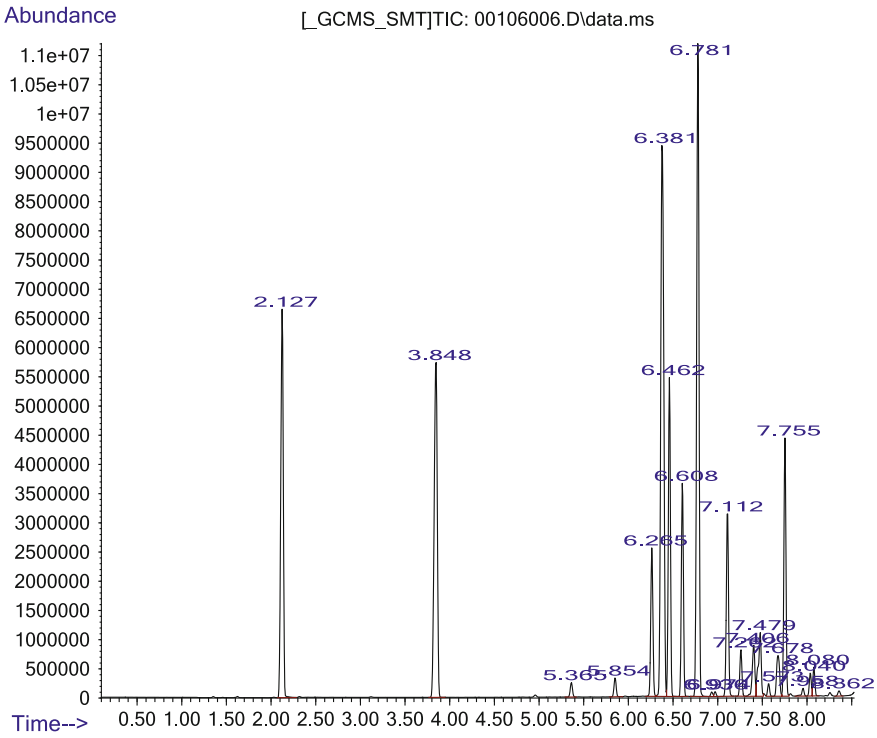


Fig. 35.4 GC chromatogram for commercial ethyl lactate solvent

According to Nie et al. [13], the effect of external mass transfer diffusion limitation between the liquid components and the resin catalyst can be avoided by increasing the agitation time of the esterification reaction. The stirring time was set between 400 and 800 rotation per minute (rpm). During the esterification process, it was observed that when the stirring process was increase above 800 (rpm), there were no noticeable changes in the rate of the reaction. Thus the esterification reaction was further conducted within agitation speed range of 400–800 rpm with 800 rpm at the highest agitation time to ensure that there is no existing of mass transfer between the catalysts and the solvent liquid.

Figure 35.8 present the calibration curve of the esterification product catalysed by amberlyst 36 at 60 °C. From Fig. 35.8, it was found that the ester product exhibited R^2 value of up to 1.00000 which was suggested to be a good correlation value of the graph. It was also found that the solvent showed a positive slope of 4,639,249 with an intercept of 9.54E11.

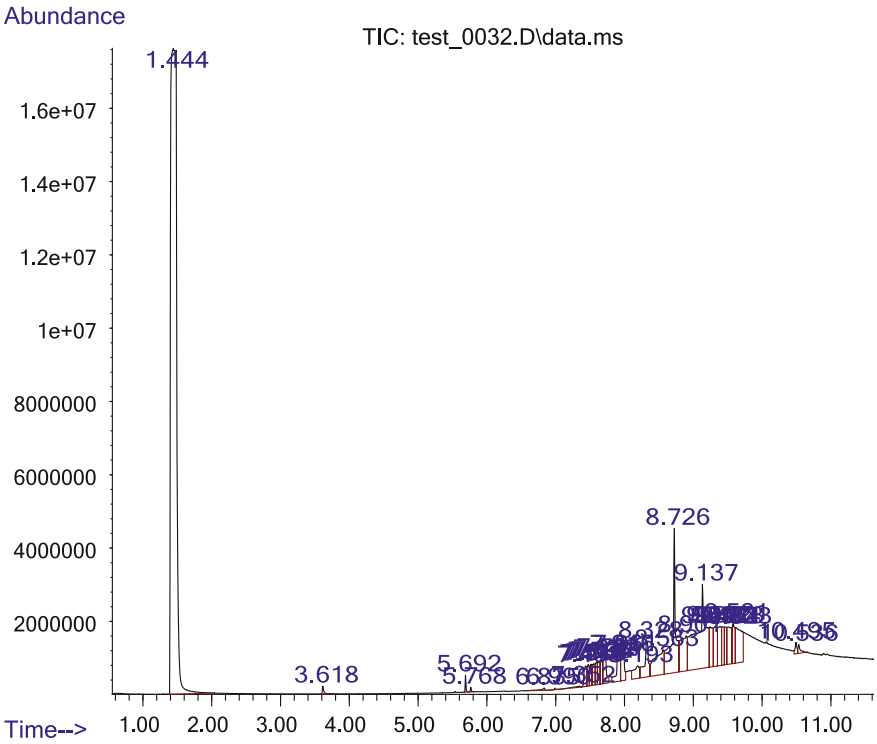


Fig. 35.5 GC-chromatogram of reaction product catalysed by amberlyst 36 at 60 °C

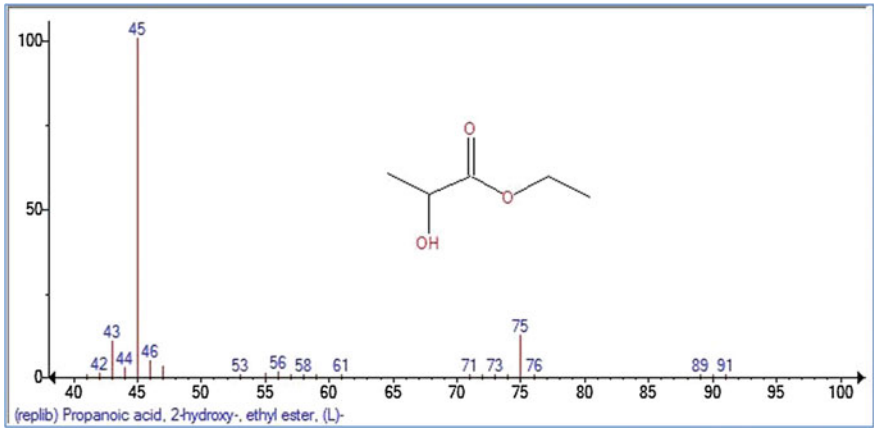


Fig. 35.6 MS Library spectra for ethyl lactate solvent

Table 35.1 Comparison of retention time and peak area for the commercial ethyl lactate and reaction product for amberlyst 36 at 60 °C

Retention time ethyl lactate/reaction product		Peak area of ethyl lactate/reaction product	
Commercial ethyl lactate	Reaction product with amberlyst 36	Commercial ethyl lactate	Reaction product catalysed by amberlyst 36
2.127	1.444	13,152,320	107,063,789
3.846	3.618	14,006,785	129,264
5.365	5.692	446352.4	5,600,390
5.854	8.726	57427.69	21,436,293
6.265	9.137	4,295,677	100,636,546
6.381	10.436	22,542,187	14,917,454

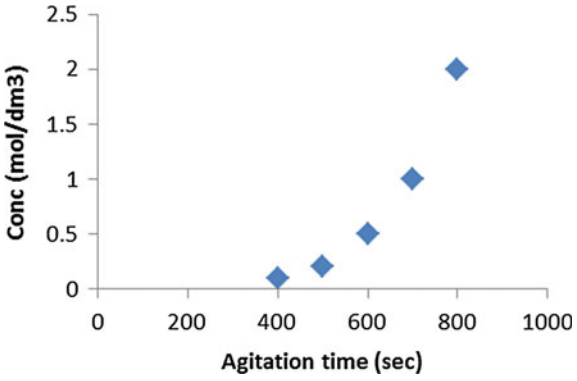


Fig. 35.7 Concentration (mol/dm^3) against agitation speed (s)

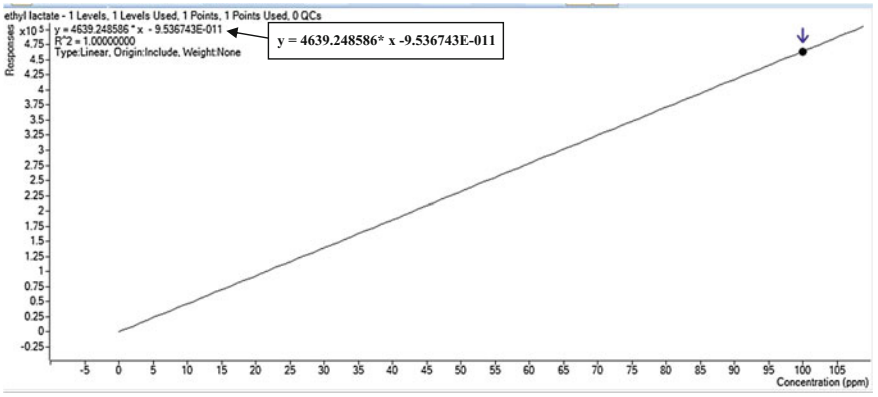


Fig. 35.8 Calibration curve for the esterification product at 60 °C

35.3.2 SEM Surface Morphology of Resin Catalysts

Figure 35.9a–d, presents the SEM images of the fresh commercial resin catalysts before and after esterification at 60 °C. It can be seen that the surfaces of dowex 50W8x (Fig. 35.9a) and amberlyst 36 (Fig. 35.9b) before the esterification reaction showed a very smooth surface indicating that the resin catalysts were defect-free. A similar result was obtained by Zhang et al. [3]. From Fig. 35.9c (dowex 50W8x after esterification) it was observed that the catalysts exhibited some tiny bubble-like structure which was attributed to the effect of temperature on the resin catalyst after the esterification process [14]. However, from Fig. 35.9d (amberlyst 36 after esterification) it was observed that amberlyst 36 showed no sign of crack on the surface indicating that this catalyst has a greater thermal stability at 60 °C in contrast to the dowex 50W8x catalyst at the same temperature.

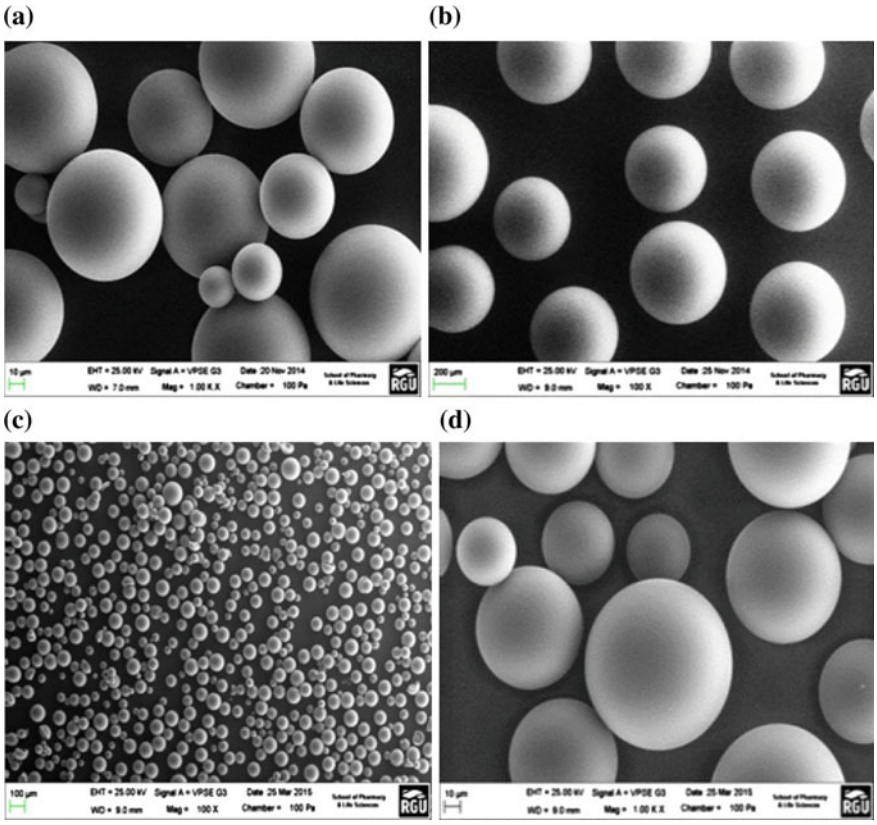


Fig. 35.9 **a** (dowex 50W8x before), **b** (amberlyst 36 before), **c** (dowex 50W8x after) and **d** (amberlyst 36 after): SEM surface morphology of Amberlyst 36 and Dowex 50W8x resin catalysts before and after esterification reaction at 60 °C

35.3.3 FTIR of Esterification Products

Figure 35.10 presents the FTIR spectra for amberlyst 16 cation-exchange resin catalysts respectively. From Fig. 35.10, it was observed that the FTIR analysis of the resin catalysts showed the band at 1723 cm^{-1} corresponding to C=O stretching strong adsorption bond, and also the bands at 2986 cm^{-1} representing O–H functional groups which correspond to stretching vibration bond. However, the band at 1212 cm^{-1} was suggested to represent =C–O–C– functional group. This (=C–O–C–) structure was suggested to originate from the carboxylic acid functional group while the O–H are thought to arise from the alcohol functional group. It was suggested that the strong adsorption bond of C=O, O–H and =C–O–C– could indicate the fact that ethanol, water and lactic acid were the adsorbed components on the surface of the resin catalysts [15]. The rationale of the result and the adsorption parameters were further validated using Langmuir Hinshelwood model.

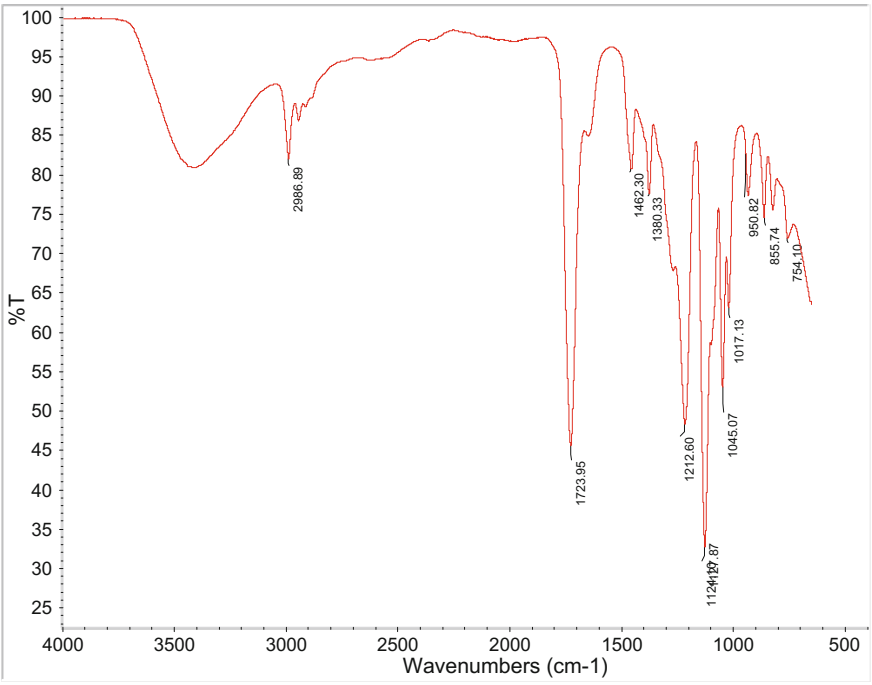


Fig. 35.10 FTIR spectra of esterification reaction product catalysed by amberlyst 16 at 80 °C and the wavelength region of 500–4000 cm^{-1}

35.3.4 Model Validation and Parameters Estimation

According to Khajavi et al. [9], heterogeneous reaction can be presented with many models including LM, ER and pseudo-homogenous model. Among them, LH model seems to be more appropriate for this type of reaction. The FTIR-ATR results were further tested using simplified mechanism of the Langmuir Hinshelwood model as thus:

$$\mathrm{LA} + \mathrm{S} \leftrightarrow \mathrm{LA} - \mathrm{S} \tag{35.1}$$

$$\mathrm{E} + \mathrm{S} \leftrightarrow \mathrm{E} - \mathrm{S} \tag{35.2}$$

$$\mathrm{E} - \mathrm{S} + \mathrm{LA} - \mathrm{S} \leftrightarrow \mathrm{EL} - \mathrm{S} + \mathrm{W} - \mathrm{S} \tag{35.3}$$

$$\mathrm{EL} - \mathrm{S} \leftrightarrow \mathrm{EL} + \mathrm{S} \tag{35.4}$$

$$\mathrm{W} - \mathrm{S} \leftrightarrow \mathrm{W} + \mathrm{S} \tag{35.5}$$

where LA = lactic acid, E = ethanol, EL = Ethyl lactate, W = Water and S = Vacant site on catalyst surface [3]. Therefore, the LH model which describes the reaction rate as initial molar amount of lactic acid and water in the esterification process could be written as follows:

$$r = \frac{n_{LA,0}}{w}\left(\frac{dx}{dt}\right) \tag{35.6}$$

From Eq. 35.1, the reaction rate constant (k) involving the esterification parameters (lactic acid, ethanol, ethyl lactate and water) can be written as shown in Eq. 35.7:

$$\mathrm{k} = \frac{a_{LA}a_E - a_{EL}a_W}{\left(1 + k_{WaW} + k_{EaE} + k_{LA}a_{LA} + k_{EL}a_{EL}\right)^2} \tag{35.7}$$

The adsorption coefficient (i) as well as the equilibrium constant (K_{eq}) of the esterification parameters in the esterification reaction can be written using the Eq. 35.8:

$$k_i = \frac{c_i - z}{a_i c_s}, \quad K_{eq} = \left(\frac{a_{EL} \times a_W}{a_{LA} \times a_E}\right) eq. \tag{35.8}$$

In the Eqs. 35.6–35.8, $n_{LA,0}$ = initial molar concentration of lactic acid and w = the catalyst loading, k = represents the reaction rate constant, k_i = adsorption coefficient, C_s = the concentration of vacant site on catalyst surface, $C_i - s$ = the concentration of component i at the catalyst surface, a_i = the activity for component and K_{eq} = the reaction equilibrium constant. Although ethanol and water were identified as the most adsorbed components on the surface of the resin catalysts,

lactic acid was also suspected to also adsorb on the surface of the resins. Two mechanisms (A and B) of the Langmuir model were tested in order to determine which one of these was more suitable to describe the adsorption components. Mechanism "A" assumes that ethanol and water adsorbed much stronger than other components in the esterification solution and as such, the adsorption of lactic acid and ethyl lactate were ignored. The kinetic equation for the two mechanisms (A and B) were written by combining Eqs. 35.6 and 35.7.

In mechanism A, it was assumed that water and ethanol adsorbed on the surface of the resin catalysts as in the denominator of Eq. 35.9.

$$r = k\frac{a_{LA}a_E - \frac{u_{EL}u_W}{keq}}{\left(1 + k_{WaW} + k_{EaE}\right)^2}. \tag{35.9}$$

In comparison to the first mechanism (A), in the second mechanism (B) it was assumed that water and lactic acid adsorbed most as shown in the denominator of Eq. 35.10.

$$r = k\frac{a_{LA}a_E - \frac{u_{EL}u_W}{keq}}{\left(1 + k_{WaW} + k_{LA}a_{LA}\right)^2} \tag{35.10}$$

In both A and B mechanisms, three parameters were considered for evaluation at a constant reaction temperature. In the 1st mechanism, the parameters include: k, k_W and k_E whereas for the 2nd mechanism, the corresponding parameters are k, k_W and k_A. Overall, these results further confirmed Langmuir model as the fitted model for the description of the adsorption components on the surface of the cation-exchange resins. This model was based on a similar work by Zhang et al. [3].

35.4 Conclusion

The validation method for ethyl lactate production using Langmuir Hinshelwood model was successfully achieved. The SEM surface of the cation-exchange before esterification showed a defect-free surface whereas the surface of the catalyst after esterification showed a bit of crack indicating the effect of temperature on the surface of the resin catalyst. The mass spectra result generated from the ion chromatogram of the reaction product confirm the presence of ethyl lactate solvent which matches the library spectra of that of the commercial ethyl lactate used as the reference sample. The SEM results of the amberlyst 36 confirm that the catalyst has more resistance to temperature in contrast to Dowex 50W8x. This also suggests that the produced reaction product with amberlyst 36 can be used for industrial purposes. The order of the catalytic activity with respect to the SEM analysis was amberlyst 36 > Dowex 50W8x. The concentration of the reaction product was found to increase with increase in the agitation time. The spectra for the FTIR

showed different functional groups including C=O, O–H and C–H. The Langmuir model was found to be the best fitted model for the description of the adsorption component on the resin surface.

Acknowledgements The author wishes to express sincere thanks to The Centre for Process Integration and Membrane Technology (CPIMT) for supplying the infrastructures used for the study. School of Pharmacy and Life Science at RGU for the SEM examination and FTIR analysis. Additionally, CCEMC is gratefully acknowledged for their financial support towards the research work.

References

1. Engin A, Haluk H, Gurkan K (2003) Production of lactic acid esters catalyzed by heteropoly acid supported over ion-exchange resins. Green Chem 5(4):460–466
2. Troupe RA, DiMilla E (1957) Kinetics of the ethyl alcohol—lactic acid reaction. Ind Eng Chem 49(5):847–855
3. Zhang Y, Ma L, Yang J (2004) Kinetics of esterification of lactic acid with ethanol catalyzed by cation-exchange resins. React Funct Polym 61(1):101–114
4. Altıokka MR, Çıtak A (2003) Kinetics study of esterification of acetic acid with isobutanol in the presence of amberlite catalyst. Appl Catal A 239(1):141–148
5. Lomba L, Giner B, Zuriaga E, Gascón I, Lafuente C (2014) Thermophysical properties of lactates. Thermochim Acta 575:305–312
6. Pereira CS, Pinho SP, Silva VM, Rodrigues AE (2008) Thermodynamic equilibrium and reaction kinetics for the esterification of lactic acid with ethanol catalyzed by acid ion-exchange resin. Ind Eng Chem Res 47(5):1453–1463
7. Mitsutani A (2002) Future possibilities of recently commercialized acid/base-catalyzed chemical processes. Catal Today 73(1):57–63
8. Toor AP, Sharma M, Thakur S, Wanchoo RK (2011) Ion-exchange resin catalyzed esterification of lactic acid with isopropanol: a kinetic study. Bull Chem React Eng Catal 6(1)
9. Khajavi S, Jansen JC, Kapteijn F (2010) Application of a sodalite membrane reactor in esterification—coupling reaction and separation. Catal Today 156(3):132–139
10. Ameri E, Moheb A, Roodpeyma S (2010) Vapor-permeation-aided esterification of isopropanol/propionic acid using NaA and PERVAP® 2201 membranes. Chem Eng J 162(1):355–363
11. Sharma M, Wanchoo R, Toor AP (2014) Amberlyst 15 catalyzed esterification of nonanoic acid with 1-propanol: kinetics, modeling, and comparison of its reaction kinetics with lower alcohols. Ind Eng Chem Res 53(6):2167–2174
12. Sanz MT, Murga R, Beltrán S, Cabezas JL, Coca J (2004) Kinetic study for the reactive system of lactic acid esterification with methanol: methyl lactate hydrolysis reaction. Ind Eng Chem Res 43(9):2049–2053
13. Nie H, Xu R, Zhang F, Zhou Z, Zhang Z, Yang G (2015) Thermodynamic and kinetic studies on alkoxylation of camphene over cation exchange resin catalysts. AIChE J 1–8
14. Okon E, Habiba S, Gobina E (2015) Batch process esterification of lactic acid catalysed by cation-exchange resins for the production of environmental-friendly solvent. In: Proceedings of the world congress on engineering and computer science 2015, WCECS 2015. Lecture notes in engineering and computer science, 21–23 Oct 2015, San Francisco, USA, pp 623–628
15. Sert E, Buluklu AD, Karakuş S, Atalay FS (2013) Kinetic study of catalytic esterification of acrylic acid with butanol catalyzed by different ion exchange resins. Chem Eng Process 73:23–28

Chapter 36
Potential Impacts of Leachate From Waste Mixed With CCA-Treated Wood Destined to Landfills

Emmanuel Emem-Obong Agbenyeku, Edison Muzenda and Innocent Mandla Msibi

36.1 Introduction

Chromated Copper Arsenate (CCA)-treated wood have been in use for the past few decades as structural elements with short to long term service life. The CCA-treated wood became widespread in about 1970 to 1980 with the life spans of the structures made with it ranging from 10 to 40 years. This implies that disposal quantities are insistently on the rise as the wood continues to come out of service. Notwithstanding, in the United States alone as recorded by [1] an estimated 180 million m^3 of CCA-treated wood is still used. Wood treated with CCA is often used in outdoor structures such as gazebos, decks, garden-bed borders, play-structures, playground equipment, picnic tables, docks, residential fencing, patios and walkways.

CCA is a water-soluble inorganic pesticide normally used as a wood preservative to make it resistant to attack by termites and fungi that cause decay. The wood is soaked in a solution of CCA and subjected to vacuum pressure forcing penetration of the chemical into the wood. As such, CCA-treated wood is also called pressure-treated wood [2]. Considering the unrestricted use of CCA-treated wood, the disposal rate is expected to continue to increase and eventually level out.

E.E.-O. Agbenyeku (✉) · E. Muzenda
Chemical Engineering Department, University of Johannesburg, Johannesburg, South Africa
e-mail: kobitha2003@yahoo.com; emmaa@uj.ac.za

E. Muzenda
e-mail: muzendae@biust.ac.bw; emmaa@uj.ac.za

E. Muzenda
College of Engineering and Technology, Botswana International University of Science and Technology, Palapye, Botswana

I.M. Msibi
Innovation and Impact, Water Research Commission, Pretoria, and Research and Innovation Division, University of Johannesburg, Johannesburg, South Africa
e-mail: mandlam@wrc.org.za

S.-I. Ao et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-981-10-2717-8_36

However, phase-outs or embargos on CCA-treated wood have begun to occur. Moreover, with the amount of awareness been cast on the direct effects of CCA-treated wood on human and environmental health, the clear impact on users recorded by [3] can be appreciated following that; freshly CCA-treated wood may retain pesticide residues on the wood surface after the treatment process. Then, since CCA is water-soluble, rainwater can seep in and leach CCA onto wood surfaces. Furthermore, as the wood ages it cracks which worsens the leaching process. Hence, CCA residue can stick to hands or clothing from treated wood surfaces which could pose health risks when ingested. Plus in and around the immediate environment, CCA when in contact with rainwater or by virtue of weathering effects leach from CCA-treated wood into the soil beneath and adjacent to CCA-treated wood structures. The soil around such areas has been found to be contaminated by arsenic, chromium and copper. However, waterproofing sealants applied to structural components have been recorded to lower the metal concentrations in soil around waterproofed CCA-treated wood [1].

The impacts of CCA together with the vast quantities of generated construction and demolition (C&D) waste, increased disposal amounts of CCA-treated wood are expected to occur which may peak disposal rates sooner than anticipated thereby turning increased interest towards proper disposal and containment. One crucial aspect that could fast-track the disposal process of CCA-treated wood is the concern over its effect on the health of children. Young children are more at risk of exposure to CCA because they tend to spend more time playing outdoors and since they have frequent hand-to-mouth contacts, it becomes a critical issue deserving dire attention. Children playing on playground equipment or decks constructed with CCA-treated wood can be exposed to CCA by touching the CCA leachate on the wood surface with their hands thereby, unconsciously ingesting the chemical by hand-to-mouth activity. It is however noted by [3] that the amount of CCA leached on the surface of the wood depends upon the type of wood and the age of the structure. The amount ingested is also reliant on the frequency of hand-to-mouth contact. Children may also be exposed to CCA in contaminated soil by accidentally ingesting the chemical when playing around areas with these structures through similar hand-to-mouth activities. Currently, it is expected that CCA-treated wood in the United States would have phased-out as scheduled to take place in early 2004 [4]. A similar goal may not be actualized any time soon in South Africa and other developing African countries as pressing needs such as poverty alleviation, health sector improvement, education, housing and circular sustainable economic development are also desperately in need of urgent intervention. Nevertheless, one propelling force leading to bans and discontinued use of CCA-treated wood could be the report on the increased cancer risks to children in contact with CCA-treated wood decks and play sets as recorded by [5, 6].

The recycling and reuse options in South Africa and most developing African countries are currently lacking or infeasible. This implies that most CCA-treated wood will be destined to landfills or openly dumped. CCA-treated wood needs to be properly managed as a solid waste once removed from service. Under standardized tests, CCA-treated wood may show characteristics of a hazardous waste because of

the leachability of the toxic metals in the CCA chemical. However, when CCA-treated wood is used for its intended purpose and discarded by the user as noted by [7] it is exempted from being a hazardous waste. Considering the high metal concentrations in CCA-treated wood, it should not be mulched or burnt as it emits arsenic to the atmosphere causing air pollution and concentrates arsenic, copper and chromium in the ash [8]. Landfills, monofills and open dumps are often where waste wood are disposed. It is therefore generally known in South Africa that landfills will remain the major system of waste management for various waste types [9], including CCA-treated wood in the foreseeable future or at least till economical recycling options are initiated. Studies by [10] have shown that C&D waste containing CCA-treated wood leached concentrations of arsenic above groundwater standards. As such, in contribution to ongoing interests in waste containment and leachate control, this work through laboratory investigations, explores the leachate quality and subsequent potential impact of waste mixed with CCA-treated wood on the health of humans and the environment. This work simulates using a laboratory bespoke percolation column device an account of a breakaway from containment systems, particularly, in unlined facilities fully reliant on the natural geology of the site.

36.2 Practical Exploration

The main objective of the study presented herein was to simulate the generation of leachate from the co-disposal of CCA-treated wood in a laboratory bespoke percolation column device as simulated landfills. The small scale bespoke device of about 600 mm long with internal diameter of about 160 mm was used in series of tests for the study. A pictorial view of the laboratory bespoke device is shown in Fig. 36.1 and for the purpose of this study, the device was coupled to comprise of two sections:

1. The bottom chamber called the bucket section; which contained the waste to simulate different CCA-treated wood disposal cases: a C&D debris landfill, a MSW landfill and a wood monofill. Series of leaching tests were conducted for the study using the bespoke device. The testing series were split into pairs, each

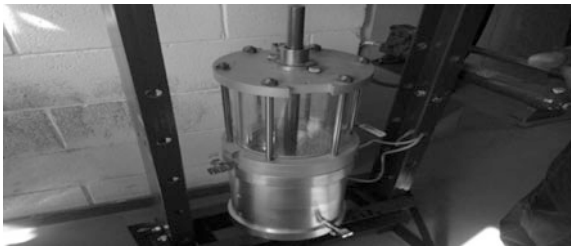
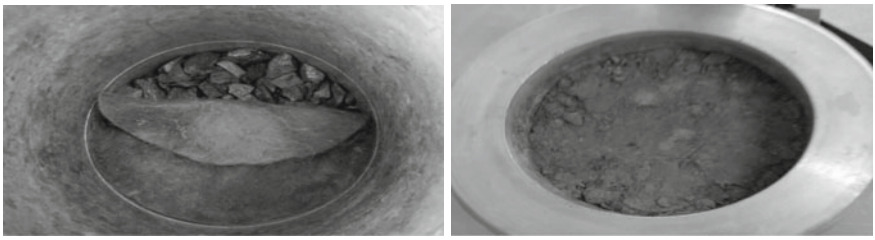


Fig. 36.1 Bespoke percolation column test device

pair representing a different disposal situation, having an experimental and control series in each pair summing up to six testing sets in total. The bottom chamber was setup using selected gravel laid as drainage path in the bucket section and a wet geotextile which served as a filter bed in the system was placed over it to prevent clogging of the outlet by moving fines. The simulated waste was chopped/crushed into relatively proportionate sizes and placed in the bottom chamber to a height of 230 mm overlain by a 20 mm thin layer of lightly compacted clayey soil.

The size reduction of larger waste materials into smaller particle sizes was done in order to fit into the small scale bespoke device whereas the thin soil layer was introduced to prevent a rapid flush out of the influent during the percolation process and ensure lasting waste-water interaction towards the production of leachate. Figure 36.2a, b shows the experimental arrangement for the bucket which held the resized waste.

The particle size distribution curves for the clayey soil, selected gravel used as filter bed, the resized C&D debris, MSW and wood waste are shown in Fig. 36.3. Whereas Fig. 36.4a–c shows the pictorial view of the resized C&D debris, MSW and wood waste respectively. The respective waste materials were appropriately weighed for their masses, mixed together, placed and simultaneously compacted in layers within the bottom chamber.



(a) Filter bed (b) Soil overlying waste material

Fig. 36.2 **a** and **b** Wet geotextile placed over selected gravel and lightly compacted clayey soil overlying resized waste material placed and compacted in layers within the bucket

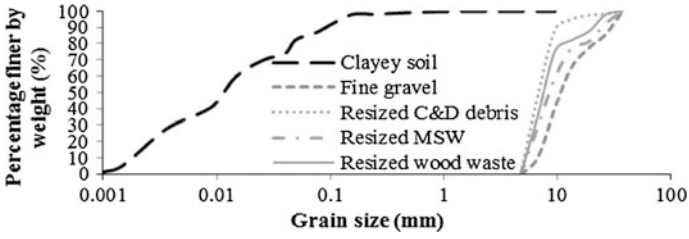


Fig. 36.3 Particle size distribution for the respective materials used for the study

(a) Resized C&D debris (b) Resized MSW (c) Resized Wood waste

Fig. 36.4 **a–c** Pictorial view of the resized waste materials used in the study

The wood monofill represents a worst-case disposal scenario for CCA-treated wood since it contains the highest percentage of treated wood. This case factor will also provide information on the leaching of metals that will occur from wood without the interference of other materials. Since the relative contribution of CCA-treated wood from C&D activities varies and has not been sufficiently quantified, a 100 % waste ratio by weight was used.

2. The upper chamber above the bucket section called the reservoir; contained de-ionized water as influent which flowed through a perforated plate simulating low intensity rain droplets pouring into the bottom chamber which held the waste body and slowly saturating the waste body till leachate was formed. The concentration of the collected effluent was tested at intervals to monitor the concentration levels of the targeted chemical ions. The system was coupled firmly using O-rings, gasket corks and silicon sealants to ensure a leakage free assemblage. The reservoir held a water head of 250 mm which was manually topped as the water level dropped since a mechanism for retaining constant head was not designed for in the device. Figure 36.5a, b show the upper chamber of the assembled device used in the study. As earlier stated, the choice of CCA-treated wood monofill offers the worst-case disposal scenario in unlined sites considering that the disposal of chemical treated wood will in the foreseeable future insistently come out of service and be destined to dump sites. This case will also aid a clearer understanding on the chemical ions generated thereof without intrusion from contaminant species of other sources.

(a) Perforated plate (b) Assembled device with de-ionized water in reservior

Fig. 36.5 **a** and **b** Perforated plate through which de-ionized water as influent in the reservoir section flowed simulating low intensity rain droplets

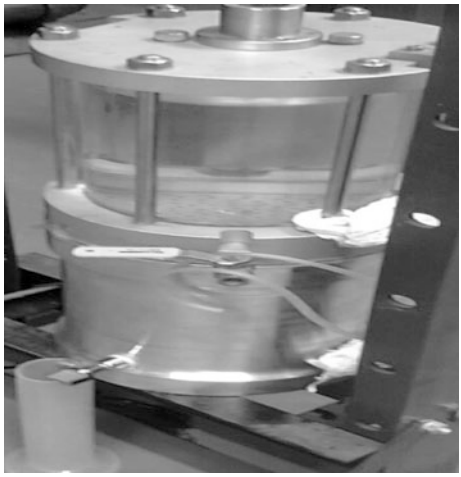
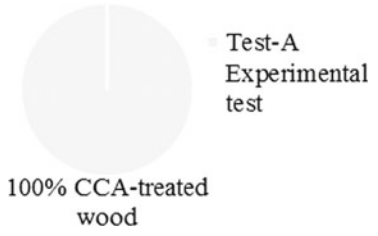


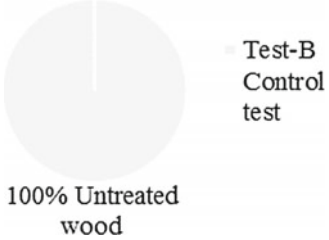
Fig. 36.6 Percolation test under study using de-ionized water as permeant for the waste material with the effluent collected and tested at intervals to track the targeted parameters

Outside other confirmatory percolation tests conducted in the study, six main tests were executed and the data recorded. Test-A represented the experimental case for CCA-treated wood in a monofill while Test-B represented the control case with untreated wood. Test-C represented the experimental case for MSW+CCA-treated wood while Test-D represented the control case for MSW+ untreated wood. Test-E represented the experimental case for C&D waste+CCA-treated wood whereas Test-F represented the control case for C&D waste+untreated wood. The fully assembled percolation column test setup as initiated in the study is shown in Fig. 36.6 while Fig. 36.7a–f shows the composition by percentage mass of the waste materials for the respective test.

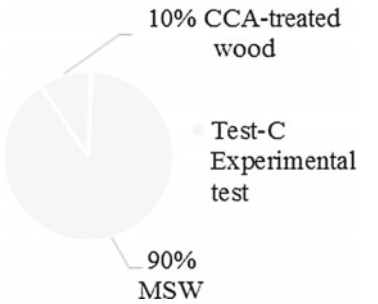
Running and monitoring the bespoke percolation column device for each testing period lasted up to 15 days. The respective tests were conducted at temperatures varying between 5–25 °C. The varying temperatures simulated seasonal changes and ambient fluctuations towards assessing the impact of changing temperatures on gas production. The gas composition in the bespoke device was analyzed using a Portable Multi-flex Beckman GEM-500 that measured methane, carbon dioxide and oxygen content. The gas composition readings were collected daily for Tests-C and D containing 90 % MSW and Tests-E and F containing 70 % C&D waste. During higher temperatures up to 25 °C, the methane concentrations in the device containing MSW rose to roughly 70 % with carbon dioxide at roughly 30 %. The primary gas in the device containing C&D waste was carbon dioxide. Over lower temperatures up to 5 °C, both methane and carbon dioxide decreased whereas oxygen increased in the devices containing MSW and C&D waste. The gas fluctuations is plausibly an indication that microorganisms that produce methane and carbon dioxide are more active in warmer conditions than in cooler conditions. On formation and breakthrough of leachate, a 50 ml cylinder was used to constantly collect the effluent. General water quality parameters i.e., pH, dissolved oxygen (DO), conductivity, temperature and oxidation-reduction potential (ORP) were measured in the study



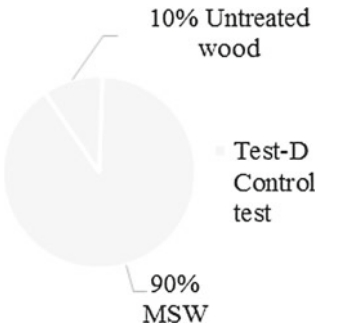
(a) 100% CCA-treated wood



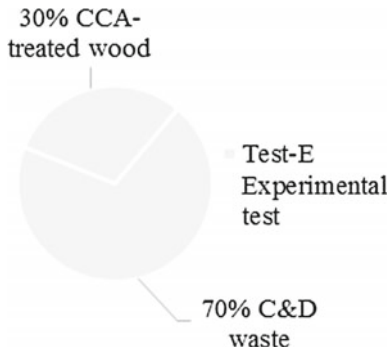
(b) 100% Untreated wood



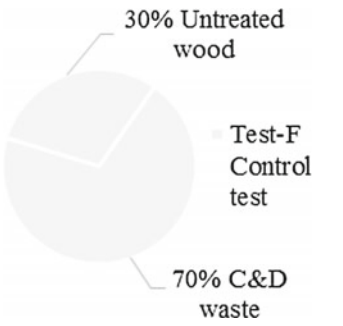
(c) 90%MSW+10%CCA-treated wood



(d) 90%MSW+10%Untreated wood



(e) 70%C&D+30%CCA-treated wood



(f) 70%C&D+30%Untreated wood

Fig. 36.7 a–f Composition by percentage mass of the waste materials used in the study

using a Portable Multi-flex Beckman Century pH/mV/FT/DO-915A-1202-TM SS-1 m Model. The outcome of these measurements are specifically discussed for the tests involving 100 % CCA-treated and untreated wood. Collected leachate were put in plastic containers and stored in a cooling chamber at 2 °C to prevent any form of chemical or biological activities. The effluent samples were then analyzed by full spectral method and were compared to South African standard of drinking water in accordance with [11].

36.3 Results and Discussion

Test-A and Test-B indicate cases of leachate generation from the deposition of CCA-treated and untreated wood in a monofill respectively. This was done with a view to ensuring that clear understanding could be made of the targeted ions without the intrusion of other ions from different waste bodies. Each test setup ran for 15 days using de-ionized water as permeant for proper interaction with the waste body as well as for sufficient leachate formation. Breakthrough for Test-A and Test-B occurred at very early stages. The collection of leachate and measurement of general water quality parameters and targeted ions (i.e., arsenic, copper and chromium) were done for both tests up to 15 days, after which a steady state concentration level was reached. Test-A had a pH value of 6.08 as against Test-B with a considerably higher pH value of 6.45 as shown in Fig. 36.8.

Observing and measuring the general water quality parameters provides an indication of the activities within the bespoke column device and assists in characterizing the generated leachate. From results and analysis therefore, the pH of both tests was found to minimally decrease over the test duration. The arsenic, copper and chromium ions from the formed leachate in Test-A showed higher concentration levels as against Test-B with almost over two magnitudes lower as shown in Fig. 36.9a–c.

The chromium and copper ions leached two and one order of magnitude more in Test-A as against Test-B respectively. The high concentrations of metals plausibly tie to the increasing trend recorded and correspond to the low and decreasing pH. These outcomes were found to be reasonably in line with similar studies by [4].

Considering that the study herein was a short-term laboratory investigation as against works from other authors which were conducted in fields over long periods, values represented are not indicative of real life expectations. Some level of microbial action in both Test-A and B is suspected to have occurred over the testing period as indicated by the ORP and DO outcomes. Although, as reported by [12] certain bacteria have been discovered to flourish on CCA-treated wood thereby, extracting the metals present. Also, as observed by [4] some fungi creates dicarboxylic acid and other organic acids making chromium and arsenic remain in water soluble forms. This can trigger precipitation of copper as copper-oxalate having low water solubility.

This could possibly explain why copper concentration levels were relatively lower in both tests for CCA-treated and untreated wood wastes, in contrast to,

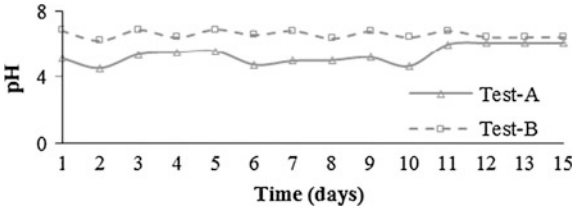


Fig. 36.8 Measured pH for tests on CCA-treated and untreated wood

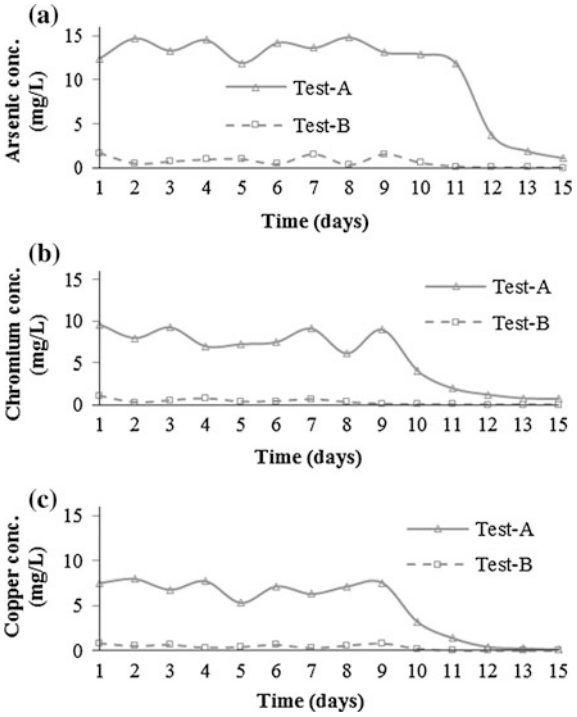


Fig. 36.9 **a** Measured arsenic conc. on CCA-treated and untreated wood. **b** Measured chromium conc. on CCA-treated and untreated wood. **c** Measured copper conc. on CCA-treated and untreated wood

arsenic and chromium respectively. DO levels began in the range of 4–6 mg/L but eventually stabilized at roughly 1 mg/L on termination of the test as shown in Fig. 36.10. Conductivity was found in the range of 300–800 µS/cm showing a decreasing trend over the duration of the test as shown in Fig. 36.11.

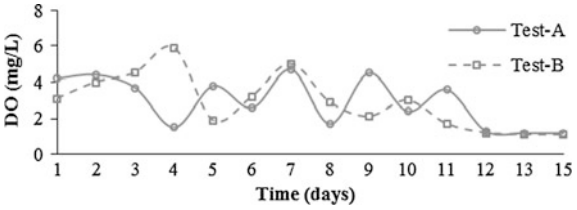


Fig. 36.10 Measured DO level for CCA-treated and untreated wood

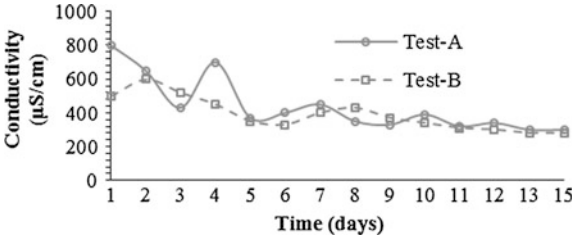


Fig. 36.11 Measured conductivity for CCA-treated and untreated wood

In the case of the ORP results, Test-A started out positive with roughly 90 mV whereas the ORP of Test-B started out negative between −180 to −370 mV. The ORP of Test-A eventually decreased to the range of Test-B with both Tests-A and B progressively decreasing (remaining negative) right to the termination of the experiment.

Generally, external changing temperatures also led to temperature fluctuations inside the column device as well as in the generated leachate.

36.4 Conclusion and Future Work

It is clear that CCA-treated wood will continually come out of service and be disposed in monofills and/or landfills as landfilling still remains the major form of waste management in South Africa and other developing countries. This study has therefore, explored and discussed the potential environmental impact of disposing CCA-treated wood in the scenario of a monofill. Other scenarios on co-disposal of CCA-treated wood waste were also investigated herein, however, discussions on those scenarios will be addressed in future works. The laboratory bespoke column device used in the study simulated a monofill scenario of CCA-treated and untreated wood waste. High concentrations of the targeted chemical ions from analysis were found in Test-A as the experimental case. The respective tests were ran such that, a nonintrusive situation; where no other waste material was co-disposed to interact with the main subject was done. Hence, from results and analysis the following conclusions were drawn:

- That the arsenic, chromium and copper concentrations were found to slightly increase while the pH decreased.
- The concentration levels of arsenic and copper in the generated leachate for Test-A were both found to constitute environmental risks as per [10] since they were above 10 and 1 mg/L respectively during the early stages of permeation.
- Based on the amount of leachate produced in a short-term investigation, there could be an advantage to disposing of CCA-treated wood separately if proper containment and management procedures can be guaranteed.
- From the concentration levels over such a short testing period, it appeared that the generated leachate had the potential to negatively impact human and environmental health. The disposal of this leachate however, may be unrealistic as it could be expensive to manage.

In summary, leachate from CCA-treated wood waste contains arsenic, chromium and copper. Reliable studies suggests that exposure to the arsenic from CCA-treated wood poses the greatest potential health risk. Nonetheless, it is unsure what exposure level threatens health from contact with CCA-treated wood. However, since CCA-treated wood cannot completely come out of service particularly in

developing countries, a few measures have been highlighted to reduce exposure to CCA:

i. When working with CCA-treated wood dust masks, gloves and protective clothing must be worn to reduce chances of exposure to sawdust.
ii. Sealants should be applied to CCA-treated wood structural members at intervals of 1–2 years to prevent direct contact with treatment chemicals.
iii. Children should be guided from playing in areas contaminated by CCA-treated wood as well as reminded to wash up after contact with CCA-treated wood components or playground equipment.
iv. Clients and firms are advised to consider greener optional and eco-friendly building materials i.e., hardwood and plastics as outdoor structural members.
v. Retail CCA-treated wood stores should have consumer information sheets describing safe handling recommendations. Furthermore, CCA-treated wood may be disposed as ordinary household trash but should not be burnt [8] because toxic chemicals would be released into the air or remain in the ashes. Finally, CCA-treated wood must not be mulched and sawdust from CCA-treated wood must not be added in composting piles [4].

Acknowledgements The Authors appreciate the University of Johannesburg where the study was carried out.

References

1. CPSC (2006) Evaluation of the effectiveness of surface coatings in reducing dislodgeable arsenic from new wood pressure-treated with chromate copper arsenate (CCA). Draft final report. consumer product safety commission. http://www.cpsc.gov/library/foia/foia07/os/cca.pdf
2. CDPH (2007) Pesticides used in pressure-treated wood. Hartford: connecticut department of public health. http://www.ct.gov/dph/lib/environmental_health/echa/pdf/pressure_treated_wood.pdf
3. ATSDR (2007) Toxicological profile for arsenic (update). Agency for toxic substances and disease registry, Atlanta
4. Jambeck RJ, Townsend T, Solo-Gabriele H (2008) Leachate quality from simulated landfills containing CCA-treated wood. United States, Florida, Miami
5. Agbenyeku EE, Muzenda E, Msibi IM (2015) Potential risks of CCA-treated wood destined to landfills. In: Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering and Computer Science 2015, 21–23 October, 2015. San Francisco, USA, pp 593–597
6. United States Environmental Protection Agency (US EPA) (2003a) A probabilistic risk assessment for children who contact CCA-treated playsets and decks. Draft preliminary report, office of pesticide programs, antimicrobials division, November 10, 2003
7. Code of Federal Regulations-CFR (2003) Title 40–Protection of the Environment, Chapter 1–Environmental Protection Agency, Part 261–Identification and Listing of Hazardous Waste
8. Solo-Gabriele H, Townsend T, Cai Y, Khan B, Song J, Jambeck J, Dubey B, Jang Y (2003) Arsenic and chromium speciation of leachates from CCA-treated wood. Draft report, Florida center for solid and hazardous waste management, Gainesville, FL

9. Agbenyeku EE, Muzenda E, Msibi IM (2014b) Buffering of TOC-contaminant using natural clay mineral liner. In: International conference on earth, environment and life sciences (EELS-2014) Dec. 23–24, 2014 Dubai (UAE)
10. Weber WJ, Jang YC, Townsend TG, Laux S (2002) Leachate from land disposed residential construction waste. J Environ Eng 128(3):237–245
11. Water Services Act No. 108 of 1997. Monitoring requirements and regulations of the South African National Standard (SANS). 241 Drinking Water Specification
12. Cole FA, Clausen CA (1996) Bacterial biodegradation of CCA-treated waste wood. In: Forest products society conference proceedings, September 1996. Madison, Wisconsin

Chapter 37
The Use of Mathematical Modeling for the Development of a Low Cost Fuzzy Gain Schedule Neutralization Control System

Rodrigo Sislian, Flávio Vasconcelos da Silva, Rubens Gedraite, Heikki Jokinen and Dhanesh Kattipparambil Rajan

37.1 Introduction

Water is an essential natural resource for any human activity and, for that reason, there is a need to minimize its consumption, as well as return it to the environment with minimal contamination due to its limited capacity for self-purification [1]. Thus, there is a need of balance between the increase of production without depleting natural resources, thus generating smaller amount of waste as well as recovering and reusing water as much as possible.

The pH neutralization process is used in different industrial processes such as wastewater treatment, chemical and biotechnological processes. This process make it possible to reuse the water for wide range of applications in the production, such

R. Sislian (✉)
Federal Institute of São Paulo, Avenida Salgado Filho,
3501, 07115-000 São Paulo, Brazil
e-mail: rodrigo.sislian@gmail.com; rodrigo@ifsp.edu.br

R. Sislian · F.V. da Silva
University of Campinas, Av. Albert Einstein, 500, 13083-852 Campinas, Brazil
e-mail: flavio@feq.unicamp.br

R. Gedraite
Federal University of Uberlândia, Avenida João Naves de Ávila, 2121,
38408-100 Uberlândia, Brazil
e-mail: rgedraite@gmail.com

H. Jokinen · D.K. Rajan
Tampere University of Technology, Korkeakoulunkatu, 10, FI 33720 Tampere, Finland
e-mail: heikki.jokinen@tut.fi

D.K. Rajan
e-mail: dhanesh.kr@tut.fi

S.-I. Ao et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-981-10-2717-8_37

as the first rinse for cleaning equipment in the food industry, or any need that allows the application of recycled water.

Due to its non-linearity, it is difficult to control a pH process with adequate performance [2], which bring the need of the process behavior knowledge. It can be done based on mathematical models.

Since any system has an inherent cost, with financial advantages in reducing the cost in industry, the aim of this paper is the development of a low cost prototype neutralization system using a Fuzzy Gain Schedule Controller and an Arduino board based on the system's dynamic behavior, based on an approximate mathematical model.

37.2 Theoretical Framework

Different processes in industry use the pH neutralization process, such as wastewater treatment and chemical process. A large amount of research projects has been conducted regarding the pH neutralization process, which is justified by the high non-linearity and dead time inherent in the process. That reinforces the need of studies regarding modeling and control strategies [3].

Directly related to the wastewater neutralization control, there is a need of balance, not exhausting the natural resources with the production increase, generating less waste, as well as recovering and reusing water as much as possible, which strengthens the idea of the control strategy studies.

The above mentioned is complemented by [4], describing that climate change, the limits of water supply and the continued growth of the population have intensified efforts to reduce water consumption, as well as the amount of water wasted.

Searching for different control strategies, the development of Artificial Intelligence (AI) techniques in recent years increasingly occupies a prominent position in research of industrial processes control and gradually used in industrial plants successfully, which is the case of the fuzzy control [5].

One important step for the fuzzy control system project is the system behavior response knowledge, which can be based in several methods such as the process knowledge (based on experience) or obtained from the step response curves, where it is possible to extract low order approximate models, which describes the dynamic behavior of the process [6].

Many technological processes can be approximated reasonably based on a FOPDT dynamic model [7]; this dynamic model information can be used for a fuzzy controller project.

A typical fuzzy system is described by [2], where the controller is composed by: (i) the input signal fuzzyfication (converting the sensor signal in fuzzy values); (ii) a fuzzy engine, which handles the rule inference (based on input-output relations conditions); and (iii) defuzzyfication, generating a continuous output signal for the actuators.

There are different possibilities of implementing fuzzy controllers, such as the fuzzy controller presented by [8], where the authors simulate the fuzzy controller with one input and two outputs.

A different fuzzy controller structure has been developed and simulated by [2] using an auxiliary variable to detect in which region the process is operating, with the objective of compensating the gain process non-linearity; this system has been implemented by [9] using low-cost instrumentation.

37.3 Materials and Methods

37.3.1 Experimental System

The prototype developed for the experimental tests was mounted on the Laboratory E209-B of the Department of Automation Science and Engineering of the Tampere University of Technology.

A simplified general diagram of the studied system is presented on the Fig. 37.1, where AE1 represents the pH measured sensor (Haoshi—FIT0348 low cost glass electrode pH meter). The signal converter (and AT1) consists on a motor shield (L298 N chip and two output channels—to amplify the analog output signal) and a

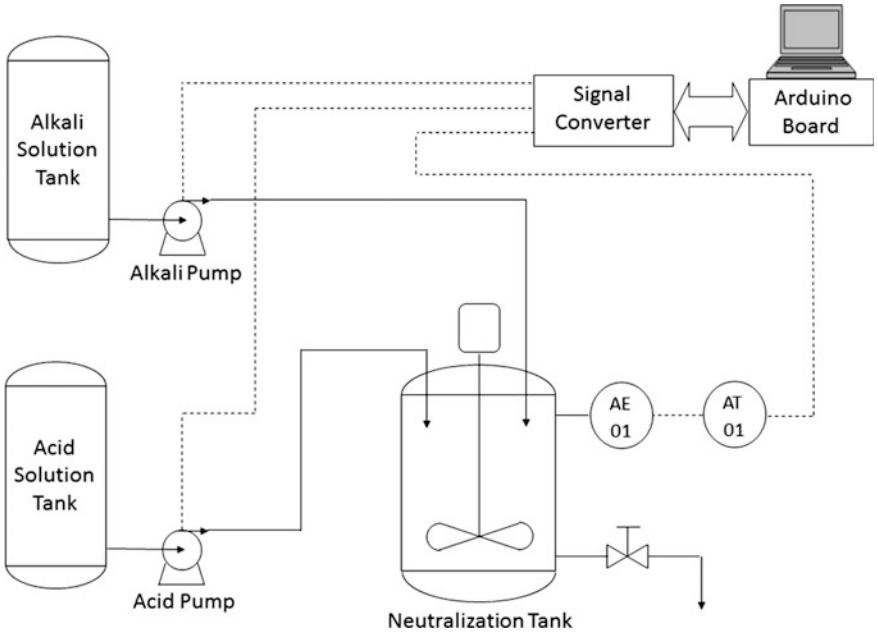


Fig. 37.1 Simplified general diagram for the studied system

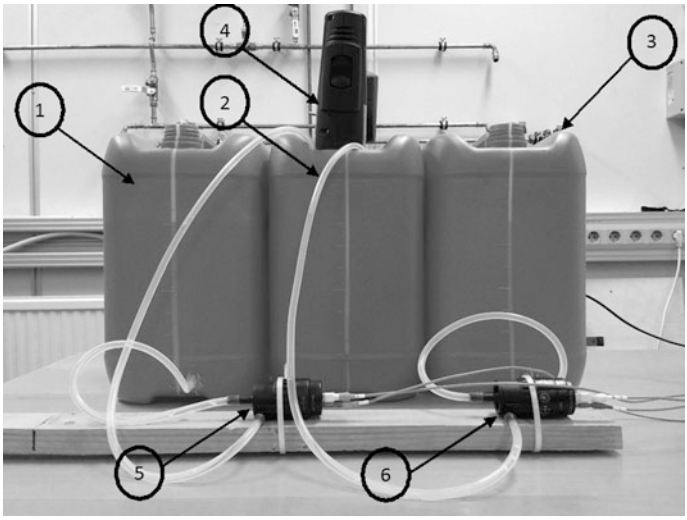


Fig. 37.2 Neutralization system prototype for the studied system

pH-Voltage converter (to convert the pH meter signal into Volts). For the acid and alkaline solution injection, it has been used 2 pumps (Drift—12VDC/2A).

The system works as described in the sequence (referring to the Fig. 37.2). The Neutralization Tank (2) collects the water that is going to be neutralized. The pH meter sends the pH value for the Arduino board and this signal is sent to the computer. Based on the control system, the acid pump (5) or the alkali pump (6) controls the acid and alkali solution injection from the tanks—acid (1) and alkali (3)—to the Neutralization Tank (2). The mixer (4) constantly homogenizes the neutralization tank solution.

The electronic system used to collect data is composed by a computer, an Arduino Mega board 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 for the proceeded experiments, from which the data for the system behavior study and the control of the system has been developed.

To make it possible the integration between the Arduino board and the Labview tool it was installed the Arduino Toolkit, then the Interface for Arduino Firmware was loaded in the Arduino board. After that, it was possible to develop the application.

An important step was the calibration of the pH meter (based on buffer solutions). For that, the pH meter sensor was inserted in each solution used, and the voltage sent from the transmitter was measured.

Then, it was generated the converting equation, making it possible for the system to calculate the pH, based on the voltage sent from the transmitter.

37.3.2 Open Loop Tests

In order to know the system behavior for the instrumentation used, mainly the low-cost pumps and pH meter, it has been developed an application using the Labview tool.

Then, after calibrating the pH meter, it has been proceeded open-loop tests applying step disturbances on the acid and alkali pumps separately, making it possible to acquire the pH system response for each different step on the pumps.

The first step tests applied on the input were from 0 to 60 %, 0 to 80 % and 0 to 100 % on the acid pump with the alkali pump closed. Then, the same step tests were applied to the alkali pump (0 to 60 %, 0 to 80 % and 0 to 100 %) with the acid pump closed. The system response for the step from 0 to 100 % is presented on Fig. 37.3 to illustrate the open loop tests response.

37.3.3 FOPDT Model Development

After obtaining the open loop test responses, the process models were calculated using the FOPDT (first order plus dead time) approach, which is largely used to represent industrial processes dynamics [10].

The non-parametric identification employs the response curves of the process when excited by input signals as step changes. From the curves obtained, it is possible to extract low order approximate models, which describes the dynamic behavior of the process [6].

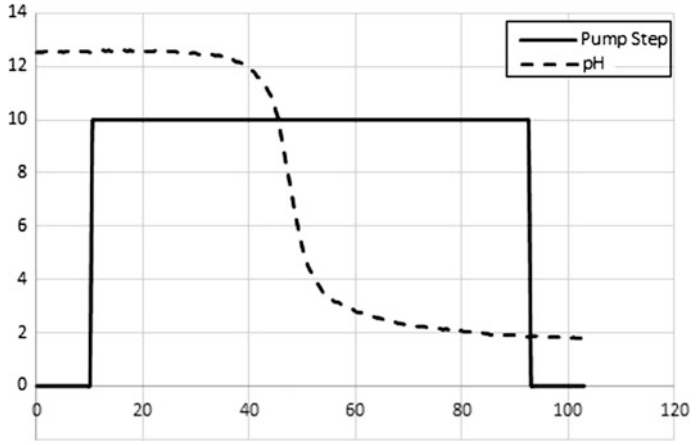


Fig. 37.3 Acid pump step response 0 to 100 %

Since the process is non-linear, it has been defined three regions for the acid and three regions for the alkaline step, and calculated the transfer functions using the Smith Method in order to develop the rules and the membership functions.

For the Smith Method application, it necessary to measure the variation of the controlled variable (Pump Voltage) and divide it for the process variable variation (pH) obtaining the gain K. After that, times values $t_{0.632}$ (time for the process variable in 63.2 % from steady state value) and $t_{0.284}$ (time for the process variable in 28.4 % from steady state value) are calculated. Based on that, τ (process time constant) and θ (process transport delay) are calculated as presented by the equations:

$$\begin{cases} \tau = 1.5 * (t0.632 - t0.284) \\ \theta = t0.632 - \tau \end{cases}$$

The process parameters, K_p (process gain—that represents the relationship of the pH in response to the variation of the pump voltage), τ_p (process time constant) and θ_p (process transport delay) were calculated using the process response curve method (for the pH response) to a step change in the pump voltage.

After calculating the parameters, the next step was to compare the collected data with the FOPDT models calculated and proceed with the manual adjust of the parameters. It was simulated using the SIMULINK/MATLAB® software and the results are presented on the Figs. 37.4, 37.5, 37.6, 37.7, 37.8 and 37.9.

The empirical models parameters obtained after the simulation and adjust are presented on Table 37.1 (acid pump) and Table 37.2 (alkali pump).

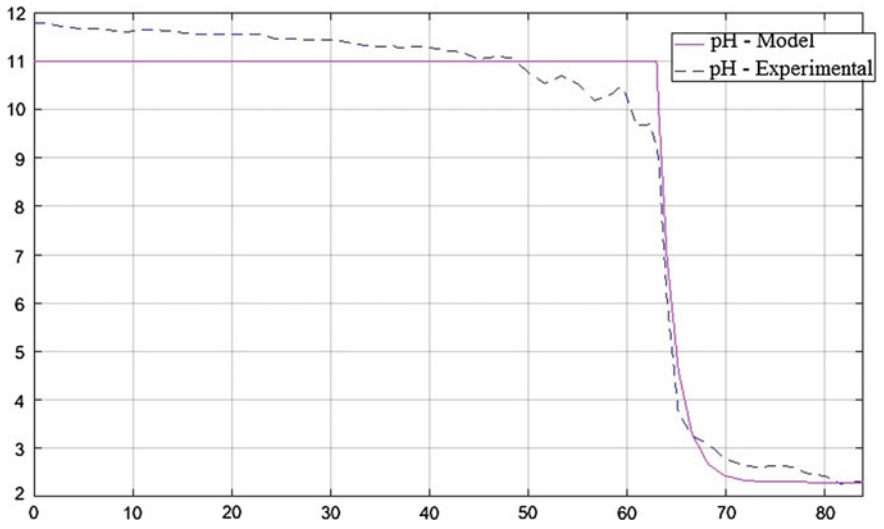


Fig. 37.4 Model versus experimental data for a step from 0 to 60 % in the acid pump voltage

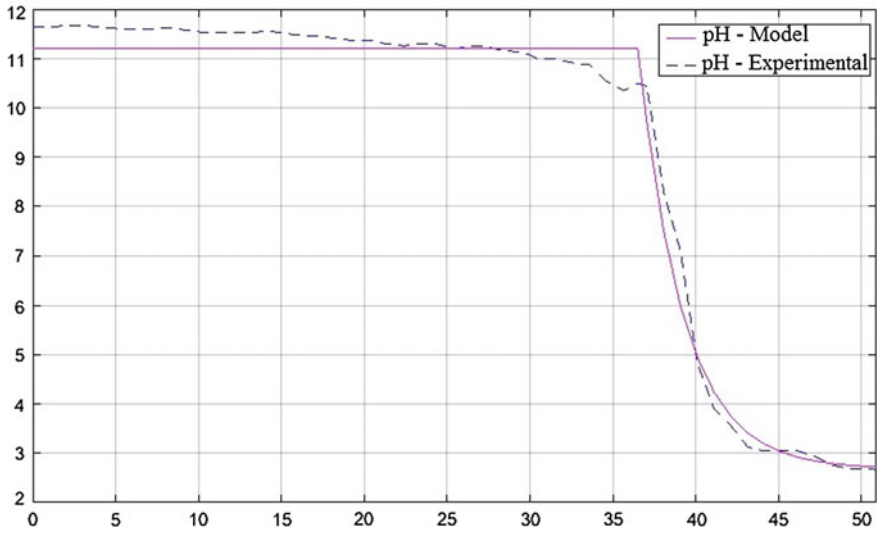


Fig. 37.5 Model versus experimental data for a step from 0 to 80 % in the acid pump voltage

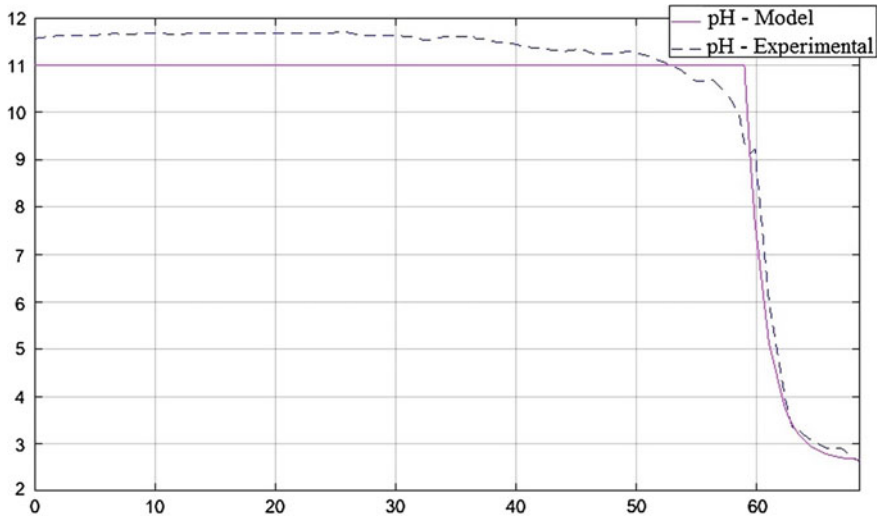


Fig. 37.6 Model versus experimental data for a step from 0 to 100 % in the acid pump voltage

Based on the obtained pH responses and the models presented, it was possible to identify the system behavior response. This information was used for the Fuzzy Controller project.

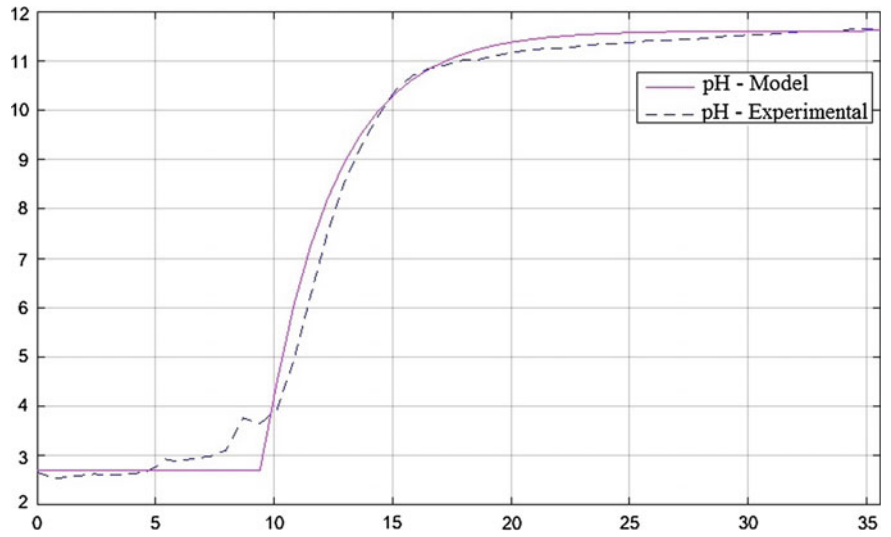


Fig. 37.7 Model versus experimental data for a step from 0 to 60 % in the alkali pump voltage

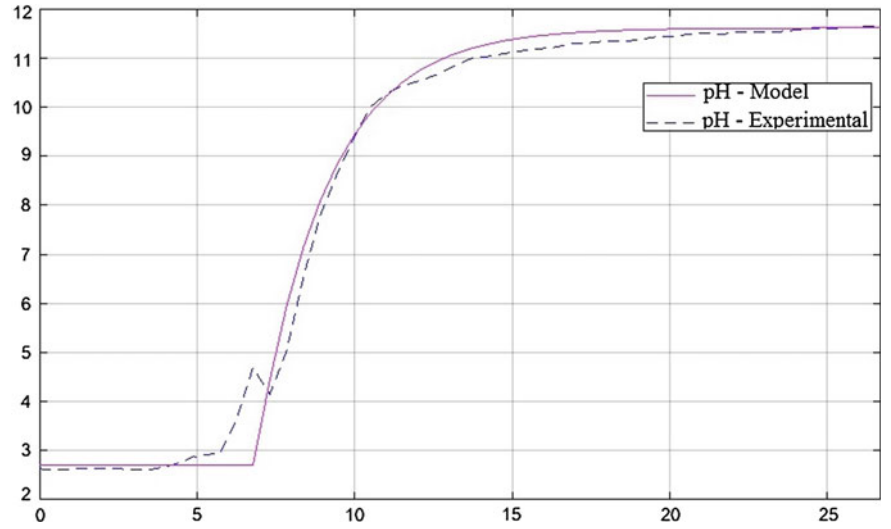


Fig. 37.8 Model versus experimental data for a step from 0 to 80 % in the alkali pump voltage

37.3.4 Fuzzy Controller Membership Functions Project

Once known the system behavior, the next step was the Fuzzy Controller project. After a research on the Neutralization System Controller and after testing some

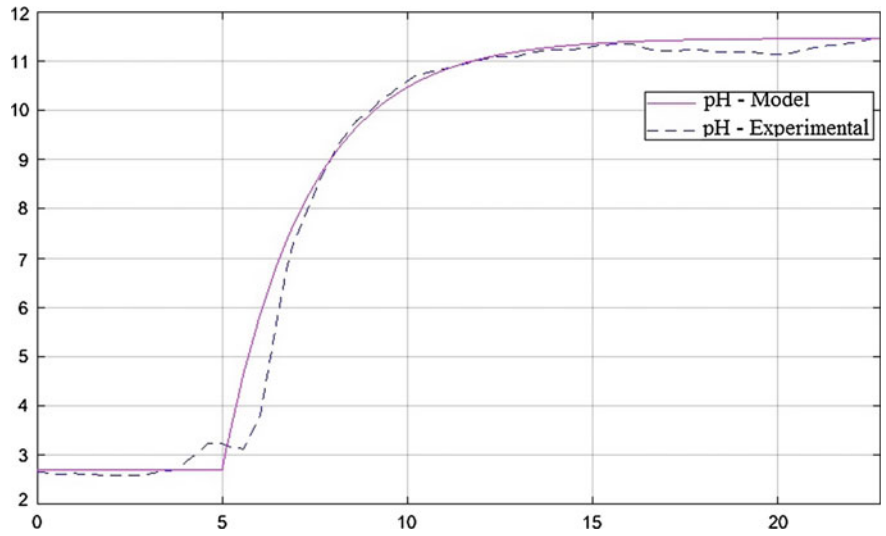


Fig. 37.9 Model versus experimental data for a step from 0 to 100 % in the alkali pump voltage

Table 37.1 Empirical models parameters—acid pump

Step change (%)	Gain Kp (pH/%)	Transport delay θp (s)	Time constant τp (s)
0–60	1.45	311.5	8.25
0–80	1.075	184.5	13.5
0–100	0.76	297.25	8.25

Table 37.2 Empirical models parameters—alkali pump

Step change (%)	Gain Kp (pH/%)	Transport delay θp (s)	Time constant τp (s)
0–60	1.31	49.5	13.5
0–80	0.99	35	9.75
0–100	0.71	28	7.5

different techniques, the fuzzy controller structure proposed by [8] was used as a starting point for the development of the Low Cost pH Neutralization System proposed in this paper.

The process has been divided into three regions of non-linear gains: pH Acid, pH Neutral and pH Alkali. The process gain, as [8] states, is low in the pH Acid and pH Alkali regions (i.e. considerable changes have to be made in the acid/alkali pump, to make appreciable changes in the pH value); and the process gain is high in the pH Neutral region (i.e. a small change in the acid/alkali pump, results in a large change in the pH value).

This division was used for the Auxiliary Variable (AV) definition as an input for the Fuzzy Gain Schedule Controller. The three trapezoidal fuzzy membership

Fig. 37.10 Membership function for the auxiliary variable

functions for this variable (based on the process knowledge and after practical tests) were defined with the linguist terms as Acid, Neutral and Alkaline and the discourse universe between [1, 14] (presented on Fig. 37.10).

The next step was to define the other fuzzy membership functions related to the controller inputs. The control error has been defined by five triangular membership functions and the linguist terms defined were NL (negative large), NS (negative small), Z (zero), PS (positive small) and PL (positive large) with the discourse universe between [−7, 7]—since pH range is 14; and the change in the control error has the same membership functions but with discourse universe between [−2, 2]—based on practical tests.

Finally it has been defined the fuzzy membership functions related to the controller outputs for the Acid and Alkali pumps (18 triangular membership functions) and the linguist terms defined were: Z0Alkali, AlkaliPumpOff, SlowAlkali, AverageAlkali, FastAlkali, zz0Alkali, zSlowAlkali, zAverageAlkali, zFastAlkali, Z0Acid, AcidPumpOff, SlowAcid, AverageAcid, FastAcid, zz0Acid, zSlowAcid, zAverageAcid and zFastAcid. The discourse universe has been defined between [0, 90]—using the pump range from 0 to 90 %.

The membership function for the controller outputs are presented on Fig. 37.11.

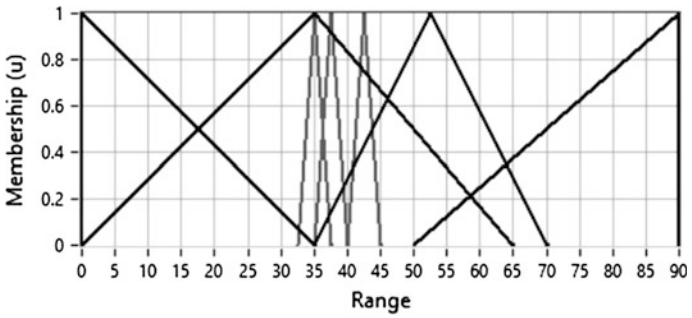


Fig. 37.11 Membership function for the controller output

37.3.5 *Fuzzy Controller Rules Definition*

The next and final step for the Fuzzy Gain Schedule Controller development was the definition of the rules. The defuzzification method defined was Center of Area.

It was defined the rules for each process region. As [5] presented, by means of the Auxiliary Variable (AV), it was possible to simplify the rules, and it was defined a set of 25 fuzzy rules for the AV = neutral and 25 rules for AV = not-neutral. The tool for the rules implemented on the Labview® is presented on Fig. 37.12.

The rules are divided in two different kinds, one group for AV = neutral and the other for AV = not-neutral. Two rules examples (from the set of 50 rules) are presented below, to illustrate the rules defined:

(1) IF 'AV' IS NOT 'Neutral' AND 'Error' IS 'NL' AND 'DeltaError' IS 'NL' THEN 'Acid Pump' IS 'FastAcid' ALSO 'AlkaliPump' IS 'AlkaliPumpOff'
(2) IF 'AV' IS 'Neutral' AND 'Error' IS 'NS' AND 'DeltaError' IS 'NS' THEN 'Acid Pump' IS 'zFastAcid' ALSO 'AlkaliPump' IS 'AlkaliPumpOff'

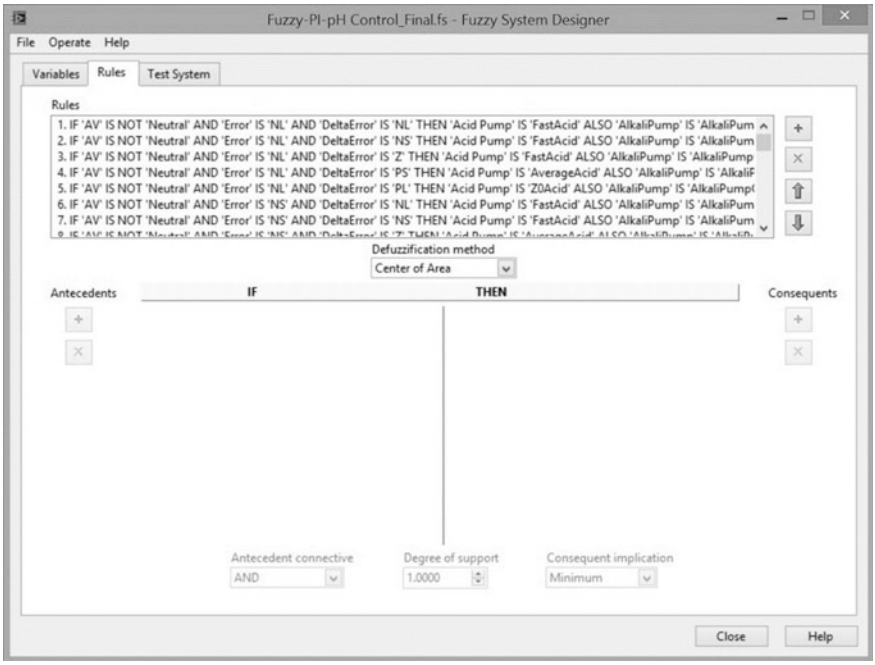


Fig. 37.12 Tool for the rules implementation on labview®

37.3.6 Fuzzy Gain Schedule Controller Implementation and Application

After defining and implementing the Fuzzy Gain Schedule pH Neutralization Controller using the Labview tool, the block diagram is presented on Fig. 37.13 and the data acquisition system front panel with all the variables defined is presented on Fig. 37.14.

To evaluate the control performance, it has been set-up the pH set-point to 7 (for pH neutralization) and inserted a alkali solution (pH 6) in the Neutralization Tank.

After the water inside the tank was neutralized, it was inserted pure acid, waiting for the system to neutralize the water and then pure alkaline solution, simulating system disturbances. It is possible to see in Fig. 37.15—where it is presented the percentage of voltage sent to the pump, the pH and pH set point—the system neutralizing the pH of the water inside the Neutralization tank, showing that the neutralization control system is functional.

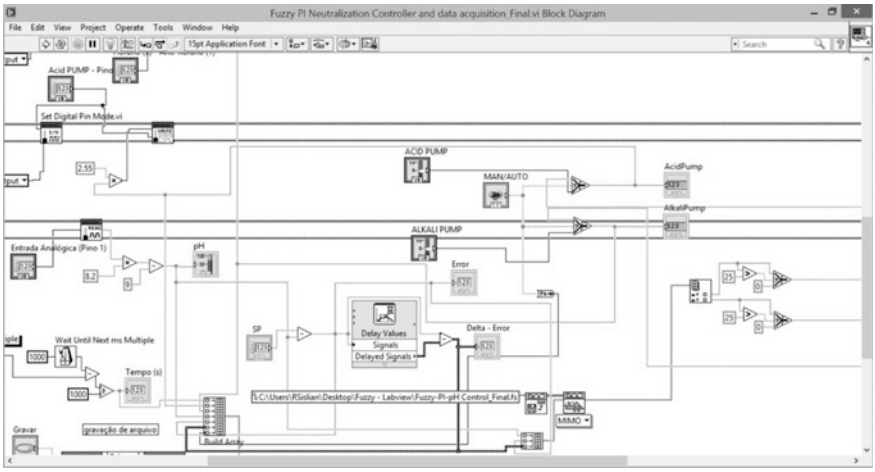


Fig. 37.13 Block diagram of the fuzzy gain schedule pH neutralization control

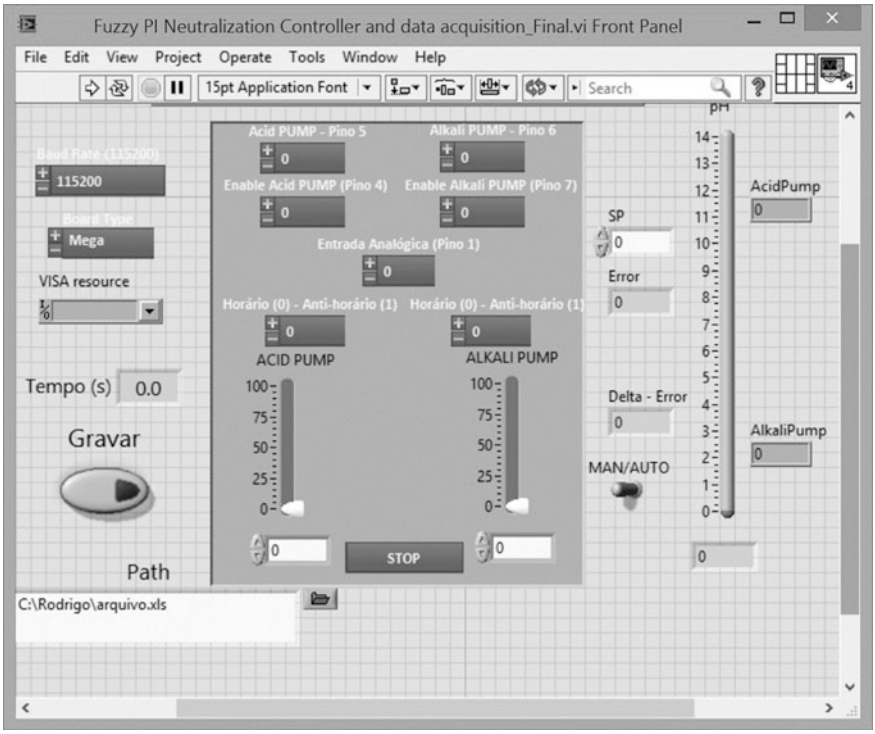


Fig. 37.14 Front panel of the fuzzy gain schedule pH neutralization control system

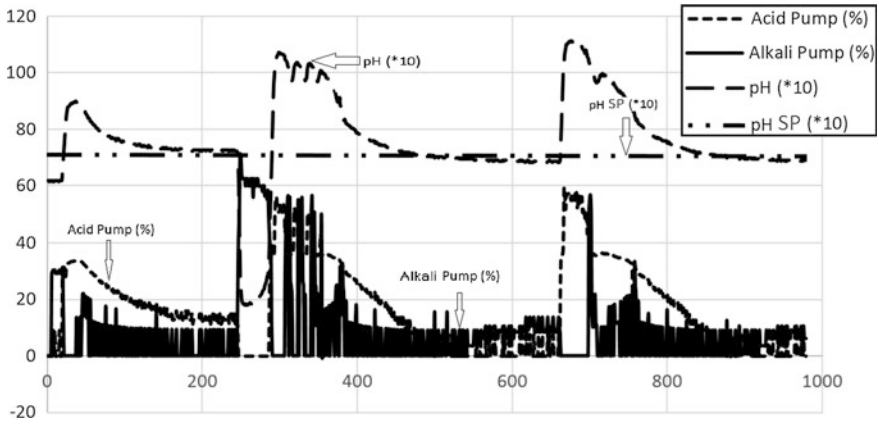


Fig. 37.15 pH neutralization system response

37.4 Conclusion

The implementation and instrumentation of a typical pH Neutralization System, with an appropriate monitoring, control and data acquisition of the process variables has been successfully implemented, as well as the Fuzzy Gain Schedule pH neutralization controller.

The fuzzy controller implementation was based on the process models calculated using the FOPDT (first order plus dead time) approach, making it possible to describe the dynamic behavior of the process.

The development of the control system considered in this paper reveals an application of the non-parametric identification (mathematical modeling) and the Fuzzy Gain Schedule Control, and an attractive Industrial application perspective, representing a potential application for water consumption reduction in industry, based on low cost elements, whit a rapid return on investment.

Acknowledgements The authors acknowledge the financial support provided by Federal Institute of São Paulo, by CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) SETEC/MEC Nº 015/2014, by Tampere University of Applied Sciences and by Tampere University of Technology—Department of Automation Science and Engineering.

References

1. Alvarez D, Garrido N, Sans R, Carreras I (2004) Minimization–optimization of water use in the process of cleaning reactors and containers in a chemical industry. J Cleaner Prod 12(7): 781–787, Set
2. Fuente MJ, Robles C, Casado O, Tadeo F (2002) Fuzzy control of a neutralization process. In: Proceedings of the 2002 IEEE international conference on control applications, vol 2. Glasgow, Scotland, pp 1032–1037
3. Morales C, Lourenço L, Garcia C (2013) Predictive Controller Applied in a pH Neutralization Pilot Plant (Controle Preditivo Aplicado a uma Planta Piloto de Neutralização de pH). Simpósio Brasileiro de Automação Inteligente. Fortaleza, Brazil, 2013
4. Bansal B, Chen XD (2006) A critical review of milk fouling in heat exchangers. Compr Rev Food Sci Food Saf 5(2): 27–33
5. Leite MS, Fileti AMF, Silva FV (2010) Development and experimental application of fuzzy and conventional controllers in a bioprocess (Desenvolvimento e Aplicação Experimental de Controladores Fuzzy e Convencional em um Bioprocesso). Revista Controle and Automação 21(2):147–158
6. Liu T, Gao F (2012) Industrial process identification and control design. Springer, London—UK
7. Vodencarevic A (2010) Design of PLC-based smith predictor for controlling processes with long dead time. In: proceedings of the international multiconference of engineers and computer scientists 2010, Vol, IMECS, 2010, 17–19 March, 2010, Hong Kong, pp 940–945
8. Mohd IM, Amin MKM (2013) Fuzzy controller modeling of a pH level control system. In: 32nd JSST annual Conference, International Conference on Simulation Technology, Tokio, Japan, Sep., 2013

9. Sislian R, Silva FV, Gedraite R, Jokinen H, Rajan DK (2015) Development of a low-cost fuzzy gain schedule neutralization control system. In: Lecture notes in engineering and computer science: proceedings of the world congress on engineering and computer science 2015, WCECS 2015 -23 October, 2015, San Francisco, USA, pp 575–578
10. Âström K, Hägglund T (1995) PID controllers: theory, design and tuning. ISA. North Carolina—US

Chapter 38
Recovery of VOC from Offshore and Onshore Shuttle Tankers Using Structured Y-Type Zeolite Membranes

Shehu Habiba, Adebayo Ajayi, Okon Edidiong and Edward Gobina

Nomenclature

FPSOs	Floating Production, Storage Offloading units
HC	Hydrocarbon
IG	Inert gas
NMVOC	Non-methane volatile organic compound
PI	Process Intensification
RVP	Reid Vapour Pressure
VOC	Volatile organic compound

Symbols

A	Area of the membrane (m^2)
P	Permeability (mol m m^{-2} s^{-1} Pa^{-1})
Q	Molar gas flow rate (mol s^{-1})
R	Molar gas constant (8.314 J mol^{-1} K^{-1})
ΔP	Pressure drop across the membrane (Pa)
$\alpha_{i,j}$	Selectivity of component *i* over *j*

S. Habiba · A. Ajayi · O. Edidiong · E. Gobina (✉)
Centre for Process Integration and Membrane Technology (CPIMT), School of Engineering, The Robert Gordon University Aberdeen, AB10 7GJ, Aberdeen, UK
e-mail: e.gobina@rgu.ac.uk

S. Habiba
e-mail: h.shehu@rgu.ac.uk

A. Ajayi
e-mail: a.e.ajayi@ru.ac.uk

O. Edidiong
e-mail: e.p.okon@rgu.ac.uk

S.-I. Ao et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-981-10-2717-8_38

38.1 Introduction

Volatile organic compounds are chemicals that have high vapor pressure at room temperature. Due to their high vapor pressures, large numbers of molecules evaporate from the initial form of the compound and enter into the atmosphere. These compounds are harmful to the environment; methane is a potent greenhouse gas, which contributes to ozone layer depletion. Other Non-methane VOCs (NMVOCS) such as ethane, butane, hexane, pentane and propane react with nitrous oxide to form ground level ozone, which affects both human and plants [1].

Apart from the VOC's being harmful to the environment, they constitute significant economic value that should not be wasted. The sources of the release of VOCs in the onshore and offshore facilities to the atmosphere mostly occur during production, transportation and storage of crude oil [1]. The light hydrocarbons mostly vaporize out of the crude oil during the loading and unloading operations of shuttle tankers and also from Floating Production Storage and Offloading units (FPSOs).

Due to the adverse effects of Volatile Organic Compounds (VOCs) in the environment and also the economic loss of these hydrocarbon compounds, some technologies are used for the abatement of VOCs that are released from both onshore and offshore facilities. Some of the commercially viable methods for treating VOCs emissions include; absorption, adsorption, thermal oxidation, membrane separation and cryogenic condensation.

38.1.1 VOCs Emission Sources

VOCs are emitted to the atmosphere during, production, storage and transportation of crude oil. The two main sources of light hydrocarbon emissions in oil and gas production can be found in onshore and offshore facilities.

38.1.2 Factors affecting the Rate of VOC Emission from Onshore and Offshore Facilities

The rate of emission of Volatile Organic Compounds depends on various factors. These factors are discussed below:

38.1.3 Loading Time

The emission of hydrocarbon gas from a cargo tank is a non-equilibrium process and hence the emission will increase with increase in loading time. The longer the loading operation, the higher the percentage of HC gas emitted [2].

38.1.4 Effects of Ship Movement

During loading operations, the movement of the ship can alter the emission rates, also movements due to weather tends to increase circulation in the cargo tank between the crude oil and the atmosphere thus increasing a higher blending between HC vapor and IG which in turn increases the vaporization rate of the HC. This is more prevalent when the shuttle tanker is short loaded and the rolls or pitch could cause more splashing [2].

38.1.5 Nature of Hydrocarbon

A high concentration of light hydrocarbon affects the rate of VOC emitted, since the vaporization rate of the hydrocarbon is increased which thus lead to more hydrocarbon being emitted to the atmosphere. Crude oil composition with high concentration of light hydrocarbons such as methane and ethane tend to have high hydrocarbon vapor, which can mix with inert gas as a result of convection, and then cause high rate of hydrocarbon vaporization from the loading facilities [3].

38.1.6 Temperature of Crude Oil

The temperature of crude is a very important factor in oil and gas production. The Reid Vapor Pressure (RVP) of crude oil is determined at a specific temperature (37.8 °C), which makes RVP independent of temperature. The difference in temperature of the ship and the shore tank leads to a variation between the volumes of liquid loaded and vapor displaced. The vapor displaced into the shore tank expands and warm up when the ship's tank is colder than the shore tank [4]. This leads to subsequent emission and increase in pressure inside the tank.

VOC Reduction Techniques

There are various available techniques for treating VOC emissions from onshore and offshore facilities during the loading and unloading operations. Some of these technique include; absorption, thermal oxidation, adsorption, condensation, reducing volatility and membrane separation [4].

38.1.7 Absorption

Absorption method for VOC recovery is a technology developed by cool sorption. It is basically used to recover non-methane VOCs in chilled liquid or pressurized crude oil (8–11 bar) [5]. This process involves feeding the bottom of the packed

column with vapor from the tank during loading operation. The vapor moves upward and is in counter current contact with the chilled liquid absorbent flowing downwards. The absorbent dissolves the hydrocarbon from the vapor/air liquid and removes it from the mixture. The residual air is vented to the atmosphere as it moves out of the top of the column. There is regeneration of the absorbent liquid in the stripping section of the absorption system [4]. Methanol is injected into the absorption system to prevent the formation of hydrates in the vent gas.

38.1.8 Thermal Oxidation

Thermal oxidation method of controlling VOC emission is also called combustion method and is widely used in USA. The combustion system ranges from simple enclosed fares to catalytic oxidizers with internal heat recovery. The issue of safety and combustion emission of CO_2 are the major negative attributes of thermal oxidation. Energy recovery is used to minimize the implication of CO_2 emission while the use of enrichment dilution, detonation arrestors coupled with effective management procedures ensures safety of the system [4].

38.1.9 Adsorption

Adsorption system is mainly used to separate inert gas from hydrocarbon fractions. There are various adsorption systems, one of which is the use of activated carbon. In this type of adsorption technique, the carbon absorbs the organic molecules and the gases like CO_2 and air move through the bed unabsorbed and are emitted to the atmosphere [4]. Before the bed becomes saturated, regeneration is carried out either by vacuum steam stripping, but once the bed becomes saturated, the adsorption process stops and the vapor moves through the bed directly without being absorbed. The use of two beds increases the efficiency of the adsorption process. This continuous operation process involves the use of one bed for the operation, while the other bed is being regenerated.

38.1.10 Cryogenic Condensation

This is method of VOC reduction involving the passing a mixture of VOC containing gas through a liquid nitrogen cooled condenser [4]. This technology is often used in pharmaceutical industry. It is a cost-effective technique of VOCs emission control when compared with the other technologies already discussed [6]. A mathematical description of the process can be developed in order to design a counter current single tube condenser using nitrogen vapor as coolant [7].

38.1.11 Reducing Volatility

This is one of the simplest ways to reduce VOC emissions. It entails the reduction of the volatility of the cargo. Although the vapor pressure of pure substances cannot be altered, changing the composition of the crude to include more of heavier molecular weight compounds and less of lighter molecular weight ones can reduce the gasoline vapor pressure. The volatility of the crude should be reduced before loading and storage.

38.1.12 Sequential Transfer

In this system, additional pipelines are installed on the shuttle tank and are used during loading and offloading of the tank. The shuttle tank is divided into sections and the loading and discharging is done sequentially. The gas out-flow from the first loaded section is piped to the bottom of the next section to be loaded, this is repeated sequentially for the remaining sections and the cargo tank is connected to a VOC recovery plant, thus the IG content emitted is considerably low in this system which improves the operating conditions of the recovery plant [2].

38.1.13 Membrane Separation

Membrane technology can be used for the separation of hydrocarbons from inert gases and the different hydrocarbon gases emitted can be individually separated. The separation concept of hydrocarbons from a shuttle tanker using membrane technology requires another process for the recovery of the VOC such as condensation or adsorption [2]. Modern membrane technology can also be used in implementing Process intensification (PI) which is an innovative design method that is aimed at decreasing production cost, waste generated and size of equipment used as well as energy utilization [8]. This process is a pressure driven process that has numerous industrial applications in chemical and petro chemical industries including petroleum refineries [9]. In recent years there has been an increase in demand for light hydrocarbons for use as petrochemical feed stocks and for fuel. This has been met to a large extent by the recovery of these components from petroleum gases and natural gas by the use of a depropanizer [10]. The use of membranes for gas separations is growing at a slow but steady rate [11]. Baker in 2002 made an estimate that the market scale of gas separations using membrane technology by the year 2020 will be five times that of year 2000 [11]. Microporous inorganic membranes having pore sizes of less than 1 nm have been studied extensively for gas separation applications because of their good resistance to harsh chemicals, good thermal and mechanical stability as well as stability under high

pressure when compared to polymeric membranes [12]. Today much of the research work is being directed towards the investigation of new membrane material and the development of new membrane structures that exhibit both higher selectivity and permeability of the target gases [13].

Different types of membranes have been studied for the separation of VOCs from inert gases. This process involves the dissolving of vapor molecules in the membrane, which moves by the principle of diffusion to the other end and desorb into the membrane support material [14]. The principle of pressure differential drives the diffusion process. For mesoporous membranes, separation is based on the collision between the gas molecule and the membrane pore wall and hence the mean free path of the gas molecules is greater than the pore size. The diffusion here is governed by Knudsen mechanism and the rate of transport of any gas is inversely proportional to the square root of its molecular weight [14]. However, for a micro porous membrane with pore size less than 2 nm, separation of gases is based mostly on molecular sieving. The transport mechanism in these membranes is often complex and involves surface diffusion that occurs when the permeating species exhibit a strong affinity for the membrane surface and thus adsorbed on the walls of the pores [14].

A membrane's permeance and selectivity has an influence on the economics of a gas separation process [13]. Permeance is the rate at which a substance permeates through a membrane and is dependent on several factors like the pore size and material of the membrane. The selectivity of a membrane is the fundamental parameter to achieving high product purity and high recoveries; hence for the potential growth of membrane gas separation process, the production of highly selective membranes for the desired gas is essential. The development of inorganic membranes like silica and zeolite has increased the potential of membrane gas separation applications as they can withstand aggressive chemicals as well as high temperatures. However, there are drawbacks on the use of such membranes, which includes their high cost, modest reproducibility.

The permeance P ($molm^{-2}s^{-1}Pa^{-1}$) represents the proportionality coefficient with the flux at steady state of a particular gas through a membrane and is given by:

$$P = \frac{Q}{A.\Delta p}$$

Where Q is the molar gas flow rate through the membrane (mol s^{-1}), A is the membrane surface area (m^2) and Δp is the pressure difference across the membrane (Pa). The permeance is therefore a measure of the quantity of a component that permeates through the membrane [15].

The calculated gas selectivity is the ratio of the permeability coefficients of two different gases as they permeate independently through the membrane is given by:

$$\alpha_{ij} = \frac{P_i}{P_j}$$

Where P_i and P_j is the permeance of the single gases through the membrane.

The selectivity is the measure of the ability of a membrane to separate two gases and it is used to determine the purity of the permeate gas as well as determine the quantity of product that is lost.

Gas separation can be used for various applications such as pollution control, photochemical process, oxygen enrichment, pharmaceutical process and many more [16].

This research adapts the use of Y-type Zeolite membrane for the separation and subsequent recovery of hydrocarbon gases under varying conditions of temperature and pressure since zeolites have the ability to withstand high temperature and pressure as well as fine pore size distribution that is highly selective to hydrocarbon gases.

38.2 Experimental

38.2.1 Permeation Setup

A schematic diagram of the membrane flow apparatus used for the permeation test for the gases is presented in Fig. 38.1. Four different gases: carbon dioxide, helium, nitrogen, methane and propane were used for the permeability through a porous zeolite membrane at various transmembrane pressures.

38.2.2 Membrane Preparation

A solution containing Silicone oxide, aluminum oxide, sodium oxide and deionized water was prepared and homogenized at room temperature for 20 h; the amount of each substance used is given in Table 38.1. Zeolite crystals were deposited on

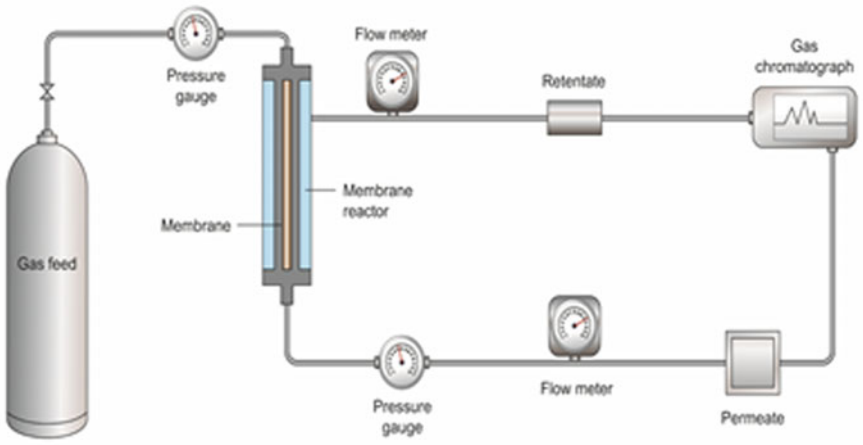


Fig. 38.1 Schematic diagram of a membrane permeation flow system

Table 38.1 Composition of the modification solution for zeolite membrane

Chemical	Amount (ml)
Aluminium oxide	10
Sodium hydroxide	14
Deionised water	798
Silicone oxide	1

Table 38.2 Optimum operating conditions of the quanta chrome gas analyzer

Parameter	Value
Area (A^2mol_{-1})	16.2
Non-ideality (1/mmHg)	6.58×10^{-5}
Sample cell type (mm)	12
Analysis time (mins)	237
Mol weight ($gmol^{-1}$)	28.0134
Ambient temperature (K)	300
Bath temperature (K)	77

alumina support, which is subsequently dipped into the solution and kept for 20 h at 343 K. The membrane was washed with deionized water and the pH of the rinse water was monitored. When the rinse water pH was neutral the membrane was air dried for 20 min and thermally treated in the oven at 338 K for 2 h prior to permeation test [17].

38.2.3 Membrane Characterization

The morphology of the membrane was determined by the use of the Zeiss EVO LS10 scanning electron Microscope. Nitrogen physisorption measurements were carried out at 77.35 K using a Quanta chrome adsorption gas analyzer. The operating conditions of the instrument in given in Table 38.2.

The experimental rig (Fig. 38.1) was used to determine the permeance of the gases. The gases were fed to the zeolite membrane from the gas cylinder through the gas inlet; the pressure was controlled at the inlet port by a pressure gauge. The permeate flow rate of the individual gases was measured by a digital flow meter in liters per minute.

38.3 Results and Discussion

The zeolite membrane showed permeance in the range of 10^{-6} $molm^{-2}s^{-1}Pa^{-1}$ for CO_2, CH_4, He and N_2, but in the range of 10^{-7} for propane. These permeances are relatively high when compared to literature values [18]. The maximum selectivity for this membrane was calculated and presented in Table 38.3.

Table 38.3 Maximum selectivity of methane through a zeolite membrane at 293 K

Gas mixture	CH_4/CO_2	CH_4/C_3H_8	CH_4/N_2	CH_4/He
Selectivity	2.9	3.3	1.4	1.2

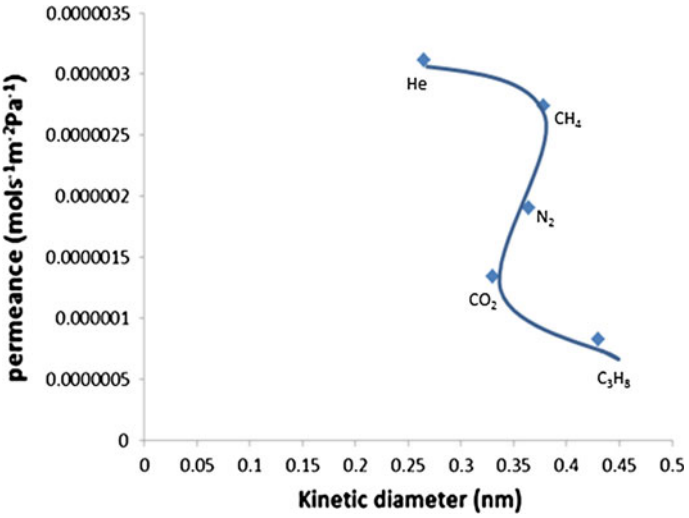


Fig. 38.2 Effect of kinetic diameter on gas permeance at 293 K and 10^4 Pa

The selectivity of methane over propane (Table 38.3) is higher than the values ranging from 1.42 to 2.56 obtained from the work of Tirouni et al. [19].

The effect of the kinetic diameter on the permeance of the gases is depicted in Fig. 38.2. The order of the kinetic of the gases is: He < CO_2 < N_2 < CH_4 < C_3H_8. Hence, the separation of helium, nitrogen and propane as observed in Fig. 38.2 was based on molecular sieving properties of zeolite. Carbon dioxide and methane deviated from the expected pattern. This could indicate the presence of inter-crystalline defects in the zeolite membrane [20].

The molar fluxes are linear functions of the pressure drop across the zeolite membrane [19]. Figure 38.3 shows the molar fluxes of the gases increase linearly with the increase in pressure. Pressure has significant effect on the gas flux (Fig. 38.3). The difference between the fluxes of the gases increases with increase in pressure, rate of increase of the flux of carbon dioxide, propane, methane, oxygen and nitrogen was observed to have slightly reduced at a higher of pressure of about 1.0×10^5 Pa and higher. The contribution of viscous flux to the overall mass transfer at higher pressure for zeolite membrane might have caused this. A good linear regression value in the range of $R^2 = 0.99$ was observed.

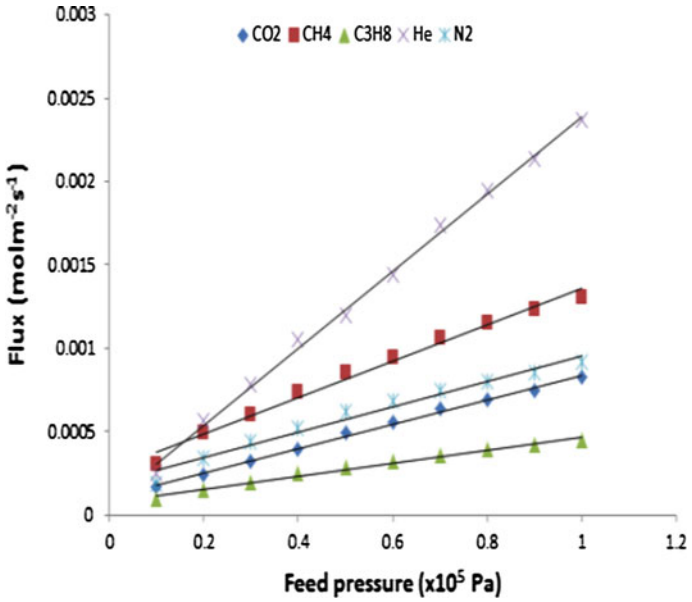


Fig. 38.3 Molar flux of the gases through zeolite membrane at 293 K

Figure 38.4a, b shows the cross sectional view and outer surface of the zeolite membrane. The surface of the membrane is covered with a dense layer of zeolite crystals.

Figure 38.4 clearly shows the zeolite crystals that are deposited on the alumina support. The cross sectional and outer images show a good bonding of the zeolite to the support. The cross sectional view showed the formation of an intermediate layer on the support. This might have led to an improvement in the affinity amidst the

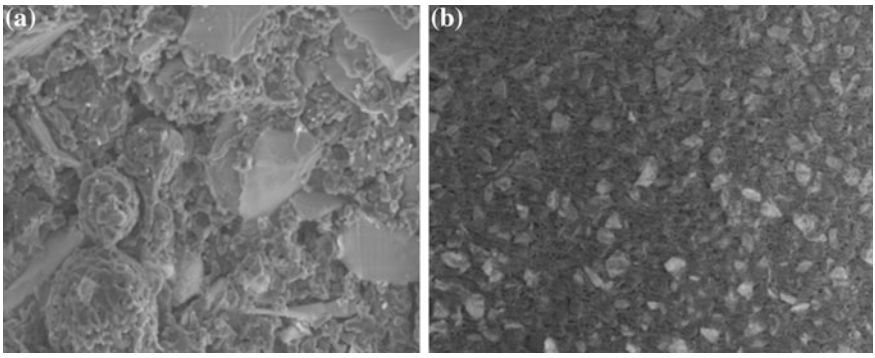


Fig. 38.4 SEM images of the cross sectional view **a** and outer surface **b** of zeolite membrane

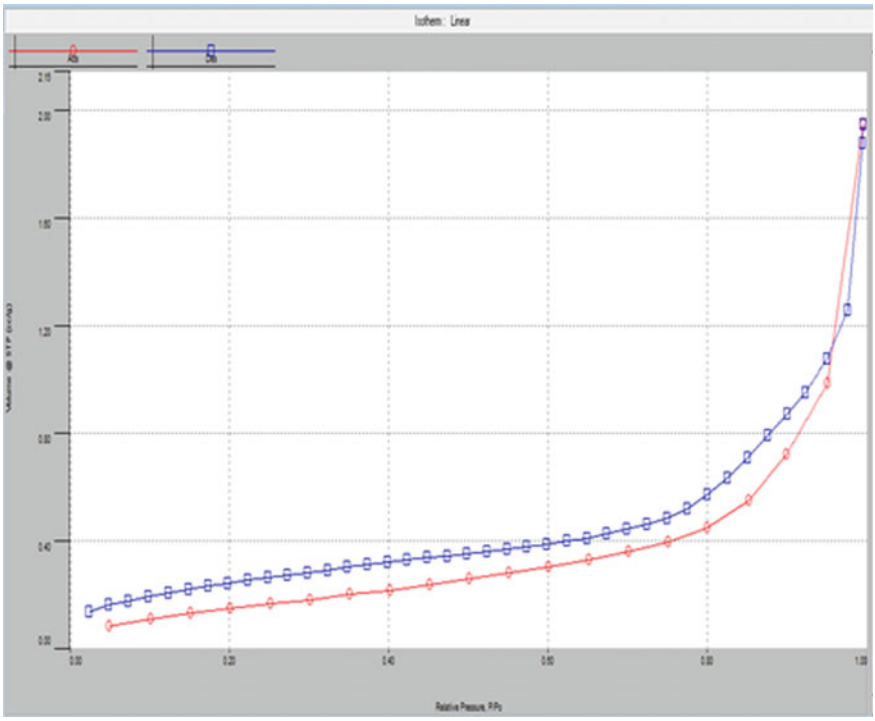


Fig. 38.5 Physisorption isotherm of the zeolite membrane

Table 38.4 Desorption summary of the zeolite membrane

Parameter	Value
Pore volume (cc/g)	0.003
Pore diameter Dv(d) (nm) Surface area (m^2g^{-1})	3.94 0.520

zeolite top layer and the alumina support. The zeolite crystals deposited on the surface of the alumina support is visible from the micrograph in the outer surface image in Fig. 38.4b. The pore sizes seen on the outer image was observed to be larger, this could have been caused by the result of interaction of the solution used to coat the membrane surface.

It can be observed that the pore diameter is not uniformly distributed.

The hysteresis isotherm in Fig. 38.5 implies the membrane is mesoporous and could undergo capillary condensation during hysteresis. Table 38.4 shows the desorption summary of the membrane.

38.4 Conclusion

The emission of volatile organic compounds from shuttle tankers does not only cause harmful effects to the environment but also causes huge monetary loss. The use of membrane technology is one of the emerging technologies that can be used for the recovery of volatile organic compounds. This recovery is based on the permeances and separation factor of the gases through the membrane. Previous studies have shown that membrane material used for gas separation affects the separation factor of that membrane. Zeolites in this work have proved to be a good choice of membrane material for the possible use on offshore and offshore storage facilities. Further work needs to be done for the synthesis of a defect-free membrane that is reproducible and can be introduced into the petroleum and gas industries for the separation of lower hydrocarbons at a competitive cost.

Acknowledgment The Authors wish to acknowledge the center for Process Integration and Membrane Technology at RGU for providing the research infrastructure and the School of Pharmacy Life Science for the SEM/EDAX analysis.

References

1. Tamaddonia M et al (2014) Experimental study of the VOC emitted from crude oil tankers. Process Saf Environ Prot 92(6):929–937
2. Huglen O, Vik R (2001a) Reducing VOC emissions from large crude carriers. In: Offshore Technology Conference
3. Huglen O, Vik R (2001b) Reducing VOC emissions from large crude carriers. In: Offshore Technology Conference, Houston, Texas
4. Howard R, Nikolas H (2001) Measures to reduce emissions of VOCs during loading and unloading of ships in the EU. AEA Technology Environment
5. IPIECA (2013) VOC recovery systems. London: IPIECA. http://www.ipieca.org/energyefficiency/solutions/60501. Accessed 4 July 2015
6. Robert JD, Robert FZ (2002) Cryogenic condensation: a cost-effective technology for controlling VOC emissions. Environ Prog Sustain Energy 21(2):111–115
7. Gupta VK, Verma N (2002) Removal of volatile organic compounds by cryogenic condensation followed by adsorption. Chem Eng Sci 57(14):2679–2696
8. Dautzenberg F, Mukherjee M (2001) Process intensification using multifunctional reactors. Chem Eng Sci 56(2):251–267
9. Bernardo P, Drioli E, Golemme G (2009) Membrane gas separation: a review/state of the art. Ind Eng Chem Res 48(10):4638–4663
10. Ahmed MJ, Theydan SK (2014) Modeling of propane separation from light hydrocarbons by adsorption on 4A molecular sieve zeolite. J Nat Gas Sci Eng 18:1–6
11. Baker RW (2002) Future directions of membrane gas separation technology. Ind Eng Chem Res 41(6):1393–1411
12. Anderson M, Wang H, Lin Y (2012) Inorganic membranes for carbon dioxide and nitrogen separation. 28(2–3):101–121
13. IMPEL (2000) Diffuse Voc Emissions. IMPEL Network
14. Basile A (2013) Handbook of membrane reactors: fundamental materials science. Elsevier, Design and Optimisation

15. Freeman B, Yampolskii Y, Pinnau I (2006) Materials science of membranes for gas and vapor separation. Wiley
16. Abedini R, Nezhadmoghadam A (2010) Application of Membrane in gas separation process: its suitability and mechanisms. Pet Coal 52(2):69–80
17. Ajayi A, Shehu H, Gobina E (2015) Recovery of VOC from onshore and offshore shuttle tankers using structured membranes. In: Lecture notes in engineering and computer science: proceedings of the world congress on engineering and computer science 2015, 21–23 October, 2015. San Francisco, USA, pp 637–642
18. Kusakabe K, Kuroda T, Murata A, Morooka S (1997) Formation of a Y-type zeolite membrane on a porous α-alumina tube for gas separation. Ind Eng Chem Res 36(3):649–655
19. Tirouni I, Sadeghi M, Pakizeh M (2015) Separation of C 3 H 8 and C 2 H 6 from CH 4 in polyurethane–zeolite 4 Å and ZSM-5 mixed matrix membranes. Sep Purif Technol 141:394–402
20. Algieri C, Bernardo P, Barbieri G, Drioli E (2009) A novel seeding procedure for preparing tubular NaY zeolite membranes. Microporous Mesoporous Mater 119(1):129–136

Chapter 39
Effect of Operating Pressure on Fischer-Tropsch Synthesis Kinetics over Titania-Supported Cobalt Catalyst

Kalala Jalama

39.1 Introduction

Fischer-Tropsch synthesis is a reaction that converts carbon monoxide in presence of hydrogen on a transition metal catalyst to form a mixture of hydrocarbons of different chain lengths. In addition to the main products, i.e. paraffins and olefins, some side products such as branched compounds and oxygenates (alcohols, aldehydes, ketones and carboxylic acids) are also formed [1]. Oxygenates are dissolved in water as the latter is also formed in large amounts during FT reaction. Only a small proportion of oxygen that enters the process as CO is converted to oxygenate; most of it leaves as water. Liquid hydrocarbons, i.e. those with five carbon atoms and above, referred to as C_{5+}, are of most economic interest. Therefore, optimizing the reaction product distribution is crucial. It is influenced by operating conditions such as temperature, pressure, etc., the type of reactor and the catalyst used [1]. The influence of these parameters have been reported in literature [2–14]. The increase in operating pressure for FT reaction over cobalt-based catalysts has been reported to have an insignificant effect [2, 14] or to increase the reaction rate and C_{5+} selectivity [5, 7, 8, 10–12]. The effect seems to be influenced by the type of support used. As part of an investigation that is on-going in our laboratory on the effect of total pressure on cobalt catalysts with various support materials, this study aims at reporting the findings for Co/TiO_2.

K. Jalama (✉)
Department of Chemical Engineering, University of Johannesburg, Doornfontein, P.O. Box 17011, Johannesburg 2028, South Africa
e-mail: kjalama@uj.ac.za

S.-I. Ao et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-981-10-2717-8_39

39.2 Experimental Procedure

39.2.1 Catalyst Preparation

The catalyst was prepared by incipient wetness impregnation. Firstly, the support was prepared by mixing TiO_2 powder (P25 Degussa) with deionized water, drying at 120 °C for 1 h followed by calcination in air at 400 °C for 16 h. The prepared support was subsequently impregnated with a solution of cobalt nitrate, dried at 120 °C overnight and calcined in air at 400 °C for 6 h.

39.2.2 Catalyst Evaluation

Some fresh catalyst (0.5 g) was loaded in a fixed-bed reactor and reduced around atmospheric pressure in 5 % H_2/Ar (30 Nml/min) at 350 °C for 14 h. The FT reaction was commenced, after catalyst reduction and reactor cooling below 100 °C, by introducing a pre-mixed synthesis gas (10 ml/min, H_2/CO ratio of 2/1, with 10 % N_2 as internal standard). The reaction was first performed at 1 bar and 220 °C and the pressure was further increased to 10 and 20 bar respectively while all other operating conditions (temperature, feed flow rate and feed gas composition) were kept unchanged. The details on product analysis have been reported in an earlier study [15].

39.3 Results and Discussion

The effect of operating pressure on CO conversion and CH_4 selectivity is summarized in Fig. 39.1. CO conversion of ca. 22 % was measured at 1 bar. This values increased to ca. 28 % at 10 bar and decreased to ca. 19 % upon further

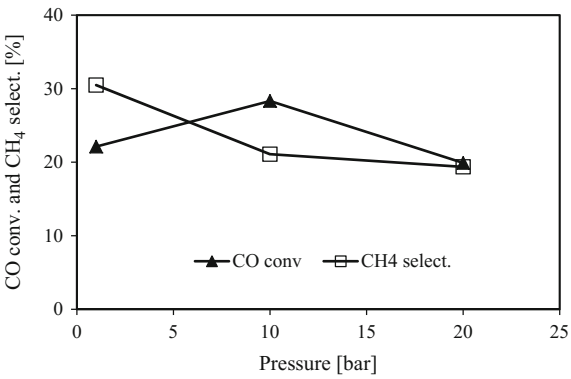


Fig. 39.1 CO conversion and CH_4 selectivity as function of the operating pressure [16]

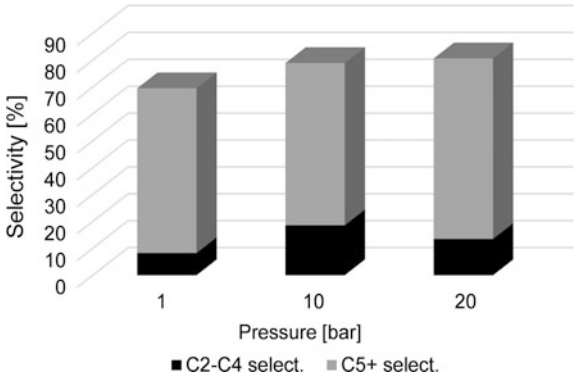


Fig. 39.2 C_2–C_4 and C_{5+} selectivities as function of the operating pressure [16]

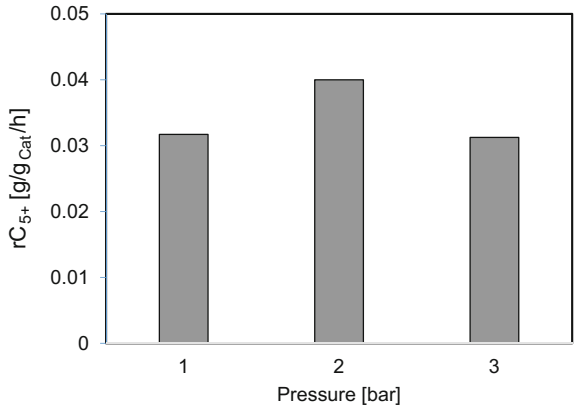


Fig. 39.3 C_{5+} production rate as function of the operating pressure [16]

increase of the operating pressure to 20 bar. The methane selectivity decreased from ca. 30 % at 1 bar to ca. 21 and 19 % when the operating pressure was increased to 10 and 20 bar respectively.

C_2–C_4 and C_{5+} selectivities as function of the operating pressure are reported in Fig. 39.2. The C_2–C_4 selectivity increased from ca. 8.3 % at 1 bar to ca. 18.6 % at 10 bar and decreased to ca. 13.5 at 20 bar. The C_{5+} selectivity did not significantly change as it remained around 60 % when the operating pressure was increased from 1 to 10 bar and increased to the highest value of ca. 67 % when the pressure was further increased to 20 bar. These data and those in Fig. 39.1 indicate that at higher operating pressure, for example 20 bar in this study, the methane selectivity is minimized and the selectivity to C_{5+} is high. However, when the rate of C_{5+} hydrocarbons formation is plotted as function of the operating pressure (Fig. 39.3), an operating pressure of 10 bar yields the highest productivity.

0.032 g of $C_{5+}/g_{Cat}/h$ were produced at 1 bar, increased to ca. 0.040 $g/g_{Cat}/h$ at 10 bar and decreased to ca. 0.031 $g/g_{Cat}/h$ at 20 bar (Fig. 39.3). Even though an operating pressure of 20 bar led to better CH_4 and C_{5+} selectivity, it had the lowest

rate of C_{5+} production. This is due to a lower CO conversion measured at 20 bar compared to 10 bar as shown in Fig. 39.1.

Zennaro et al. [17] have conducted kinetics study on titania-supported cobalt catalyst under conditions where pore-diffusional mass transfer were avoided (3–5 % CO conversion) and found that their data were best fitted by a simple Langmuir-Hinshelwood rate equation (Eq. 39.1).

$$-r_{CO} = \frac{aP_{H_2}^{0.74}P_{CO}}{(1+bP_{CO})^2} \tag{39.1}$$

where $-r_{CO}$ is the CO turnover frequency [TOF], P_{H2} the hydrogen partial pressure, P_{CO} the CO partial pressure and a, and b temperature-dependent constants.

This model was used in this study to compare the predicted and measured trends of CO rate as a function of the operating pressure. No CO TOF was determined in our study but since our experiments were all performed on a single catalyst sample at different pressures but same temperature, and space velocity, the CO TOF as function of the operating pressure should yield a similar trend as that for CO conversion rate. The predicted CO rates are summarized in Table 39.1.

The model predicts a significant increase in CO conversion rate as the partial pressures of CO and H_2 (consequently the total operating pressure) are increased. This does not totally agree with the CO conversion trend measured in this study (Fig. 39.1). An increase in CO conversion was only observed when the pressure was increased from 1 to 10 bar before decreasing when the pressure was further increased to 20 bar. A similar trend was observed by Feyzi et al. [18] on a 40 % Fe/60 %Ni/40 %Al_2O_3 where a significant decrease in CO conversion was observed when the total operating pressure was increase above 7 bar. To further understand these observations, a short review of some earlier studies that reported the kinetics of cobalt-based catalysts at different operating pressures has been compiled in Tables 39.1, 39.2.

Geerlings et al. [5] found that at low pressure (1 bar) the catalyst was operating in a dry mode while at high pressure some liquid film formed on the catalyst surface. O'Shea et al. [7] explained their results by suggesting that at high pressure, more gas dissolved in the solvent resulting in increment of CO concentration in the liquid phase. This favors the formation of carbonyl species and the increase in cobalt particles surface area through segregation.

Table 39.1 Predicted CO consumption rate as function of the operating pressure [16]

Total P (bar)	P_{CO} (bar)	P_{H2} (bar)	a	b	$-rCO$ (s^{-1})
1	0.3	0.6	0.0011	0.04	0.0002
10	3	6			0.0099
20	6	12			0.0270

The values for a and b are from [17]

Table 39.2 Brief literature review on the effect of the operating pressure on cobalt-based catalysts

Catalyst	Reactor details	Operating conditions	Findings	Reference
3.9 %Mg-21.4 %Co/Diatomaceous earth	Continuous stirred autoclave (1 L)	5–15 bar 220 °C H_2:CO = 2:1 0.017–0.018 L/g_{Cat}/min	Selectivity to C_1 and C_{10+} products remained constant over the considered range of total reactor pressure	[2]
Cobalt foil	Differential	1 and 18 bar 220 °C H_2:CO = 2:1 Residence time: 10 min	Higher TOF and chain growth probability were measured at higher pressure (18 bar)	[5]
10 %Co/SiO_2	Autoclave	20–40 bar 230 °C H_2:CO = 2:1 0.2 L/g_{Cat}/min	Catalyst activity and C_{5+} selectivity increased with the increase in pressure	[7]
15 %Co/SiO_2	Fixed-bed	2–20 bar 230 °C H_2:CO = 2:1 0.033 L/g_{Cat}/min.	CO conv., chain growth probability and olefins content increased and the selectivity for CH_4 and branched products decreased when the pressure was increased	[8]
Model Co/SiO_2	Batch high-pressure cell (0.2 L)	0.33–1 bar 240 °C H_2:CO = 1:1	Reaction rate and the chain growth probability increased with the increase in total pressure	[11]
1 %-15 %Co/Al_2O_3	Continuous stirred tank (1 L)	20–35 bar 230 °C H_2:CO = 2:1 0.008 L/g_{Cat}/min	CO conv., C_{5+} hydrocarbons selectivity and the chain growth probability increased whereas the selectivity toward CH_4 and (C_2–C_4) decreased with an increasing total pressure	[10]

(continued)

Table 39.2 (continued)

Catalyst	Reactor details	Operating conditions	Findings	Reference
0.5 %Ru-16.6 %Co-7.5 %ZrP/SiO_2	Fixed-bed	10–25 bar 230 °C H_2:CO = 2:1 0.05 L/g_{Cat}/min	CO conv. increased at higher pressures. CH_4 selectivity first decreased, then did not significantly change with further pressure increase	[12]
80 %Co-20 %Ni-30 %La_2O_5-1 %Cs	Fixed-bed	1–10 bar 260 °C H_2:CO = 2:1 5400 h^{-1}	In general no significant change in product selectivity and the CO conversion remained in the rage 15–20 % when the operating pressure was increased	[14]

It is also possible that the positive effect of high operating pressure is related to the promoting effect of water at some extent. As it can be seen from Table 39.2, most of the studies that have reported an increase in both CO conversion rate and/or C_{5+} selectivity at high total pressures, used SiO_2 as support [7, 8, 11]. Interestingly, a review of water effects on cobalt-catalyzed FT reaction by Davis and Dalai [19] has shown that the conclusions depend on a number of factors including catalyst properties. Water showed a positive effect on cobalt catalysts supported on SiO_2 compared to other supports such as Al_2O_3 and TiO_2.

The increase in CO conversion observed in our study when the operating pressure was increased from 1 to 10 bar can be attributed to an increase in reactants (H_2 and CO) pressures in the reactor. The decrease in CO conversion as the pressure was further increased to 20 bar could be explained by significant mass transfer limitations due to liquid products on the catalyst surface. Further investigations are on-going in our laboratory to confirm these assumptions. The hypothesis is that, if liquid products are responsible for the decrease in catalyst activity at high pressure, these effect must also be temperature-dependent.

39.4 Conclusion

The total operating pressure has been found to influence the CO conversion and product distribution for FT reaction over a Co/TiO_2 catalyst. As expected, the methane selectivity decreased and the C_{5+} selectivity increased when the operating pressure was increased. However, the CO conversion first increased and went through a maximum at 10 bar where a maximum production rate for C_{5+} hydrocarbons was also measured. The decrease in CO conversion above 10 bar was explained in terms of mass transfer restrictions and is expected to be influenced by the operating temperature.

Acknowledgements Financial support from the National Research Foundation (NRF) is acknowledged.

References

1. Steynberg A, Dry M (2004) Fischer-Tropsch technology. Stud Surf Sci Catal 152:1–700
2. Yates IC, Satterfield CN (1992) Hydrocarbon selectivity from cobalt Fischer-Tropsch catalysts. Energy Fuels 6(3):308–314
3. Dry ME (1996) Practical and theoretical aspects of the catalytic Fischer-Tropsch process. Appl Catal A 138(2):319–344
4. Bukur BB, Lang X, Akgerman A, Feng Z (1997) Effect of process conditions on olefin selectivity during conventional and supercritical Fischer–Tropsch synthesis. Ind Eng Chem Res 36(7):2580–2587

5. Geerlings JJC, Wilson JH, Kramer GJ, Kuipers HPCE, Hoek A, Huisman HM (1999) Fischer-Tropsch technology—from active site to commercial process. Appl Catal A 186 (1–2):27–40
6. Dry ME (2002) The Fischer-Tropsch process: 1950–2000. Catal Today 71(3–4):227–241
7. de la Penã O'Shea VA, Alvarez-Galvan MC, Campos-Martin JM, Fierro JLG (2005) Strong dependence on pressure of the performance of a Co/SiO_2 catalyst in Fischer–Tropsch slurry reactor synthesis. Catal Lett 100(1–2):105–110
8. Zheng S, Liu Y, Li J, Shi B (2007) Deuterium tracer study of pressure effect on product distribution in the cobalt-catalyzed Fischer-Tropsch synthesis. Appl Catal A 330:63–68
9. Farias FEM, Sales FG, Fernandes FAN (2008) Effect of operating conditions and potassium content on Fischer-Tropsch liquid products produced by potassium-promoted iron catalysts. J Nat Gas Chem 17:175–178
10. Sari A, Zamani Y, Taheri SA (2009) Intrinsic kinetics of Fischer-Tropsch reactions over an industrial $Co–Ru/\gamma\text{-}Al_2O_3$ catalyst in slurry phase reactor. Fuel Process Technol 90(10):1305–1313
11. Yan Z, Wang Z, Bukur DB, Goodman DW (2009) Fischer-Tropsch synthesis on a model Co/SiO_2 catalyst. J Catal 268(2):196–200
12. Kwack S-H, Park M-J, Bae JW, Ha K-S, Jun K-W (2011) Development of a kinetic model of the Fischer-Tropsch synthesis reaction with a cobalt-based catalyst. Reac Kinet Mech Cat 104(2): 483–502
13. Sauciuc A, Abosteif Z, Weber G, Rauch R, Hofbauer H, Schaub G, Dumitrescu L (2012) Influence of operating conditions on biomass-based Fischer-Tropsch synthesis. Biomass Conv Bioref 2(3):253–263
14. Mirzaei AA, Shiradi B, Atashi H, Mansouri M (2012) Modeling and operating conditions optimization of Fischer-Tropsch synthesis in a fixed-bed reactor. J Ind Eng Chem 18(4): 1515–1521
15. Jalama K (2016) Fischer-Tropsch synthesis over Co/TiO_2: effect of catalyst activation by CO compared to H2. Catal Commun 74:71–74
16. Jalama K (2015) Fischer-Tropsch synthesis over titania-supported cobalt catalyst: optimum pressure for C_{5+} hydrocarbons production. Lecture notes in engineering and computer science: proceedings of the world congress on engineering and computer science 2015, 21–23 October, 2015, San Francisco, USA, pp 643–645
17. Zennaro R, Tagliabue M, Bartholomew CH (2000) Kinetics of Fischer-Tropsch synthesis on titania-supported cobalt. Catal Today 58(4):309–319
18. Feyzi M, Mizraei AA, Bozorgzadeh HR (2010) Effects of preparation and operation conditions on precipitated iron nickel catalysts for Fischer-Tropsch synthesis. J Nat Gas Chem 19:341–353
19. Dalai AK, Davis BH (2008) Fischer-Tropsch synthesis: a review of water effects on the performances of unsupported and supported Co catalysts. Appl Catal A 348(1):1–15

Chapter 40
Characterization of Gamma-Alumina Ceramic Membrane

Ifeyinwa Orakwe, Ngozi Nwogu and Edward Gobina

40.1 Introduction

Membrane technology has drawn so much attention over the past 50 years with the initial aim of employing membranes for the separation of certain gases as well as for large scale gas diffusion for uranium isotope separations [1–3]. Alumina ceramic membranes in particular have drawn the attention of researchers as a result of its unique properties for example showing high chemical resistance and mechanical strength and are also known to withstand high temperature and pressure [4].

The term "ceramic membrane" could be seen as a semi permeable barrier consisting of metallic oxides which could be porous or dense in structure [5]. These membranes are usually asymmetric in structure, consisting of a few layers of one or more of different materials with graduated pore sizes. These ceramic membranes could be grouped based on their pore sizes in to microporous, mesoporous, macroporous or dense supports [6]. The microporous supports have pore sizes less than 2 nm and are suitable for micropore diffusion mechanisms. The mesoporous supports go in a range of pore sizes between 2–50 nm and are known to play significant roles in Knudsen diffusion mechanism. The macroporous supports have pore sizes greater than 50 nm and are applicable to sieving mechanism while the dense membranes are best known for mechanisms involving diffusion [5].

I. Orakwe · N. Nwogu · E. Gobina (✉)
Centre for Process Integration and Membrane Technology (CPIMT),
School of Engineering, The Robert Gordon University Aberdeen,
Aberdeen AB10 7GJ, UK
e-mail: e.gobina@rgu.ac.uk

I. Orakwe
e-mail: i.r.orakwe@rgu.ac.uk

N. Nwogu
e-mail: n.c.nwogu@rgu.ac.uk

S.-I. Ao et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-981-10-2717-8_40

The membrane makeup constitutes either or a combination of the following: titanium oxide (TiO_2), zirconium oxide (ZrO_2), silicon oxide (SiO_2) or aluminium oxide (Al_2O_3) [4–6].

In this study, an alumina ceramic membrane was chosen. The support is made up of ceramic material composed of a gamma-alumina and a wash coat of titania. The titania creates some surface roughness responsible for enhancing adhesion and dispersion of catalyst particles if required for further modification or experimental analysis. The gamma alumina ceramic membrane is classified under the mesoporous type since its pore size was 3.3 nm. To mention but a few advantages are its chemical stability, narrow pore size distribution, mechanical strength and durability [6].

This paper presents the characterization of the gamma alumina ceramic membrane employing three techniques: Scanning electron microscopy, nitrogen adsorption desorption and gas permeation at varying temperatures.

The scanning electron microscopy (SEM) is an important parameter used to study the morphology of a membrane. This is commonly achieved through the use of a scanning electron imaging at different magnifications at nano scale levels [7]. It provides information on the nature of layers present, the surface smoothness, as well as thickness, and also the nature of grain size or shape [4]. The SEM works by beaming electrons and not light on a sample to form an image. It is a qualitative means of analysis for characterisation methods [5]. It also has the ability of detecting defects, cracks or structural deformation on a membrane. It can analyse a membrane sample from an outer, inner, edge and cross sectional point of view [7].

The nitrogen adsorption desorption technique is one of the most common techniques used to obtain parameters such as the pore size, pore size distribution and specific surface area for a porous membrane [4]. The Barrett, Joyner and Halenda (BJH) method is used for calculating the pore size and pore size distribution, while the Brunauer, Emmett and Teller (BET) method calculates the surface area by measuring how much of the gas is adsorbed per gram of a material [7]. There are different types of isotherms that could occur during a gas adsorption and desorption process as discussed by Li [5]. He also grouped them from type I–type VI. Type I is dominant with microporous solids, type II is observed with non-porous solids, types III and V are common for vapour adsorption, types IV and V are common with mesoporous solids, in which a hysteresis loop caused by the desorption occurring at a lower pressure than the adsorption can be observed. This is usually as a result of capillary condensation. Type VI is uncommon and is seen in special types of carbon materials [5].

Gas permeation is a process whereby a flow meter is used to measure the gases flowing through the porous or dense membrane in a membrane permeator at different operating pressures. The permeation measurement methods are the most reliable method for proposing a membranes' performance [4]. The process involves the use of either single or binary gas mixtures or even liquids and a membrane holder.

The objective of this report is to use the aforementioned characterization techniques to obtain vital information regarding the features of the commercial alumina membrane in order to enable further work on the possible application of the membranes in chemical reactors, such as in the impregnation of platinum metal into the mesoporous pores of the membrane for the purpose of removal of dissolved oxygen for down-hole water injection and process application.

40.2 Experimental

40.2.1 Materials

Figure 40.1 shows a photograph of the commercial gamma alumina ceramic membrane selected for this experiment (Ceramiques Techniques et Industrielles (CTI SA) France). It consists of 77 % alumina + 23 % TiO_2. The membrane is made up of 7 mm inner diameter and 10 mm outer diameter respectively with a permeable length of 318 mm and 45 % porosity. Single and mixed gases: nitrogen (N_2), helium (He), oxygen (O_2), carbon dioxide (CO_2), 14 %CO_2/N_2, 60 %CO_2/N_2, 30 %CO_2/CH_4 and 21 %O_2/N_2 were used in the permeation tests.

40.2.2 Characterization Techniques

40.2.2.1 Scanning Electron Microscope

Scanning electron microscope (Zeiss EVO LS10 Variable Pressure) was used to observe the structure of the alumina membrane. The alumina membrane sample was mounted on a stub of an aluminium disc using a suitable adhesive. The adhesive used was able to anchor the sample to the stub as well as provide electrical continuity. The sample carousel is then loaded with the samples to be examined. After visually examining the sample, it was then positioned on a metal holder and gold-coated using sputter-coating operated under vacuum. The SEM micrographs of the inner and outer areas were then taken. The elemental composition of the membrane was also obtained by the energy dispersive analysis x-ray (EDAX) which was coupled with the electron microscopy. This gives the entire composition of the sample [8].

40.2.2.2 Nitrogen Adsorption-Desorption

Nitrogen adsorption-desorption (Quantachrome®ASiQwin™) instrument was used for the surface area and pore size distribution study of the membrane. A fragment of the γ—alumina membrane was crushed into small pieces and put into a cell and weighed and then degassed for about 4 h at 338 K. This was done to remove any moisture and impurities trapped in the sample. The analysis of the degassed sample was then carried out overnight at a temperature of 77 K. The nitrogen adsorption

Fig. 40.1 Tubular ceramic membrane [8]

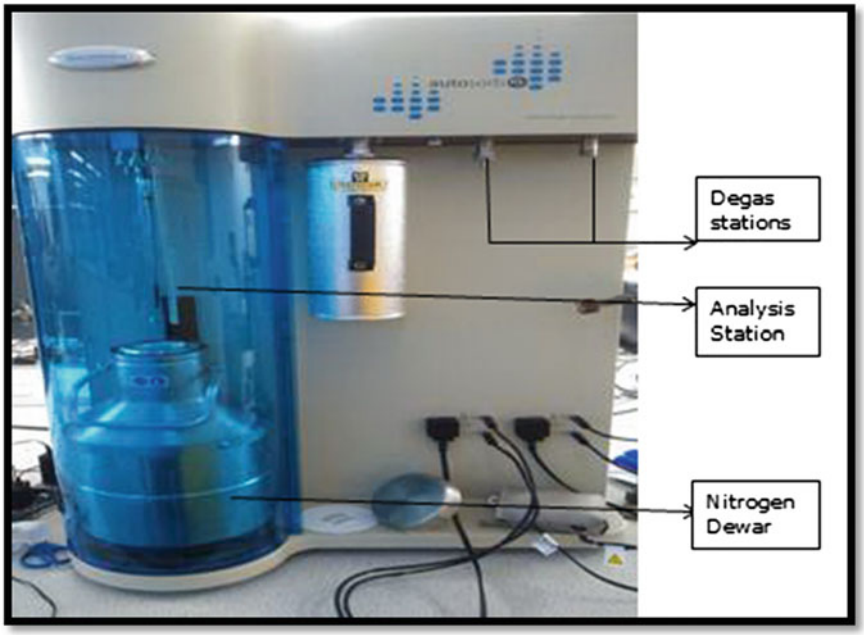


Fig. 40.2 Nitrogen adsorption desorption instrument (Quantachrome® ASiQwin™)

desorption isotherms in the case of a non-reactive gas is based on the amount of gas adsorbed as a function of the relative pressure P/P_o, where P is the applied pressure and P_o is the saturated pressure [4]. Figure 40.2 shows the pictorial diagram of the nitrogen adsorption desorption instrument used [8].

40.2.2.3 Gas Permeation Test

Gas permeation tests were carried out using single and mixed gases: nitrogen (N_2), oxygen (O_2), carbon dioxide (CO_2), 14 %CO_2/N_2, 60 %CO_2/N_2, 30 %CO_2/CH_4 and 21 %O_2/N_2 respectively in the temperature range of 25–200 °C. The experiment was carried out by regulating the flow rate of the gases using a control valve at different transmembrane pressure drop in the range of 0.05-1 bar across the membrane. The permeate flow rate was recorded using a flow meter (Cole Parmer model). Figure 40.3 shows a schematic diagram of the gas permeation setup [8].

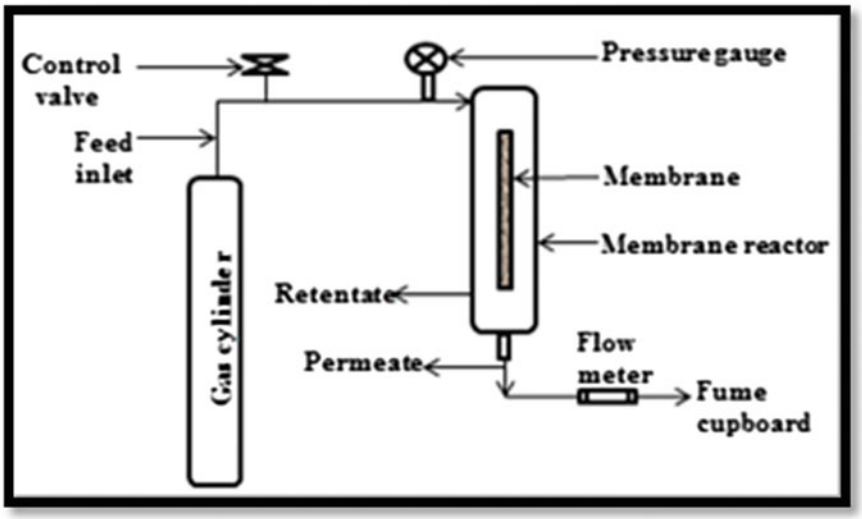


Fig. 40.3 Schematic diagram of the gas permeation setup [8]

40.3 Results and Discussion

40.3.1 SEM Characterization

A surface and structural morphology of the gamma alumina membrane has been carried out and the micrographs are shown in Fig. 40.4a, b respectively. It was also observed that the membrane does not have a smooth surface and particles were packed in a way to obtain a porous membrane. It is also observed that the particles were of varying shapes and smaller particles were deposited among larger particles and thus the effects of defects are minimised. The result shows this arrangement resulted in a mesoporous membrane and there were no cracks on the surface.

The EDAX elemental spectrum is shown in Fig. 40.5. It shows all the elements present in the membrane including zirconium, titanium, aluminium, carbon and oxygen.

40.3.2 Nitrogen Adsorption-Desorption Characterization

The nitrogen adsorption desorption instrument generates physiosorption isotherms as well as calculates and provides vital information for the surface areas, pore diameter, pore volume and pore size distribution. Figure 40.6 shows the isotherm obtained and it corresponds to the type IV isotherm [7], which is characteristic of a mesoporous membrane undergoing capillary condensation.

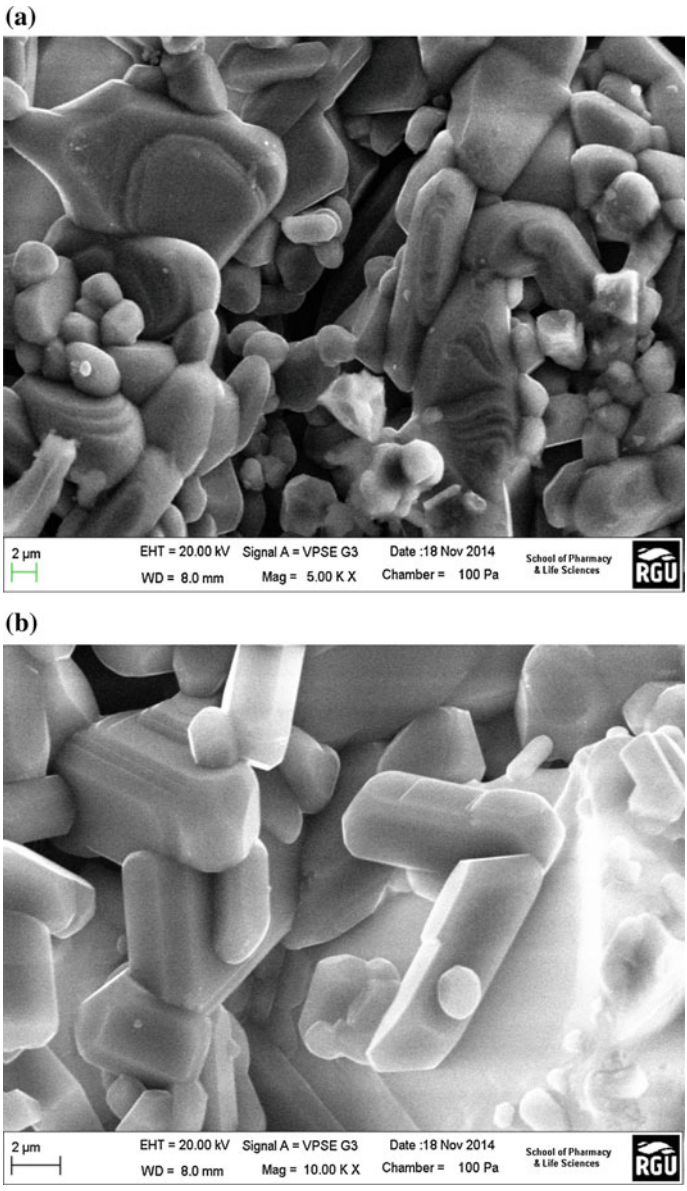


Fig. 40.4 **a** SEM micrograph of the outer section of the alumina membrane, **b** SEM micrograph of the inner section of the alumina membrane

From the isotherm in Fig. 40.6, a nitrogen desorption isotherm was generated using the Barrett, Joyner and Halenda (BJH) method as shown in Fig. 40.7. The calculated the surface area of the alumina membrane was 0.192 m^2/g, a pore volume of 0.001 cc/g and a pore diameter of 3.305 nm.

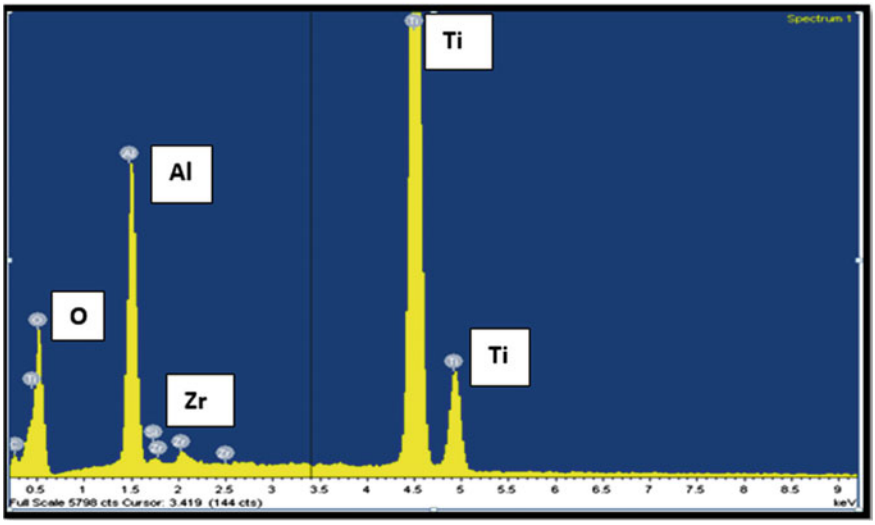


Fig. 40.5 EDAX spectrum of the elemental composition of the alumina membrane

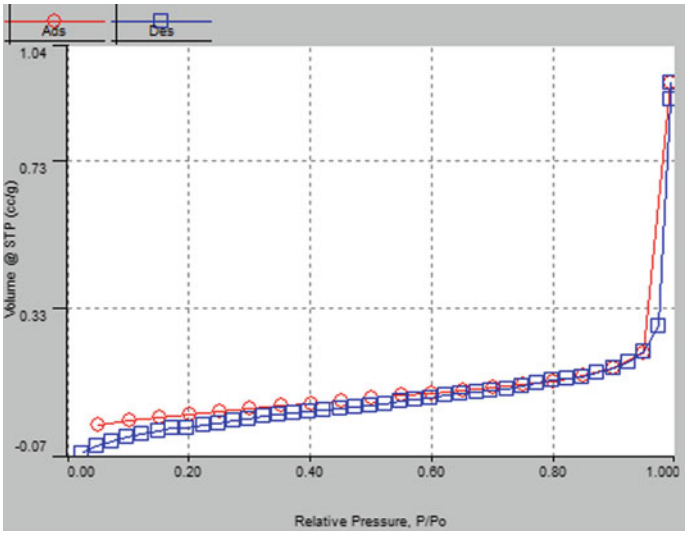


Fig. 40.6 A physisorption isotherm of the alumina membrane

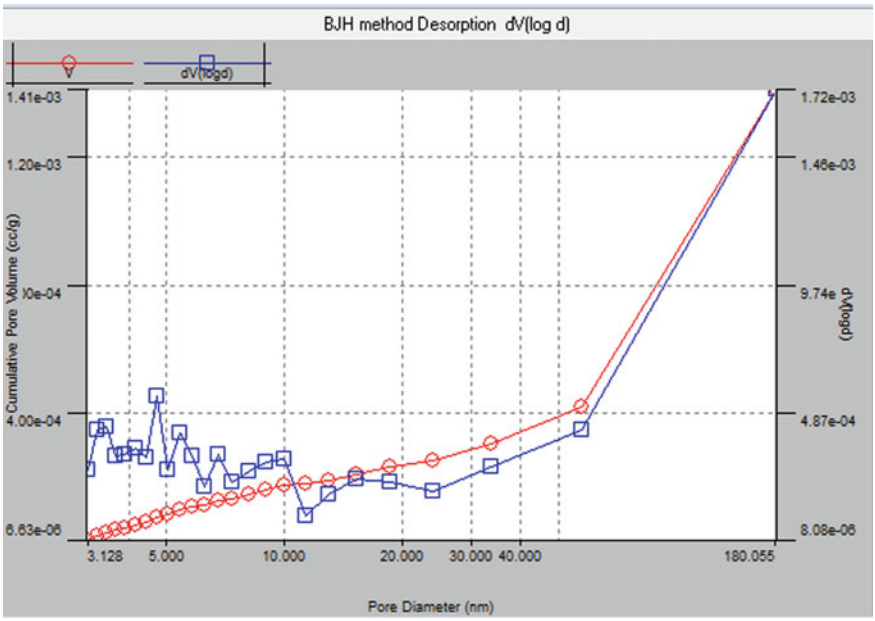


Fig. 40.7 BJH desorption Isotherm showing the pore size distribution of the alumina membrane

40.3.3 Gas Permeation Analysis

Figure 40.8a represents a plot of permeate flow rate of single gases at 0.5 bar transmembrane pressure drop against various temperatures. It was noted from the graph that the different gases behaved differently at various temperatures. O_2 showed a monotonic decrease in flow rate as the temperature increased, while nitrogen and carbon dioxide showed an initial drop in flowrate, then an increase and finally a decrease in flowrate as the temperature was increased. The gas mechanism present is highly dominated by Knudsen diffusion since the gases with lower molecular weight tended to have a higher flow rate than those with higher molecular weight. This is experienced when the molecule—pore wall collisions are greater than the molecule–molecule collisions.

Figure 40.8b represents a plot of flowrate of mixed gases at 0.5 bar pressure against temperature. The slope obtained indicated the presence of initial viscous flow, and then Knudsen flow as the temperature was increased. It was also noted that the different gases behaved differently at various temperatures.

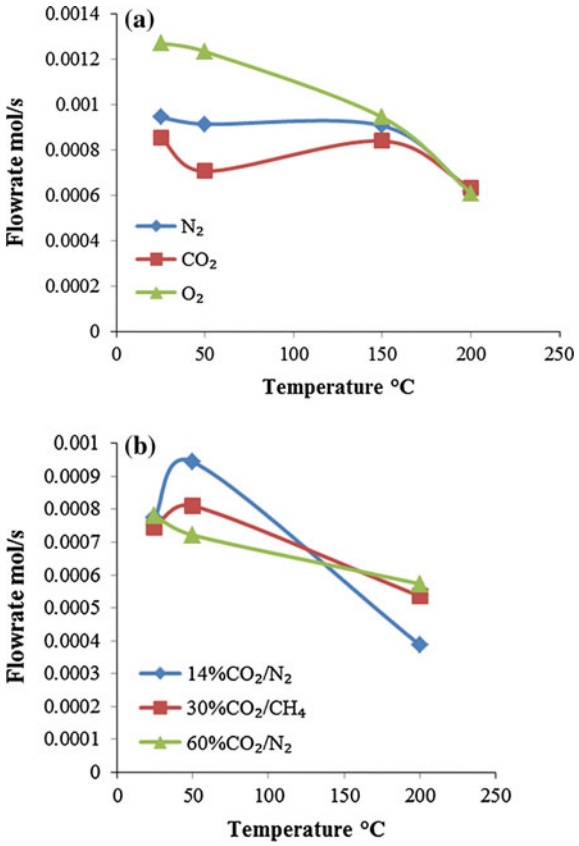


Fig. 40.8 **a** Graph of flowrates of single gases against temperatures, **b** graph of flow rates of binary gases against temperatures

40.4 Conclusion and Future Work

A characterisation study has been carried out on a commercially available ceramic membrane. It was observed from the SEM micrographs that the membrane was asymmetric with a pore size distribution. The nitrogen adsorption desorption isotherm showed an isotherm which corresponded to the characteristics of a mesoporous membrane. A mean pore diameter of 3.305 nm and surface area of 0.192 m^2/g were obtained following the BJH isotherm of the nitrogen adsorption desorption. It was also noted that the flowrate of the gases tend to decrease with increasing temperature and the lighter gases having a higher flowrate than the heavier gases. These characteristics are consistent with gas transport dominated by Knudsen flow.

Further work will involve catalyst impregnation into the pores of the membrane. This will be done to enhance hydrogen gas selectivity. Platinum and palladium catalysts are known to enhance hydrogen transport and the membrane pores will enable a wider distribution of the metallic particles.

Acknowledgements The authors are grateful to the staff of center for process integration and membrane technology and school of Pharmacy and Life sciences, Robert Gordon University for their support in some of the result analysis. The financial support of CCEMC grant is also greatly appreciated.

References

1. Lin YS (2001) Microporous and dense inorganic membranes: current status and prospective. Elsevier Seper Purif Technol 25:39–55
2. Alzahrani S, Mohammed AW (2014) "Challenges and trends in membrane technology implementation for produced water treatment". A review. J Water Process Eng 4:107–133
3. Julbe A, Farrusseng D, Guizard C (2001) Porous ceramic membranes for catalytic reactors—overview and new ideas. J Membr Sci 181(1):3–20
4. Smart S, Liu S, Serra JM, Lulianelli A, Basile A (2013) 8—porous ceramic membranes for membrane reactors. In: Basile A (ed) Handbook of membrane reactors. Woodhead Publishing, pp 298–336
5. Li K (2007) Ceramic membranes for separation and reaction. Wiley
6. Benito JM, Coesa A, Rubio F, Rodriguez MA (2005) Preparation and characterization of tubular ceramic membranes for treatment of oil emulsions. J Eur Ceram Soc 25(11):1895–1903
7. Othman M, Mukhtar H, Ahmad A (2004) Gas permeation characteristics across nano-porous inorganic membranes. IIUM Eng J 5(2)
8. Orakwe I, Nwogu NC, Gobina E (2015) Morphological characterization and gas permeation of commercially available ceramic membrane. Lecture notes in engineering and computer science: proceedings of the world congress on engineering and computer science 2015, 21–23 October, 2015, San Francisco, USA, pp 607–610

Chapter 41
Investigation of Flue Gas and Natural Gas Separation Using Silica Composite Membranes Formed on Porous Alumina Support

Ngozi Nwogu, Ifeyinwa Orakwe and Edward Gobina

Nomenclatures

Al	Aluminum
C	Carbon
Cl	Chlorine
O	Oxide
Si	Silicon
Ti	Titanium
Zr	Zirconium
M_{CO2}	Molecular weight of CO_2
M_{ti}	Molecular weight of either N_2 or CH_4
α	Separation factor

41.1 Introduction

Separation processes involving CO_2 are currently well-recognized especially with the growing concern about global warming placing greater demands on improving energy efficiency and reducing CO_2 emissions. Hence, energy sustainable CO_2 separation technologies with value-added economics are needed for industrial processing and for future to recover CO_2 from natural gas, fossil-fuel combustion, energy generation through power supply and effluents from industrial processes.

N. Nwogu · I. Orakwe · E. Gobina (✉)
Centre for Process Integration and Membrane Technology (CPIMT), School of Engineering, The Robert Gordon University, Aberdeen AB10 7GJ, UK
e-mail: e.gobina@rgu.ac.uk

N. Nwogu
e-mail: n.c.nwogu@rgu.ac.uk

I. Orakwe
e-mail: n.c.nwogu@rgu.ac.uk

S.-I. Ao et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-981-10-2717-8_41

Accordingly, some literature have reported that membrane-based separations are excellent substitute to traditional methods due to their energy efficacy and durability and therefore offers excellent separation with inherent features which are favourable in gas separation [1, 2]. Their usefulness and application as stated are in natural gas purification, separations involving flue gases and volatile organic compounds and hydrogen separation from industrial feed stock [2–4].

Some methods already exist for separating CO_2 from flue gas mixture namely cryogenic distillation, absorption and adsorption. In particular absorption of carbon dioxide through amine based solvent remains the leading process in large scale industrial processes but well known to be expensive. The cryogenic separation of CO_2 is energy intensive due to heating and refrigerating [5]. Nevertheless, membrane has become an emerging technology and holds great potentials for bulk gas separations [6]. Generally, alumina supports are macroporous in nature with large pore sizes greater than 50 nm and provides mechanical strength to the overlying asymmetric layers. Several techniques have been developed to process these supports. They include sol–gel synthesis, leaching [7–9] and chemical vapour deposition [10]. Amongst all, sol–gel processing attracts the most attention due to its excellent processing steps, simplicity and its potential for pore size and pore structure control [11]. Several approaches have been suggested by Brinker et al. [11] for the fundamental physical and chemical phenomena involved in the deposition of colloidal ceramic dispersions (sols) on porous supports for precise pore size and porosity control. Moreover, by choosing and manipulating the textural properties of the porous support, it is conceivable to control the size and shape of the resulting composite structure to improve the flow rate and the separation factor of permeating gas molecule.

In the present study, the problem of CO_2 separation from nitrogen and methane have been addressed through the design of a novel composite silica ceramic membrane separation tool to improve the ideal separation factor and diffusion characteristics at room temperature and relatively low pressure.

41.2 Experimental

A silica sol is prepared by carefully adding a curing agent to a corresponding amount of a mixture of iso-pentane and silicone elastomer. The mixture is kept under room temperature with constant stirring for homogeneity purposes. The reaction mixture had a Iso-pentane/silicone elastomer/curing agent ratio of 900/100/10 (mls) in agreement with the customary formula of silica sol preparation [12, 13]. The tubular shaped commercial support used was supplied by Ceramiques Techniques et Industrielles (CTI SA) France and comprises of (77 % α-alumina + 23 % TiO_2) with an average pore diameter of 6000 nm. The support has 20 mm and 25 mm internal and outer diameter respectively and a permeable length of 319 mm. The composite silica ceramic membranes are prepared by dip coating the alumina support in the silica-based solution followed by spin-drying and

heating. However, a complete dipping process involves repeated dipping to mend any cracking which may have occurred during support manufacture and most importantly to reduce the pore size of the porous alumina to achieve desired results. After 30 min three step successive dipping, the membrane was spin-dried at room temperature for another 1 h before final heating in an oven for 3 h at 65 °C [12, 13]. The coating through the deposition of the solution on the surface of the alumina led to the formation of a composite silica membrane.

The immersion of the support in the silica solution is shown in Fig. 41.1. The support substrate is immersed in the silica solution repeatedly and the gas transport measured after each dip.

The scanning electron microscope (Zeiss EVO LS10) was employed in determining the morphology of the membrane. The SEM recordings were made on the membrane to player section by cutting the membrane in smaller specimens.

Gas transport tests through the silica composite membrane were carried using the permeation and separation system. The apparatus set-up is depicted in Fig. 41.2. The tubular-shaped membranes were placed in stainless-steel permeance cells. At each end, graphite rings were used to hold the membrane and create a seal. The pressure difference across the membrane was measured using pressure gauges. Single gases namely; CO_2, CH_4 and N_2 (with a purity of at least 99.9 %) were used for testing. The pressure at the feed side was varied between 1–2 bar. The flow rates of all inlet and outlet streams were directly measured by digital flow meters.

Thus, for a single membrane tube, the ideal separation factor α, for pure-gas permeation can be defined by the ratio of permeances of individual pure gases. Whereas the Knudsen separation factor can be obtained using the equation below:

$$\alpha_{CO_2 t_i} = \sqrt{\frac{M_{t_i}}{M_{CO_2}}} \tag{41.1}$$

where α is the separation factor, M_{CO2} is the molecular weight of the CO_2 and M_{ti}, the molecular weight of either nitrogen or methane gas component.

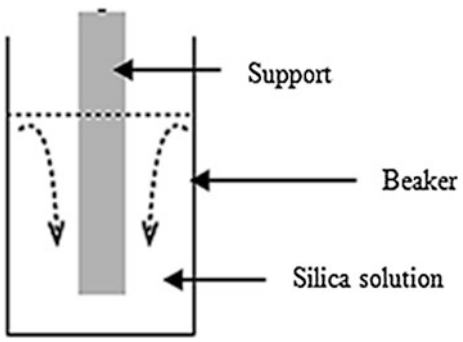


Fig. 41.1 Successive dip-coating of the membrane support

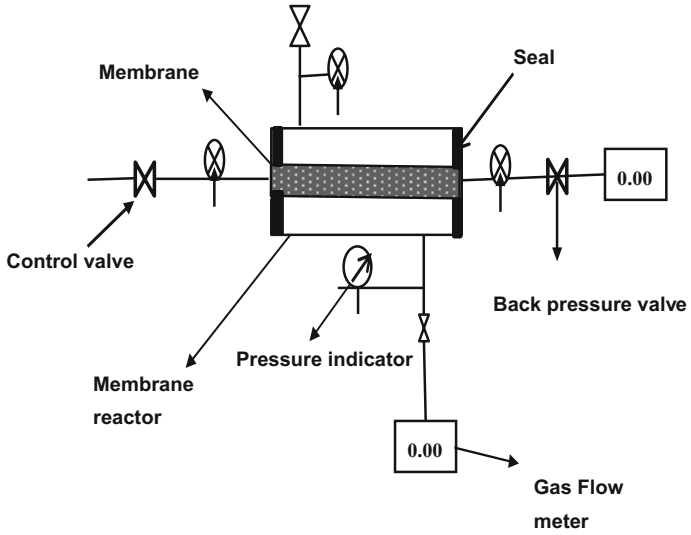


Fig. 41.2 A simplified concept of membrane separation process

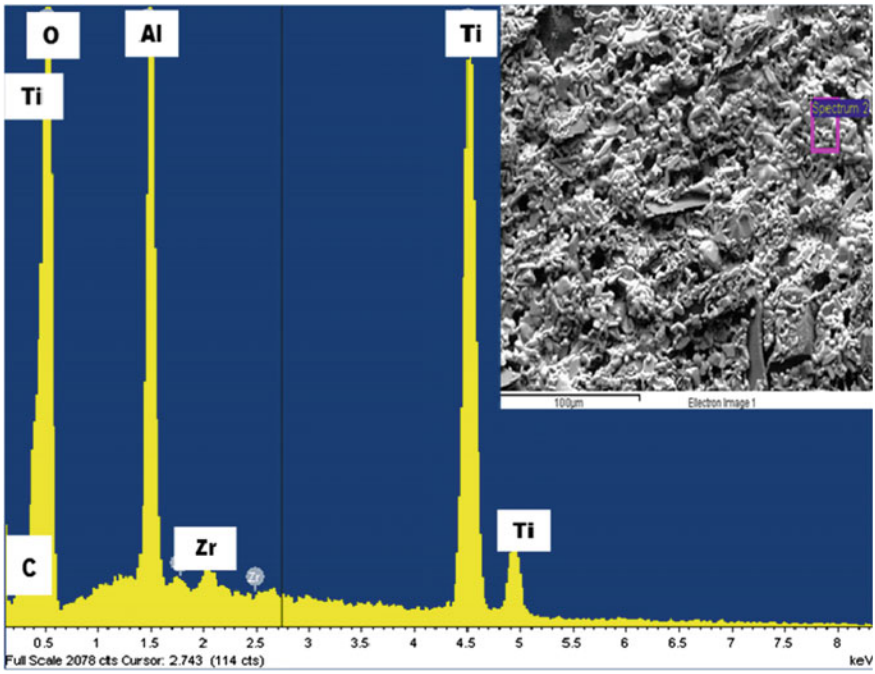


Fig. 41.3 SEM image embedded inside the EDXA micrograph of the ceramic support

41.3 Results and Discussion

The EDXA and scanning electron micrographs obtained are presented in Figs. 41.3, 41.4, 41.5, 41.6. The three step dip-coating reveals an incremental deposition of a thin silica layer obtained after 3 times dipping. The image results indicate that the silica coating was deposited on top of the alumina support layer as a distinct layer. The colour texture difference between each dipping is also very remarkable. The second and third dip has a lighter appearance than the first as a result of the silica. The support and the first dip are also coarse in nature while the second and third dip appears smoother in nature. The SEM images in Figs. 41.3, 41.4, 41.5, 41.6 are embedded in the EDXA micrographs showing clearly the absence of silica in the support and its gradual deposition as the dipping progresses.

Table 41.1 presents the elemental composition by weight of the various elements obtained from the EDXA micrograph for the support and after three successive dipping. Results confirmed that silica was not present in the support but rather C, O, Al, Zr, Ti and Cl whereas, the presence of Si with an increment was seen in the 1st, 2nd and 3rd dipping.

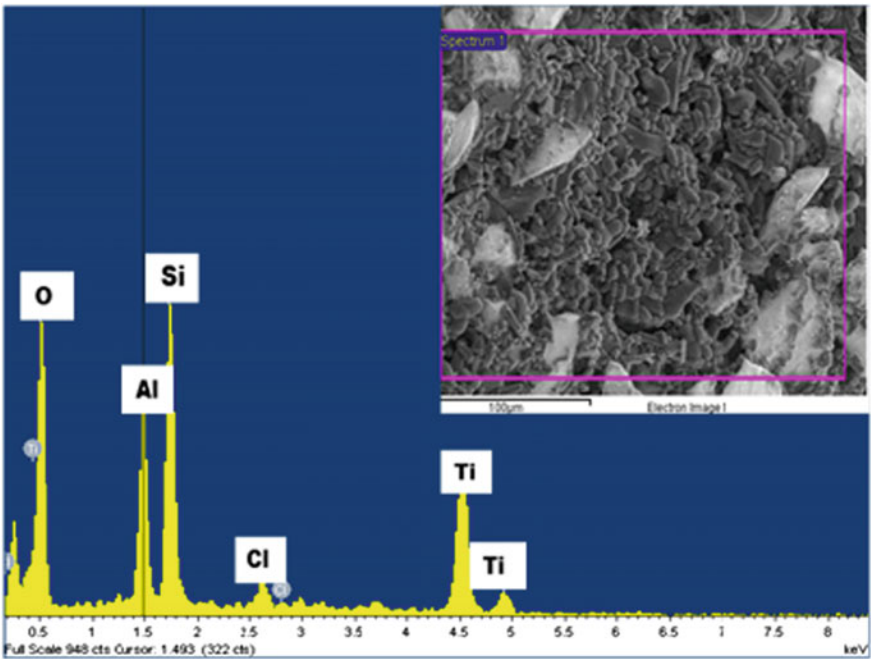


Fig. 41.4 SEM and EDXA micrograph after 1st dipping

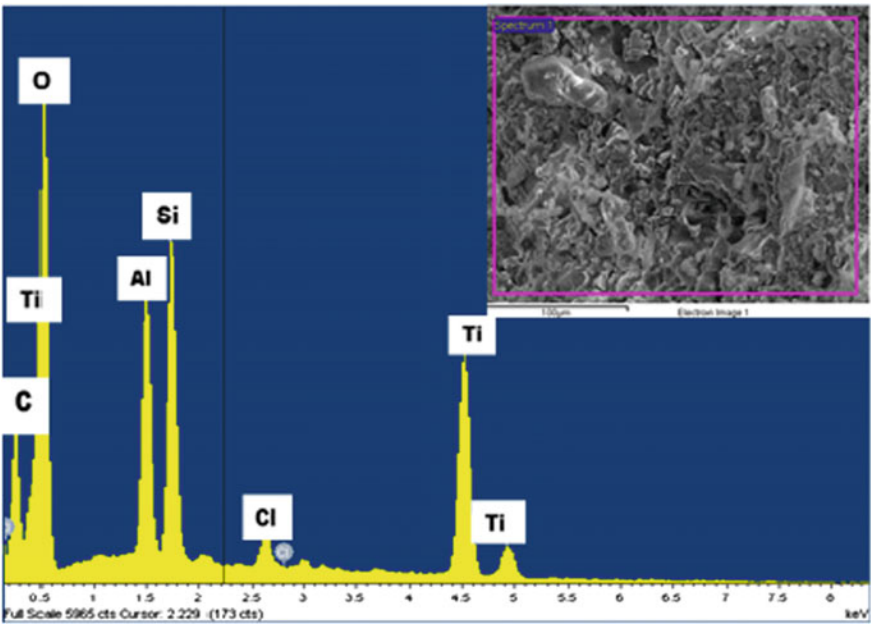


Fig. 41.5 SEM and EDXA micrograph after 2nd dipping

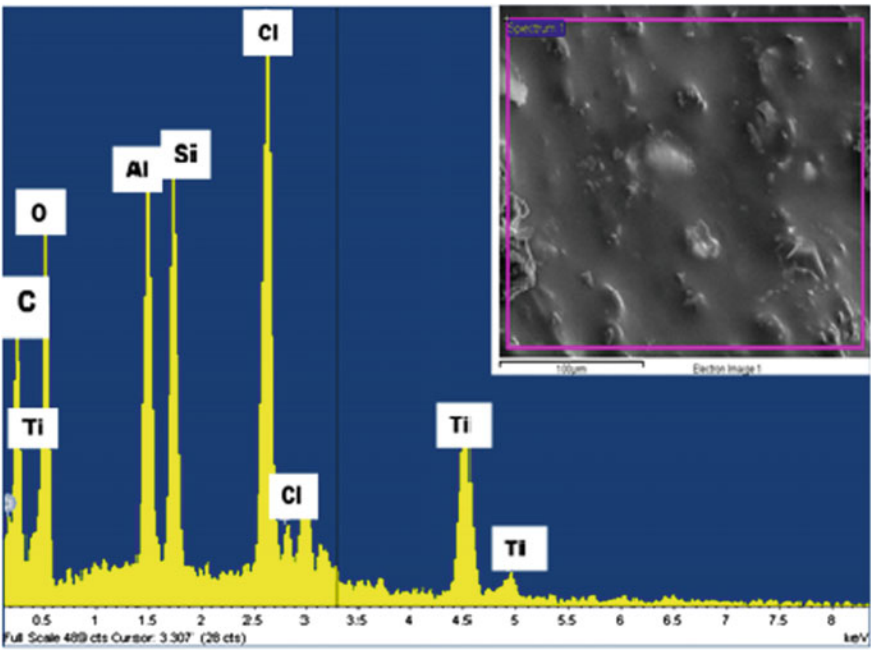


Fig. 41.6 SEM and EDXA micrograph after 3rd dipping

Table 41.1 Elemental percentage composition by weight (100 %)

Elements	Support	1st dipping	2nd dipping	3rd dipping
C	1.32	20.68	3.04	
O	37.74	21.66	38.17	36.88
Al	8.05	6.68	6.21	7.65
Si	Not present	8.05	10.19	13.93
Zr	0.98			
Ti	51.90	21.45	40.50	39.47
Cl		21.48	1.89	2.06

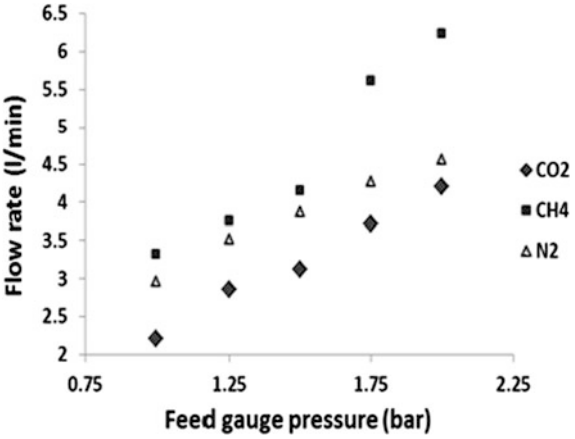


Fig. 41.7 Single gas flow rate measurement as a function of gauge pressure across the support at room temperature

41.4 Gas Permeation

A plot of CO_2, N_2 and CH_4 flow rates on the support and composite silica ceramic membrane as a function of feed gauge pressure is shown in Figs. 41.7 and 41.8 respectively. For the support, an exponential increase in gas flow rates can be observed with increase in feed pressure. CH_4 recorded the highest permeate flow rate followed by N_2 and then CO_2. This characteristic is typical of Knudsen mechanism since it appears to be molecular weight dependent. However for gas diffusion through the composite silica membrane, we observed a clear difference in the flow rate trend. CO_2 permeated faster than CH_4 and N_2. CH_4 maintained an almost constant flow rate. Interestingly, a drastic increase in flow was observed for N_2 leading to a nearly equal flow rate to that of CH_4 at a pressure of about 2 bar. A possible adsorptive type transport mechanism can be suggested on the part of CO_2. Generally, it is a well-established fact that at a relatively low temperature (room temperature), silica materials are flexible with respect to CO_2 adsorption, in which case its separation from gases like N_2 and CH4 which are less absorbable gases can preferentially result to excellent CO_2 recovery [2, 14, 15].

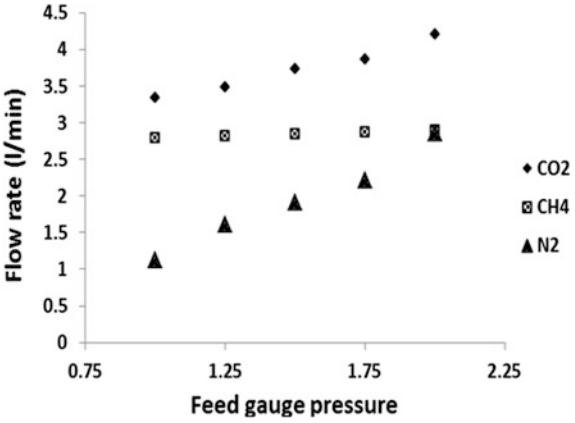


Fig. 41.8 Single gas flow rate measurement as a function of gauge pressure across multi-layer membrane at room temperature

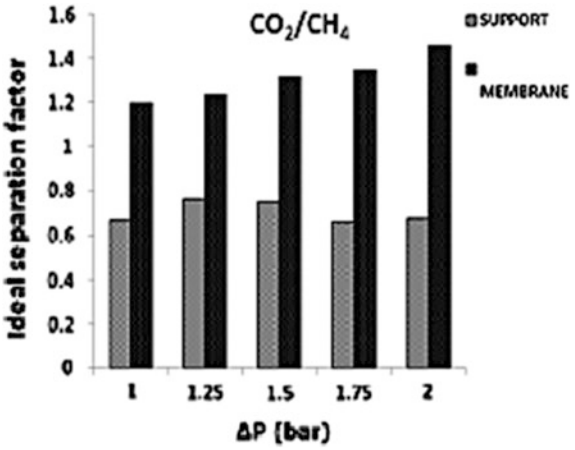


Fig. 41.9 CO_2/CH_4 Ideal separation factor as a function of difference in pressure for both support and multi-layer membrane

The transport mechanism for N_2 and CH_4 could be suggested to be partially Knudsen but with further increase in pressure, molecular- sieving mechanism can come into play since N_2 molecule size is smaller (3.64 Å), than that of CH_4 (3.8 Å).

The ideal separation factor of the various gas pairs (CO_2/CH_4 and CO_2/N_2) for the support and the silica composite membrane were obtained for the purpose of comparison and to determine how much of CO_2 can be separated from the other gases (either N_2 or CO_2). Results obtained show a very clear distinction for both scenarios. A minimal separation was witnessed in the support whereas, a higher separation factor value was obtained for the multi-layered membrane which exceeded the Knudsen separation factor as show in Figs. 41.9 and 41.10. In addition the separation factor for CO_2/N_2 (2.98) was higher than that of CO_2/CH_4 (1.19) in comparison to their Knudsen values as 0.79 and 0.6 respectively. This is

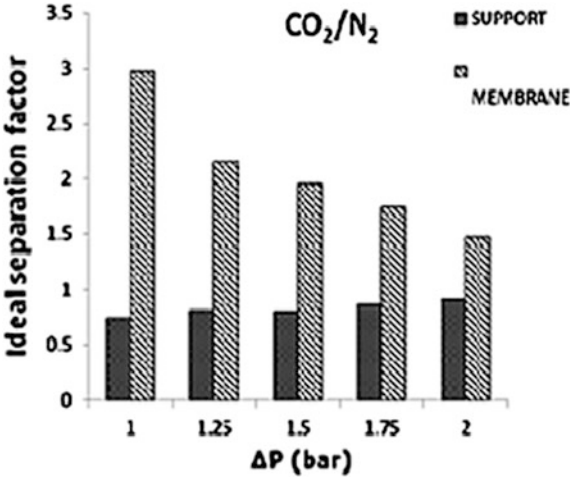


Fig. 41.10 CO_2/N_2 Ideal separation factor as a function of difference in pressure for both support and multi-layer membrane

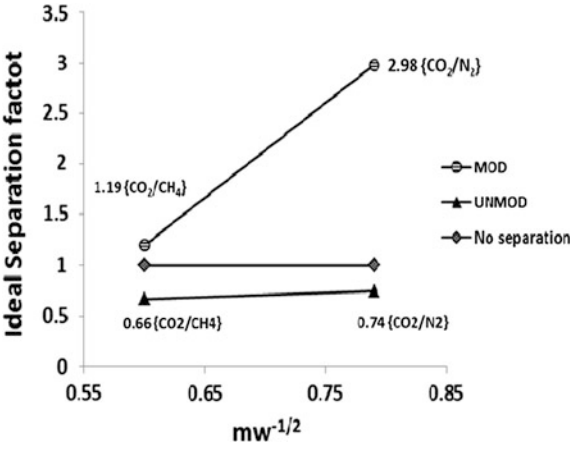


Fig. 41.11 Relationship between ideal separation factor and Knudsen for support (unmod) and silica composite membrane (mod)

presented in Fig. 41.11. Therefore, the experimental separation factor was higher than that of the ideal Knudsen, an indication of a very encouraging result which can be functionally applicable at industrial scale.

41.5 Conclusion

A silica composite membrane have been designed to enhance separation, permeation and reproducibility characteristics through a dip coating process by the deposition of a silica-based solution on the surface of a porous alumina support under a neat and dust-free environment at a temperature of 65 °c for 3 h. This was done to help reduce the large pores of the support. The membrane designed

improved gas diffusion and the separation factor especially for CO_2. The membrane favoured CO_2 with very high flow rate 4.2 l/min compared to 2.89 and 2.86 (l/min) for CH_4 and N_2 respectively. In additional, the membrane selectively separated CO_2 from N_2 and CH_4 having a separation factor of 1.19 and 2.98 higher that Knudsen separation of 0.79 and 0.6 respectively and therefore can find great application in flue gas/natural gas purification and separation processes for reduction of anthropogenic CO_2 as a known greenhouse gas. Additionally, the combined approaches are proved to be effective for achieving molecular sieving in inorganic silica membranes.

Acknowledgement The author wishes to express sincere thanks to the Centre for Process Integration and Membrane Technology of Robert Gordon University for procuring the fresh membrane and Gas sorption system analyser used for the study and to the School of Life Sciences at The Robert Gordon University for SEM observations.

References

1. Yildirim Y, Hughes R (2003) An experimental study of CO_2 separation using a silica based composite membrane. Process Saf Environ Prot 81(4):257–261
2. Tsai C, Tam S, Lu Y, Brinker CJ (2000) Dual-layer asymmetric microporous silica membranes. J Membr Sci 169(2):255–268
3. Spillman RW (1989) Economics of gas separation §membranes. Chem Eng Prog 85(1):41–62
4. Baker R, Cussler E, Eykamp W, Koros W, Riley R, Strathman R (1991) Membrane separation systems
5. Göttlicher G, Pruschek R (1997) Comparison of CO_2 removal systems for fossil-fuelled power plant processes. Energy Convers Manag 38:S173–S178
6. Venerus DC, Buongiorno J, Christianson R, Townsend J, Bang IC, Chen G et al (2010) Viscosity measurements on colloidal dispersions (nanofluids) for heat transfer applications 20(4)
7. Klein L, Gallagher D (1988) Pore structures of sol-gel silica membranes. J Membr Sci 39 (3):213–220
8. Sehgal R, Brinker CJ (1998) Supported inorganic membranes. U.S. Patent 5,772,735
9. Beaver RP (1988) Method of producing porous hollow silica-rich fibers. U.S. Patent No. 4,778,499
10. Gavalas G, Megiris C, Nam S (1989) Deposition of H_2-permselective SiO_2 films. Chem Eng Sci 44(9):1829–1835
11. Brinker C, Sehgal R, Hietala S, Deshpande R, Smith D, Loy D et al (1994) Sol-gel strategies for controlled porosity inorganic materials. J Membr Sci 94(1):85–102
12. Gobina E (2006) Apparatus and method for separating gases. U.S. Patent No. 7,048,778
13. Gobina E (2007) Apparatus and method for separating gases. U.S. Patent No. 7,297,184
14. Brinker CJ, Scherer GW (2013) Sol-gel science: the physics and chemistry of sol-gel processing. Academic Press, pp 99–108
15. Nwogu NC, Orakwe IR, Gobina E (2015) Multi-layer silica ceramic membrane for flue gas and natural gas separations. Lecture notes in engineering and computer science: proceedings of the world congress on engineering and computer science 2015, 21–23 October, 2015, San Francisco, USA, pp 651–655

Author Index

S.-I. Ao et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-981-10-2717-8

Subject Index

S.-I. Ao et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-981-10-2717-8