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

Transactions on Engineering Technologies

World Congress on Engineering and Computer Science 2014

Springer

Transactions on Engineering Technologies

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

Transactions on Engineering Technologies

World Congress on Engineering and Computer Science 2014

Springer

Editors
Haeng Kon Kim
Engineering College, Department
of Computer and Communication
Catholic University of DaeGu
DaeGu
Korea, Republic of (South Korea)

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

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

ISBN 978-94-017-7235-8 ISBN 978-94-017-7236-5 (eBook)
DOI 10.1007/978-94-017-7236-5

Library of Congress Control Number: 2013953195

Springer Dordrecht Heidelberg New York London

Printed on acid-free paper

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

Preface

A large international conference on Advances in Engineering Technologies and Physical Science was held in San Francisco, California, USA, October 22–24, 2014, under the World Congress on Engineering and Computer Science (WCES 2014). The WCECS 2014 is organized by the International Association of Engineers (IAENG). IAENG is a nonprofit international association for the engineers and the computer scientists, which was founded originally in 1968 and has been undergoing rapid expansions in recent few years. The WCECS Congress serves as an excellent platform for the engineering community to meet with each other and to exchange ideas. The Congress has also struck a balance between theoretical and application development. The conference committees have been formed with over two hundred members who are mainly research center heads, deans, department heads/chairs, professors, and research scientists from over 30 countries. The full committee list is available at the congress' Web site: www.iaeng.org/WCECS2014/committee.html. The Congress is truly an international meeting with a high level of participation from many countries. The response to the conference call for papers was excellent with more than 600 manuscript submissions for the WCECS 2014. All submitted papers went through the peer review process, and the overall acceptance rate was 51.28 %.

This volume contains 39 revised and extended research articles, written by prominent researchers participating in the congress. Topics covered include engineering mathematics, electrical engineering, circuits, 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, and also serves as an excellent source of reference for researchers and graduate students working with/on engineering technologies and physical science and applications.

Haeng Kon Kim
Mahyar A. Amouzegar
Sio-Iong Ao

The original version of the book frontmatter was revised: The spelling of the last editor's name was corrected. The erratum to the book frontmatter is available at
DOI 10.1007/978-94-017-7236-5_40

Contents

Genetic Algorithm for Energy Consumption Variance Minimisation in Manufacturing Production Lines Through Schedule Manipulation

Chris Duerden, Lik-Kwan Shark, Geoff Hall and Joe Howe

Abstract The typical manufacturing scheduling algorithms do not account for the energy consumption of each job when devising a schedule. This can potentially lead to periods of high energy demand which can be problematic for manufacturers with local infrastructure having limited energy distribution capabilities. In this book chapter, a genetic algorithm based schedule modification algorithm is introduced to optimise an original schedule such that it produces a minimal variance in the total energy consumption in a multi-process manufacturing production line. Results show a significant reduction in energy consumption variance can be achieved on schedules containing multiple concurrent jobs without breaching process constraints.

Keywords Energy consumption prediction · Energy consumption variance · Genetic algorithm · Peak energy minimisation · Production scheduling · Real value encoding

C. Duerden (✉) · G. Hall
Advanced Digital Manufacturing Technology Research Centre, University of Central Lancashire, Burnley Campus, Burnley, Lancashire BB12 0EQ, UK
e-mail: CJDuerden@uclan.ac.uk

G. Hall
e-mail: GHall5@uclan.ac.uk

L.-K. Shark
Advanced Digital Manufacturing Technology Research Centre, University of Central Lancashire, Preston Campus, Preston, Lancashire PR1 2HE, UK
e-mail: lshark@uclan.ac.uk

J. Howe
Thornton Energy Institute, University of Chester, Thornton Science Park, Pool Lane, Ince, Cheshire CH2 4NU, UK
e-mail: j.howe@chester.ac.uk

H.K. Kim et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-94-017-7236-5_1

1 Introduction

In the manufacturing sector, high throughput combined with efficient use of resources is critical for achieving an optimal cost-benefit ratio. During the generation of a production schedule the manufacturing jobs needing to be processed are assigned to the limited machinery and equipment available in such a way to ensure resource availability is never exceeded, and the scheduling constraints are satisfied [1]. Despite the crucial role it plays, many scheduling algorithms do not consider the energy demand. As jobs are executed, their demand for energy can change depending on the elemental operation they are currently performing. If this is not considered when the schedule is generated, there is the potential for very high peaks in energy demand due to the sum demand of many individual jobs running concurrently. As there is always a finite delivery rate in the energy infrastructure, this can potentially limit the availability and capacity of a manufacturing production line. Several researchers have discussed methods for including additional energy-based objectives for traditional schedulers to optimise. Fang et al. and Pechmann et al. both present methodologies for production schedules which aim to also minimise peak energy consumption [2–4]. As does the commercial scheduling software E-PPS by Transfact [5]. While the work of Fang et al. and Pechmann et al. shows promising results, it is concluded by Fang et al. that finding the optimal schedule is difficult due to the complexity and NP-hard nature of the problem. Another promising research focus is the development of intelligent machine controllers which aim to reduce overall manufacturing energy consumption by reducing the idling times of machines by putting them into energy saving modes or shut them down entirely [6–8]. As certain machines may have lengthy or costly start-up procedures, manufacturers typically leave machines idling when not in operation. In the proposed systems, by intelligently deciding when to shut down a machine or put it into an energy saving mode, the total energy consumption for the production line can be reduced.

While all these show promising results, the problem with generating energy optimised schedules has received little attention and appears to be plagued by its NP-hard nature. The use of artificial intelligence in the generation of manufacturing schedules has shown some promising results. Genetic algorithms appear to be a popular choice for solving scheduling optimisation which can include multi-objective [9, 10] and multi-project [11] problems.

In this work, a genetic algorithm is developed to modify the starting times of scheduled jobs, in order to minimise the variance in production line energy consumption without exceeding the process constraints [12]. The technique used is inspired by load-shifting, a traditional energy optimisation method in which energy intensive jobs are scheduled to run during times of low energy tariffs [13]. While the optimisation algorithm discussed below can be applied to all forms of energy consumption, in the context of this work, only electrical power is considered. Following the methodology described in Sect. 2, experiments and results are presented in Sect. 3 to demonstrate the level of potential reduction in energy consumption variance.

2 Methodology

Prior to any energy-based optimisation, a manufacturing schedule is initially generated for a list of jobs using a traditional scheduling algorithm. This will typically output a schedule optimised for processing time. Additionally, factors such as optimal intra-process job order and machine assignment have been determined by a piece of software trusted by the manufacturer. The schedule modification algorithm is then applied. This references job-specific energy profiles and alters the original job starting times in an attempt to minimise the production line energy consumption variance. Because the job start times must represent a valid schedule, not all possible combinations of job start times will be usable. To traverse such a volatile search space a genetic algorithm was selected as the basis for this schedule modification algorithm. This is a popular population-based evolutionary algorithm which has been applied to many complex optimisation problems.

A. Gene Representation and Chromosome Generation

In order for the genetic algorithm to optimise the schedules job start times, the start times themselves must be encoded into a form suitable for the algorithm to efficiently process. For this, real value encoding [14] was selected with the value of each gene in the chromosome representing the start time for a particular job. As such, for a schedule containing N jobs, the genetic algorithm will operate on a population consisting of multiple N length chromosomes. Each chromosome consists of an encoded candidate solution and an associated fitness value which represents the optimality of the candidate solution to the proposed problem. To convert job start times between the chromosome domain G and the time domain S, specialised encoding (1) and decoding (2) functions are used. Here s_e is the earliest starting time in the original schedule and it is used as a reference point for encoding and decoding job start time s_i to or from gene g_i. T is the minimum time alteration that can be applied to the job start time.

Encoding:

$$g_i = (s_i - s_e)/T \tag{1}$$

Decoding:

$$s_i = s_e + (g_i \times T) \tag{2}$$

The scheduling paradigm followed by this schedule modification algorithm defines manufacturing processes as independent of one another, allowing them to run in parallel. However intra-process jobs must be executed sequentially. At the beginning of the algorithms execution, the initial population comprises of $Np-1$ randomly generated chromosomes, where Np is the population size, along with the original encoded schedule. To ensure that the genes of each chromosome comply with this paradigm and the original job order, a relation seed is generated prior to

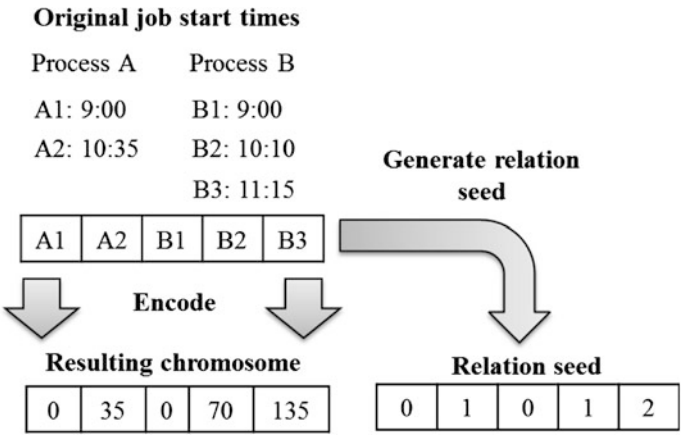


Fig. 1 Diagram showing how the original schedule is used to encode the initial chromosome and generate the relation seed. In this example, the earliest start time s_e is A1/B1, and T is set to 00:01:00

the population generation. The encoded job start times are ordered in the chromosome firstly by process, and secondly by intra-process order. The relation seed is a zero-based unit incrementing numeric vector of equal length to the chromosome, and denoted by $R = \{r_0, \ldots, r_{N-1}\}$. When an encoded job start time is the first in its process, its associated relation seed value is zero. The associated relation seed value following that is assigned a unit incrementing value until it reaches the end of that process. At this stage the count is reset to zero for the first job in the next process. An example relation seed can be seen in Fig. 1. The relation seed is subsequently utilised as one of many constraints by the random number generator to ensure job order is maintained. All these constraints aim to reduce the overall size of the search space and increase the probability of a candidate solution representing a valid schedule. For a job denoted by i and belonging to a process denoted by u with deadline d_u, then the candidate job start time is given by (3),

$$s_i = \begin{cases} 0 \leq random\ number \leq d_u & if\ r_i = 0 \\ s_{i-1} \leq random\ number \leq d_u & if\ r_i > 0 \end{cases} \tag{3}$$

B. Algorithm Overview

Based on Darwinism, a genetic algorithm is designed to locate an optimal or near-optimal solution by evolving a population of chromosomes, which are our candidate schedule solutions. This evolutionary process consists of three steps—Selection, Crossover and Mutation. At the beginning of the process, selection is used to emulate 'survival of the fittest' to generate a reproducing population from the current population. This reproducing population is of size $Np-2$ to allow for two candidate elitism. Every chromosome has a non-equal chance of being selected, with fitter chromosomes having better odds. The selection algorithm used in this

implementation is tournament selection. Here a group of chromosomes are randomly selected and the fittest in the group is assigned to the reproducing population. This process is repeated until the reproducing population reaches the required size of $Np-2$. While there are a number of different selection algorithms available, tournament selection was used due to its simplicity, and as a result, computational efficiency. The reproducing population is then operated on.

Uniform crossover is then applied to generate the new population. Designed to emulate reproduction, chromosomes are grouped into couples and are split into two at the same random point. The two halves are then swapped between chromosomes and recombined to form two new chromosomes. An illustration of this can be found in Fig. 3. Next, each chromosome has a mutation operation applied to it. Here, each gene in the chromosome has a small probability of being mutated, where its value is randomly changed. This allows the algorithm to explore new areas of the search space. After each chromosome in the reproducing population has been operated on, the next generation of the population is constructed by combining the reproduced population with a copy of the two fittest candidates from the previous generation. These are saved via elitism. The fitness of each candidate in the new population is then calculated. In a standard genetic algorithm, this process is repeated either a predetermined number of times or until the optimal value (if known) has been found.

While the algorithms objective is to locate a combination of job start times which will generate a minimal variance in production line energy consumption, this combination of start times must also represent a valid schedule so that it could potentially be executed on the proposed production line. Despite the initial candidate generation constraints, there is the potential for a large percentage of candidates to represent invalid schedules in the population. To compensate for this, additional features have been added to increase the probability of candidates representing valid schedules.

A flow diagram of the algorithm can be seen in Fig. 2. The core of the genetic algorithm is located inside the inner loop. Here the standard operators manipulate the population with an aim to locate the global optimal solution. During the uniform crossover and mutation operations there is the potential for a valid candidate schedule to become invalidated. This is demonstrated in Fig. 3.

After crossover has been applied, each child chromosome is checked at the point of crossover to see if the job order has been breached. If it has, the affected job g_i is modified according to (4).

$$g_i = (g_{i-1} + C_{i-1} + 1) \tag{4}$$

where C_{i-1} is the makespan of the previous job represented by g_{i-1}. Similarly with the mutation operator, if a gene is selected for mutation the range of possible values is limited according to (5),

$$(g_{i-1} + C_{i-1} + 1) \leq g_i \leq (g_{i+1} - C_i - 1) \tag{5}$$

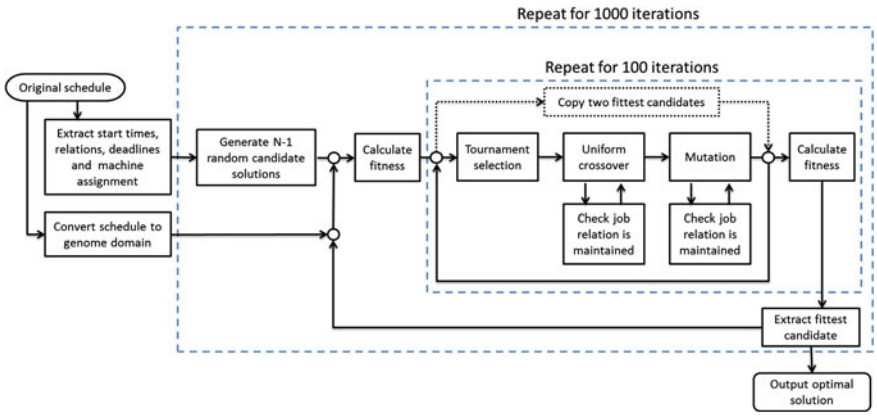


Fig. 2 Flowchart of the genetic algorithm based production schedule optimiser

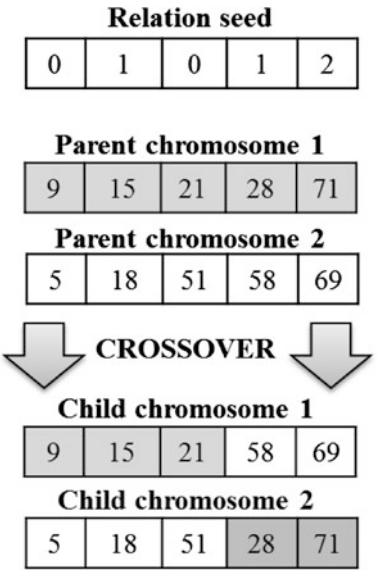


Fig. 3 Diagram showing how a candidate schedule can be invalidated during crossover. In this example, child chromosome *2* is invalid as according to the relation seed, the third job start time (*51*) must commence before the fourth job (*28*). Therefore job four has breached the job order

These checks only consider the directly affected gene and do not examine how applying these modifications will affect the remainder of the chromosome. As such applying these modifications may re-invalidate the chromosome. While it is possible to check the entire chromosome, this would result in additional computational overhead. As a chromosome may be invalid for a number of reasons, this single check is considered an efficient balance between increasing the probability of schedule validity and computational performance. The full set of constraint and validity checks are discussed in Sect. 2C.

Despite all the intra-operator checks, it is still possible for a large percentage of the population to consist of invalid candidate schedules. To prevent invalid candidates from saturating the population, after a predetermined number of iterations, the fittest chromosome is extracted and saved. The entire algorithm is then restarted from the beginning, thus allowing for a fresh start. This can be seen in the outer loop in Fig. 2. To ensure overall progress is not hindered by this, the previously extracted fittest candidate is placed into the new generation, along with N_{p-1} randomly generated candidates. For the very first iteration, the original schedule, produced by a typical schedule algorithm is encoded and input into the first generation. This also ensures that the algorithm will not produce results which are less optimised than the original. This is especially beneficial if the original schedule is also the optimal schedule. After a predetermined number of algorithm restarts, the optimal chromosome is extracted, its genes are decoded back into the time domain, and the original schedule is updated with the new job start times.

C. Fitness function

The fitness function serves two purposes: (a) determine if chromosome G represents a valid schedule that can be executed on the proposed manufacturing line, and if it is valid, (b) build a predicted energy consumption profile based on the encoded job start times, and calculate the variance. The validity conditions for each chromosome are as follows:

For a set of job starts time gene g_i with a makespan of C_i, belonging to the same parent process:

$$g_i > (g_{i-1} + C_{i-1}) \tag{6}$$

For each job with a parent process deadline d_j:

$$d_j > (g_i + C_i) \tag{7}$$

At any time t, usage on a machine of type M_k, $M_{k,\ Usage}$ with a maximum availability of $M_{k,\ X}$:

$$M_{k,Usage}(t) = < M_{k,X} \tag{8}$$

Equation (6) ensures for jobs within the same parent process, the next job does not begin until the previous one is finished. Equation (7) ensures all jobs do not exceed their parent process deadline. Finally (8) ensures that usage of each machine type at any one point in time never exceeds the total amount of that machine. If a particular chromosome fails any of these conditions it is classified as invalid and is assigned the maximum fitness value, in practice this is the maximum value of a double precision number in C#. In this implementation the fittest genome is defined as the one with the minimum fitness value, and therefore minimum energy consumption variance. If a chromosome meets all conditions, the energy consumption

variance for that represented schedule is calculated by generating a predicted energy consumption profile.

Initially an energy consumption profile is created for each machine m in the production line, E_m. This spans from s_e to D_{Max}, and is initially populated with the idling value for that particular machine m_{Idle}. Then in chronological order, every job j using that machine j_m, has its associated energy consumption profile, $j_{Profile}$ copied to E_m, beginning at the start time denoted by the appropriate gene j_g. This process overwrites the values currently assigned to the associated elements of the profile. Additionally the recorded idling consumption from that particular job j_{Idle} is subsequently assigned to the remainder of the profile. This is because the idling energy consumption could have changed if the machine is now in a different position or configuration due to the previous job. This process is repeated until the energy profiles of all jobs running on that machine have been merged. The system assumes that when not in operation, each machine is left idling. Once a machine specific energy profile has been created, the predicted production line energy consumption profile for a total of X machines can be calculated using (9).

$$E_{Total}(t) = \sum_{m=1}^{X} E_m(t) \tag{9}$$

From (9), the variance of the predicted total energy consumption profile is given by:

$$E_{Var} = \frac{T}{(D_{Max} - S_e) - 1} \sum_{t=s_e}^{D_{Max}} \left[E_{Total}(t) - \overline{E_{Total}}\right]^2 \tag{10}$$

where $D_{Max} = \max\{d_1,\ldots,d_L\}$ and $\overline{E_{Total}}$ denotes the average energy consumption. Once calculated the variance is assigned as the chromosomes fitness value.

D. Energy consumption models

The accuracy of the predicted production line energy profile and the entire system is directly related to the accuracy of the individual job-specific energy models. As such it is paramount that these models are reflective of the jobs energy consumption. In similar work by previous researchers, singular values have been used to represent job energy consumption [15–17]. While values such as average, maximum and total energy consumption give an indication as to how much energy is used, no singular value can fully represent how energy will be used over time. As such the most appropriate form would be for the optimisation algorithm to reference time series energy profiles. For the work documented here, simulated energy profiles were generated based on empirical data. These simulated profiles contained waveform features typical of many energy profiles, such as inrush currents, and transients.

To further improve accuracy, the profiles used were highly granular with a simulated power value being reported every 150 ms. This is to ensure that very short peaks in energy consumption, such as inrush peaks, are suitable captured.

While the algorithm is capable of working with any granularity of energy data, a lower granularity will result in a respectively lower accuracy. While the current aspect of industry energy monitoring is not considered in this work, it should be noted that in the manufacturing sector, not all machines have energy monitoring devices installed [18, 19]. For those that do, there is a varying array of features and capabilities with numerous reporting rates and communication methods.

3 Experiments and Results

One of the main disadvantages of a genetic algorithm is its performance is reliant on the values of its internal problem-specific operator probabilities [20]. To ensure the algorithm operates at peak efficiency, significant experimentation is required to determine these probability values.

The proposed genetic algorithm was tested with multiple schedules of increasing complexity. This test set was comprised of a combination of user-generated schedules and schedules generated from an open-source production planning software. In all cases each scheduled job was assigned an energy profile which was not considered during the initial generation.

Table 1 shows the results of an experiment to determine the optimal values for crossover probability and tournament population size—$P_{Crossover}$ and $N_{Tournament}$ respectively. The algorithm was given an $N = 8$ schedule and returned both the optimised schedule and the amount of outer loop iterations it took for the algorithm to locate it. The latter was used as the performance indicator with Table 1 showing the averaged results from ten consecutive runs of each value combination. The range of values tested for $P_{Crossover}$ was chosen as a probability any higher than 85 % may cause too much disruption to a population. This may result in a possible optimal solution being lost before it can be identified. A value less than 65 % would not allow a population to sufficiently reproduce. This has been concluded by other authors [21]. The range of $N_{Tournament}$ sizes was chosen because it was believed that any size less than $Np/8$ would not consider enough of the population to fully emulate 'survival of the fittest'. Any size above $Np/4$ would give too much precedence to the fittest values, and reduce the diversity of the reproducing population. In the original setup, the algorithm was programmed to perform 1000 outer loop iterations before returning the best value found. The results in Table 1 demonstrate that in practice, this number can be significantly reduced.

Table 1 Average number of outer loop iterations until optimal value is generated with differing crossover rate and tournament selection size

$N_{Tournament}$	$P_{Crossover}$		
	0.65	0.75	0.85
$Np/8$	161.875	128.5	165
$Np/6$	203.25	218.375	242
$Np/4$	101.625	123	186.625

Results are averaged over ten repeated experiments

Further experimentation was also carried out to determine how the values for $P_{Crossover}$, $N_{Tournament}$ and the mutation probability, $P_{Mutation}$ influenced the performance of the algorithm in terms of the global optimal solution found. Figure 4 shows the results, where the global optimal variance (GOV) refers to the best solution found by the optimisation algorithm throughout its run. It can be seen that the performance is directly related to the values of the genetic algorithms operator probabilities, as some combinations result in no progress being made. This can be seen in the second graph of Fig. 4. The graph levels off as the algorithm is not capable of locating a solution worse that the variance predicted by the original schedule.

Figure 4 demonstrates that there is no viable combination of operator probabilities which (a) permits maximum performance in returning the global optimal solution, while (b) allowing for universal compatibility with all schedules and constraints.

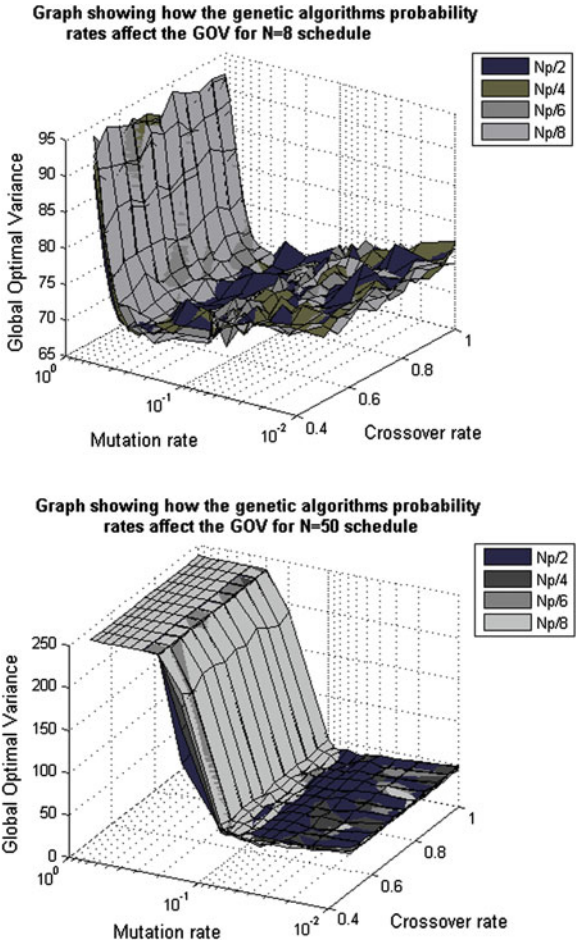


Fig. 4 Graphs showing how the probabilities of the genetic algorithms operators affect the global optimal variance for $N = 12$ and $N = 50$ schedules. Results are averaged over ten repeated experiments

Table 2 Comparison of energy consumption variance in the original and optimised schedules

N	Energy variance in original schedule	Energy variance in optimised schedule	Reduction (%)
8	143.958	52.511	63.523
10	215.111	67.185	68.767
12	237.396	69.090	70.897
15	151.928	72.077	52.558
20	76.960	26.530	65.528
30	144.673	36.395	74.843
50	236.233	42.464	82.025

As such, the optimal operator probabilities were considered to lie in the most common optimal point in both test results. These were selected as $P_{Crossover} = 0.5$, $N_{Tournament} = Np/4$, and $P_{Mutation} = 0.1$. With these settings the algorithm was tested on a range of schedules, each with different constraint values and a different number of jobs. These results, as well as the amount of energy consumption variance that can be achieved can be seen in Table 2.

Table 2 demonstrates that the level of potential variance reduction is significant. This may be due, in part to the fact that typical manufacturing schedules are generated around minimising overall makespan. This involves running many jobs in parallel and as a result, increasing the energy consumption variance for the original schedule. The optimisation algorithm aims to utilise all available time and intelligently distribute the jobs in a hope to reduce this. At this moment it is unknown how the performance of the algorithm is affected by the amount of available time. However, it is demonstrated in Table 2 that the level of variance minimisation is not directly related to the amount of jobs in the schedule, with an average variance reduction of 68.3 % over a range of 8–50 jobs. It is theorised that the performance will be primarily affected by the individual job and process constraints, and the shape of the energy profiles.

4 Conclusions

This book chapter presents an intelligent production schedule modification algorithm which aims to minimise the variance in a production lines energy consumption. The use of a genetic algorithm for individual job start time manipulation is detailed and the algorithms internal parameters are evaluated and optimised based on experimental data. Experimental data proves that the algorithm can successfully locate the optimal set of start times for a series of manufacturing jobs, such that a minimal variance in production line energy usage is produced. Additional constraints and checks also ensure that manufacturing resource limitations and process deadlines are never exceeded. While the global optimum solution is not guaranteed to be returned, a solution near the global optimum is always produced. For each

potential solution, a predicted energy consumption profile is generated based on job specific energy models. The variance of the predicted energy profile is then calculated. As the optimal solution cannot be known beforehand, the algorithm is ran for a predetermined amount of iterations. At the end, the solution which holds the minimal variance is concluded to be the most optimal. However the accuracy of the system is entirely dependent on the accuracy and resolution of the energy consumption profiles. Experiments demonstrate that with suitably accurate models, an accurate and significant reduction in energy consumption variance can be achieved regardless of the amount of jobs in the schedule. Through empirical testing an average reduction percentage of approximately 70 % has been achieved. It is also seen that the reduction percentage is independent of the schedule job count with a range between 8 and 50 jobs tested.

The methodologies described in the book chapter could potentially allow a manufacturing production line to expand without requiring the costly addition of upgrading the energy infrastructure. While only the electrical power infrastructure is discussed here, these methodologies could be practically applied to all forms of energy delivery infrastructure. With suitable power generation predictions, these methodologies could potentially allow a production line to operate entirely from limited supply resources, such as renewable energy sources.

Acknowledgments The authors would like to thank the industrial supervisors Damian Adams and Stuart Barker of BAE Systems (Operations) Ltd. Their insights and dedication have proven invaluable to the success of this work. This work is financially supported by the Engineering and Physical Sciences Research Council (EPSRC) and BAE Systems (Operations) Ltd.

References

1. Karger D, Stein C, Wein J (2009) Scheduling algorithms, algorithms and theory of computation handbook, 2nd edn. Chapman & Hall/CRC, Florida, pp 20–21, 20–34
2. Fang K, Uhan N, Zhao F, Sutherland JW (2013) Flow shop scheduling with peak power consumption constraints. Ann Oper Res 206(1):115–145
3. Pechmann A, Schöler I (2011) Optimizing energy costs by intelligent production scheduling. In: Glocalized solutions for sustainability in manufacturing. Springer, New York, pp 293–298
4. Fang K, Uhan N, Zhao F, Sutherland JW (2011) A new shop scheduling approach in support of sustainable manufacturing. In: Glocalized solutions for sustainability in manufacturing, pp 305–310
5. Transfact. Energy in Production Planning and Scheduling (E-PPS). Available: http://www.transfact.de/EN/
6. Mouzon G, Yildirim MB, Twomey J (2010) Operational methods for minimization of energy consumption of manufacturing equipment. Int J Prod Res 45(18–19):4247–4271
7. Eberspächera P, Verla A (2013) Realizing energy reduction of machine tools through a control-integrated consumption graph-based optimization method. In: Forty Sixth CIRP conference on manufacturing systems 2013, vol 7, pp 640–645
8. Orio GD, Candido G, Barata J, Bittencourt JL, Bonefeld R (2013) Energy efficiency in machine tools—a self-learning approach. In: IEEE international conference on systems, man, and cybernetics, pp 4878–4883

9. Yildirim MB, Mouzon G (2012) Single-machine sustainable production planning to minimize total energy consumption and total completion time using a multiple objective genetic algorithm. IEEE Trans Eng Manage 59(4):585–597
10. Yassa S, Sublime J, Chelouah R, Kadima H, Jo G-S, Granado B (2013) A genetic algorithm for multi-objective optimisation in workflow scheduling with hard constraints. Int J Metaheuristics 2(4):415–433
11. Gonçalves JF, Mendes JJM, Resende MGC (2008) A genetic algorithm for the resource constrained multi-project scheduling problem. Eur J Oper Res 189(3):1171–1190
12. Duerden C, Shark LK, Hall G, Howe J (2014) Genetic algorithm based modification of production schedule for variance minimisation of energy consumption, lecture notes in engineering and computer science. In: Proceedings of the world congress on engineering and computer science 2014, WCECS 2014, 22–24 Oct 2014, San Francisco, USA, pp 370–375
13. Brown N, Greenough R, Vikhorev K, Khattak S (2012) Precursors to using energy data as a manufacturing process variable. In: 6th IEEE international conference on digital ecosystems technologies, pp 1–6
14. Herrera F, Lozano M, Verdegay JL (1998) Tackling real-coded genetic algorithms: operators and tools for behavioural analysis. Artif Intell Rev 12(4):265–319
15. Bruzzone AAG, Anghinolfi D, Paolucci M, Tonelli F (2012) Energy-aware scheduling for improving manufacturing process sustainability: a mathematical model for flexible flow shops. CIRP Ann Manufact Technol 61:459–462
16. Cataldo A, Taisch M, Stahl B (2013) Modelling, simulation and evaluation of energy consumption for a manufacturing production line. In: 39th Annual conference of the IEEE industrial electronics society, pp 7537–7542
17. Kara S, Li W (2011) Unit process energy consumption models for material removal proceses. CIRP Ann Manufact Technol 60:37–40
18. Arinez J, Biller S (2010) Integration requirements for manufacturing-based energy management systems. In: IEEE PES conference on innovative smart grid technologies, pp 1–6
19. Cannata A, Karnouskos S, Taisch M (2009) Energy efficiency driven process analysis and optimization in discrete manufacturing. In: 35th Annual conference of IEEE industrial electronics, pp 4449–4454
20. Yang X-S (2014) Chapter 5—Genetic algorithm, nature-inspired optimization algorithm. Elsevier, London
21. Lin WY, Lee WY, Hong TP (2003) Adapting crossover and mutation rates in genetic algorithms. J Inf Sci Eng 19:889–903

A Unified Approach to Data Analysis and Modeling of the Appearance of Materials for Computer Graphics and Multidimensional Reflectometry

Mikhail Langovoy

Abstract Characterizing the appearance of real-world surfaces is a fundamental problem in multidimensional , computer vision and computer graphics. In this paper, we outline a unified perception-based approach to modeling of the appearance of materials for computer graphics and reflectometry. We discuss the differences and the common points of data analysis and modeling for BRDFs in both physical and in virtual application domains. We outline a mathematical framework that captures important problems in both types of application domains, and allows for application and performance comparisons of statistical and machine learning methods. For comparisons between methods, we use criteria that are relevant to both statistics and machine learning, as well as to both virtual and physical application domains. Additionally, we propose a class of multiple testing procedures to test a hypothesis that a material has diffuse reflection in a generalized sense. We treat a general case where the number of hypotheses can potentially grow with the number of measurements. Our approach leads to tests that are more powerful than the generic multiple testing procedures.

Keywords BRDF · Computer graphics · Data analysis · Light reflection · Machine learning · Metrology · Perception · Realistic image representation · Reflectometry · Statistics of manifolds

1 Introduction

Characterizing the appearance of real-world surfaces is a fundamental problem in multidimensional reflectometry, computer vision and computer graphics. For many applications, appearance is sufficiently well characterized by the bidirectional reflectance distribution function (BRDF).

M. Langovoy (✉)
The Physikalisch-Technische Bundesanstalt, Abbestrasse 2–12, 10587 Berlin, Germany
e-mail: mikhail.langovoy@ptb.de

H.K. Kim et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-94-017-7236-5_2

In the case of a fixed wavelength, BRDF describes reflected light as a four-dimensional function of incoming and outgoing light directions. In a special case of rotational symmetry, isotropic BRDFs are are used. Isotropic BRDFs are functions of only three angles. On the other hand, for modelling or describing complicated visual effects such as goniochromism or irradiance, an extra dimension accounting for the wave length has to be added.

In computer graphics and computer vision, usually either physically inspired analytic reflectance models [1–3], or parametric reflectance models chosen via qualitative criteria [4–7] are used to model BRDFs. These BRDF models are only crude approximations of the reflectance of real materials. Moreover, analytic reflectance models are limited to describing only special subclasses of materials.

In multidimensional reflectometry, an alternative approach is usually taken. One directly measures values of the BRDF for different combinations of the incoming and outgoing angles and then fits the measured data to a selected analytic model using optimization techniques. There are several shortcomings to this approach as well.

An alternative approach to fitting parametric models is in constructing more realistic BRDFs on the basis of actual BRDF measurements. This approach bridges the gap between computer graphics and industrial reflectometry. For example, [8] and [9] modelled reflectance of materials in nature as a linear combination of a small set of basis functions derived from analyzing a large number of densely sampled BRDFs of different materials.

There were numerous efforts to use modern machine learning techniques to construct data-driven BRDF models. Brady et al. [10] proposed a method to generate new analytical BRDFs using a heuristic distance-based search procedure called Genetic Programming. In [11], an active learning algorithm using discrete perceptional data was developed and applied to learning parameters of BRDF models such as the Ashikhmin—Shirley model [12].

In computer graphics, it is important that BRDF models should be processed in real-time. Computer-modelled materials have to remind real materials qualitatively, but the quantitative accuracy was not considered as important. The picture in reflectometry and metrology was almost the opposite: there was typically no need in real-time processing of BRDFs, but quantitative accuracy was always the paramount. In view of this, some of the breakthrough results from computer vision and animation would not fit applications in reflectometry and in many industries.

Another difference with virtual reality models is that in computer graphics measurement uncertainties are essentially never present. This is not the case in metrology, reflectometry and in any real-world based industry [13]. Since measurement errors can greatly influence shape and properties of BRDF manifolds, there is a clear need to develop new methods for handling BRDFs with measurement uncertainties.

In this paper, we treat BRDF measurements as samples of points from a high-dimensional and highly non-linear non-convex manifold. We argue that any realistic statistical analysis of BRDF measurements, or any parameter or manifold learning procedure applied to BRDF measurements has to account both for

nonlinear structure of the data as well as for an ill-behaved noise. Standard statistical and machine learning methods can not be safely directly applied to BRDF data. Our study of parameters for generalized Lambertian models in Sects. 5 and 6 clarifies certain pitfalls in analysis of BRDF data, and helps to understand and develop more refined estimates for generalized Lambertian models in Sect. 7.

We introduce and apply in Sect. 6 the notion of Pitman closeness to compare different estimators and parameter learning methods that could be applied to BRDF models. To the best of our knowledge, [14] and [15] were the first works where the Pitman closeness criterion was introduced to either fields of computer graphics as well as metrology. This criterion for comparison of estimators appeals to the actually observable precision of estimators and is assumption-free and loss function-independent, and thus seems to be especially appropriate for applications in metrology, as well as for comparative studies of parameter learning procedures derived for different types of loss functions.

We use the generalized Lambertian model parameter estimators from Sect. 7 to build statistical tests to test a hypothesis whether any particular material is diffuse, even if in a weak sense, or not. Testing validity of BRDF models is important for computer graphics, even though rarely done in a rigorous way, with [16] being a notable exception dealing with several types of tests for parametric models. Surprisingly, hypothesis testing for BRDF data is rarely studied in metrology and reflectometry as well. Recent works [17] and [18] deals with hypothesis testing for diffuse reflection standards. In this paper, we treat a more general case of generalized Lambertian BRDFs, which demands simultaneous testing for a set of stochastically ordered hypotheses, where the number of those hypothesis is the number of measured BRDF layers and so can potentially grow with the number of measurements available. We build a class of tests for this complicated set of hypotheses, and show that our approach leads to tests that are more powerful than the generic multiple testing procedures often recommended by default in the literature.

2 Unified Approach

In our research, we advocate the use of a universal approach to data analysis and material appearance modeling, based on those goals that are common both for computer graphics as well as for industrial applications. Accurate simulations are important for virtual applications as well, because that would allow to use computer graphics algorithms for testing and development of complicated real-life technologies, avoiding the difficulties of running expensive physical experiments. Our approach can be seen as complementing the methodology proposed in [19].

An important first step of our approach is to formulate the problems within a rigorous mathematical framework, necessarily including criteria for comparisons between simulations, predictions and actual measurements. An important advantage here is that for a mathematically formalized problem, there is a large toolbox of

methods from applied mathematics, statistics and machine learning that can be applied to this problem. For example, when a criterion for comparisons of results is not specified, the whole decision-theoretic framework is not even applicable.

Methods from different areas often rely on different types of model assumptions. In those situations where it is desirable to compare a range of methods, we advocate the use of assumption-free and loss function-independent criteria. A possible example is the closeness of estimators.

Following [19], we also split our framework into 3 specialized parts:

1. Light reflection models (via BRDFs) and their validation.
2. Light transport simulation.
3. Perception-based studies tailored for specific applications.

Notice that a proper treatment of Stage 3 can greatly reduce computational expense of the global illumination algorithms. Algorithms are substantially accelerated if one develops a perception-based metric evaluating perceptual importance of scene features, because in this case all the numerous unimportant features can be singled out and the amount of operations to process the unimportant features can be greatly reduced.

In the framework developed in [19], only physically based error metrics are used at Stage 1 and Stage 2 (see also [16, 20]). In our framework, we start applying perception-based metrics already at the Stage 1. In [21], we started developing perception-based metrics for the space of BRDFs. We believe that this approach makes a difference when studying sampling of BRDF manifolds and efficient ways to measure the light reflection.

The main goal of Stage 3 in the approach of [19] is to create photorealistic synthetic images which are perceptionally indistinguishable from real scenes. Our goal at Stage 3 is formulated in a more flexible way. Suppose that our study involves materials (or physical scenes) $\mathscr{M}_1, \mathscr{M}_2, \ldots, \mathscr{M}_m$, and we are interested to evaluate the function $\mathfrak{F}$ on m arguments, $\mathfrak{F}(\mathscr{M}_1, \mathscr{M}_2, \ldots, \mathscr{M}_m)$. A typical example here is the customer preference function, that returns the number i of the material $\mathscr{M}_\mathrm{i}$ that appears the most attractive. Then our goal at Stage 3 is in simulating the scenes $\widehat{\mathscr{M}_1}, \ldots, \widehat{\mathscr{M}_m}$ such that

$$\mathfrak{F}(\mathscr{M}_1, \mathscr{M}_2, \ldots, \mathscr{M}_m) = \mathfrak{F}(\widehat{\mathscr{M}_1}, \ldots, \widehat{\mathscr{M}_m}).$$

For the customer preference example, our simulated images are of satisfactory quality if the customer's choice can be guessed correctly by looking at the simulated images alone.

These specialized tasks are much less strict than building fully photorealistic images. Mathematically, we substantially reduce dimensionality of the problem, and relax our demands on precision of light transport studies. It can be expected that the algorithms would run faster due to the smaller and less demanding problem.

3 Main Definition

The bidirectional reflectance distribution function (BRDF), $f_r(\omega_i, \omega_r)$ is a four-dimensional function that defines how light is reflected at an opaque surface. The function takes a negative incoming light direction, ω_i, and outgoing direction, ω_r, both defined with respect to the surface normal **n**, and returns the ratio of reflected radiance exiting along ω_r to the irradiance incident on the surface from direction ω_i. Each direction ω is itself parameterized by azimuth angle ϕ and zenith angle θ, therefore the BRDF as a whole is 4-dimensional. The BRDF has units sr^{-1}, with steradians (sr) being a unit of solid angle.

The BRDF was first defined by Nicodemus in [22]. The defining equation is:

$$f_r(\omega_i, \omega_r) = \frac{dL_r(\omega_r)}{dE_i(\omega_i)} = \frac{dL_r(\omega_r)}{L_i(\omega_i)\cos\theta_i d\omega_i}. \quad (1)$$

where L is radiance, or power per unit solid-angle-in-the-direction-of-a-ray per unit projected-area-perpendicular-to-the-ray, E is irradiance, or power per unit surface area, and θ_i is the angle between ω_i and the surface normal, **n**. The index i indicates incident light, whereas the index r indicates reflected light.

In the basic definition it is assumed that the wavelength λ is fixed and is the same for both the incoming and the reflected light. In order to model complicated visual effects such as iridescence, luminescence and structural coloration, or to model materials such as pearls, crystals or minerals, as well as to analyze the related data, it is necessary to have an extended, wavelength-dependent definition of BRDFs. Fortunately, formally this new definition is relatively straightforward and is obtained by rewriting Eq. (1) for $f_r(\lambda_i, \omega_i, \lambda_r, \omega_r)$, where λ_i and λ_r are the wavelengths of the incoming and the reflected light respectively.

4 Important Models of Diffuse Reflection

Lambertian model [4] represents reflection of perfectly diffuse surfaces by a constant BRDF. Because of its simplicity, Lambertian model is extensively used as one of the building blocks for models in computer graphics. Most of the recent studies of light reflection by means of advanced machine learning methods still rely on the Lambertian model. Examples include color studies [23, 24], analytic inference [25], perception studies [26], and face detection [27].

It was believed for a long time that the so-called standard diffuse reflection materials exhibit Lambertian reflectance, but recent studies with actual BRDF measurements convincingly reject this hypothesis [17, 18, 28].

We refer to [14, 15] for a brief discussion of the Oren"-Nayar "directed–diffuse" microfacet model [1], a sophisticated model by [29], and the Lommel"-Seeliger model [30] of the lunar and Martian reflection.

5 Statistical Analysis of BRDF Models

In this section, we treat parameter estimation for BRDF models of standard diffuse reference materials. These materials are supposed to have ideal diffuse reflection with constant BRDFs. Graphically, for each incoming angle θ_i, φ_i, the resulting BRDF $f_r(\omega_i, \omega_r)$ is a (subset of) two-dimensional upper hemisphere. The radius ρ of this hemisphere is the parameter that we aim to estimate in this paper.

As we mentioned before, the Lambertian model has been shown to be inaccurate even for those materials that were designed to be as close to perfectly diffuse as possible. Therefore, parameter estimates determined for the Lambertian model can hardly be used in practice. However, there are two methodological reasons that make these estimators worth a separate study.

First, BRDF measurements represent a sample of points from a high-dimensional and highly non-linear non-convex manifold. Moreover, these measurements are collected via a nontrivial process, possibly involving random or systematic measurement errors of digital or geometric nature. These two observations suggest that any realistic statistical analysis of BRDF measurements has to account both for nonlinear structure of the data as well as for a very ill-behaved noise and heavy-tailed noise. Any type of statistical inference is more complicated in these conditions, see, e.g., [31]. Standard statistical methods typically assume nice situations like i.i.d. normal errors, and can not be safely directly applied to BRDF data. The same applies to statistical analysis of image data in general [32]. Our study of parameters for Lambertian models clarifies certain pitfalls in analysis of BRDF data, and helps to understand and develop more refined estimates for more realistic BRDF models that will be studied in subsequent papers.

Second, we would use the Lambertian model parameter estimators to build statistical tests to test a hypothesis whether any particular material is perfect diffuse or not. This will be studied in a separate paper.

Suppose we have measurements of a BRDF available for the *set of incoming angles*

$$\Omega_{inc} = \{\omega_i^{(p)}\}_{p=1}^{P_{inc}} = \{(\theta_i^{(p)}, \varphi_i^{(p)})\}_{p=1}^{P_{inc}}. \tag{2}$$

Here $P_{inc} \geq 1$ is the total number of incoming angles where the measurements were taken. Say that for an incoming angle $\{\omega_i^{(p)}\}$ we have measurements available for angles from the *set of reflection angles*

$$\Omega_{refl} = \bigcup_{p=1}^{P_{inc}} \Omega_{refl}(p), \tag{3}$$

where

$$\Omega_{refl}(p) = \{\omega_r^{(q)}\}_{q=1}^{P_{refl}(p)} = \{(\theta_r^{(q)}, \varphi_r^{(q)})\}_{q=1}^{P_{refl}(p)},$$

where $\{P_{refl}(p)\}_{p=1}^{P_{inc}}$ are (possibly different) numbers of measurements taken for corresponding incoming angles.

Overall, we have the set of random observations

$$\{f(\theta_i^{(p)}, \varphi_i^{(p)}, \theta_r^{(q)}, \varphi_r^{(q)}) \mid (\theta_i^{(p)}, \varphi_i^{(p)}) \in \Omega_{inc}, (\theta_r^{(q)}, \varphi_r^{(q)}) \in \Omega_{refl}(p)\}. \quad (4)$$

Our aim is to infer properties of the BRDF function (1) from the set of observations (4). In general, the connection between the true BRDF and its measurements is described via a stochastic transformation T, i.e.

$$f(\omega_i, \omega_r) = T(f_r(\omega_i, \omega_r)), \quad (5)$$

where

$$T : \mathscr{M} \times \mathscr{P} \times \mathscr{F}_4 \to F_4, \quad (6)$$

with $\mathscr{M} = (M, \mathfrak{A}, \mu)$ is an (unknown) measurable space, $\mathscr{P} = (\Pi, \mathfrak{P}, \mathbb{P})$ is an unknown probability space, $\mathscr{F}_4$ is the space of all Helmholtz-invariant energy preserving 4-dimensional BRDFs, and F_4 is the set of all functions of 4 arguments on the 3-dimensional unit sphere S^3 in $\mathbb{R}^4$.

Equations (5) and (6) mean that there could be errors of both stochastic or non-stochastic origin. In this setting, the problem of inferring the BRDF can be seen as a statistical inverse problem. However, contrary to much literature on this subject, we do not assume linearity of the transformation T, we do not assume that T is purely stochastic, and we do not assume an additive model with zero-mean parametric errors, as these assumptions do not seem realistic for BRDF measurements.

Of course, this setup is intractable in full generality, but for special cases such as inference for Lambertian model, we would be able to obtain quite general solutions (see also [21]).

It is also easily observable (see, e.g., [28]) that for all materials their sub-BRDFs, consisting of measurements for different incoming angles, look substantially different (no matter if we believe in the underlying Lambertian model or not). This suggests that different sub-BRDFs of the same material still have different parameter values, and this in turn calls for applying statistical procedures separately for different sub-BRDFs and for combining the results via model selection, multitesting and related techniques.

6 Means, Medians and Robust Estimators

6.1 Basic Properties of Distributions in BRDF Data

In our choice of estimators for parameters in BRDF models, we have to take into account specific properties of BRDF data. It is important to notice that, due to the complicated structure of measurement devices, outliers are possible in the data. Additionally, due to technical difficulties in measuring peak values of BRDFs (see [33, 34]), we have to count on the fact that certain (even though small) parts of the data contain observations with big errors. This also leads us to conclusion that, even for simplest additive error models, we cannot blindly assume that random errors are identically distributed throughout the whole manifold. Additionally, missing data are possible and even inevitable for certain angles. Measurement angles are often non-uniformly distributed. In view of the above arguments, a useful estimator for any BRDF model has to exhibit certain robustness against outliers and dependent or mixed errors.

An estimator has to be universal enough in the sense that it has to be applicable to BRDF samples without requiring extra regularity in the data set, such as uniformly distributed design points, pre-specified large number of measurements, or absence of missing values. This observation suggests that simpler estimators are more practical for BRDF data than complicated (even if possibly asymptotically optimal) estimators, as the later class of estimators has to rely on rather strict regularity assumptions about the underlying model.

6.2 Pitman Closeness of Estimators

Let Ω be a probability space and let $\widehat{\theta}_1 : \Omega \to \mathbb{R}$ and $\widehat{\theta}_2 : \Omega \to \mathbb{R}$ be estimators of a parameter $\theta \in \mathbb{R}$. Then the *Pitman relative closeness* of these two estimators at the point θ is defined as

$$\mathscr{P}(\widehat{\theta}_1, \widehat{\theta}_2; \theta) = \mathbb{P}(|\widehat{\theta}_1 - \theta| < |\widehat{\theta}_2 - \theta|). \tag{7}$$

The estimator $\widehat{\theta}_1$ is *Pitman closer* to θ than $\widehat{\theta}_2$, if

$$\mathscr{P}(\widehat{\theta}_1, \widehat{\theta}_2; \theta) > 1/2.$$

While this criterion for comparison of estimators is much less known as, say, unbiasedness or asymptotic variance, it appeals to the actually observable precision of estimators, and thus seems of much interest for applications in metrology.

The closeness criterion appeals to the actually observed precision of estimators and is assumption-free and loss function-independent, and thus seems to be especially appropriate for comparative studies of parameter learning procedures derived

for different types of loss functions. As a drawback, the Pitman closeness has some nontrivial properties such as non-transitivity [35], which leads to counterintuitive results in several examples [36]. On the other hand, these nontrivial properties help to clarify some classic statistical paradoxes such as the Stein paradox [37, 38].

We refer to [39] for an extensive discussion of the relative closeness of estimators and other related notions and their properties. Besides unbiasedness, asymptotic variance and relative closeness, there are many other criteria for comparing quality of statistical estimators. At least 7 of them can be found in [40].

We apply the notion of Pitman closeness to compare different estimators that could be used in BRDF models. Based on this and other criteria, we show that, in the context of the BRDF model parameter estimation and parameter learning, procedures based on either median or trimmed mean are safer to use and are often more accurate than procedures based on sample means.

6.3 Mean and Median

We refer to [14, 15] for definitions of the sample mean, the sample median and the trimmed (truncated) mean. Sample mean is known to be an asymptotically efficient estimator, as well as a uniformly minimum-variance unbiased estimator, for the expected value of the random variable. However, it is important to note that these nice properties are guaranteed only for sufficiently "nice" distributions (see [41] or [42]), while sometimes even marginal deviations from these smooth models seriously spoil performance of the sample mean estimator. In view of the above discussion of properties of BRDF data, we conclude that it is not advisable to apply the sample mean directly as an estimator of the Lambertian radius.

In this and in the next subsection, we present some results of an extensive Monte Carlo experiment comparing relative closeness of different types of basic nonparametric estimators. Each of the graphs contains values of relative closeness obtained for samples of all sizes ranging from 1 to 1000 observations. We performed 1,000,000 comparisons for each sample size. We refer to [14] and [15] for more examples and details.

However, if we are dealing with a heavy-tailed distribution, the picture changes. Suppose we are presented with a Cauchy distribution, and our goal is to estimate the mode (the mean does not exist in this case). Then Fig. 1 shows that the relative closeness of the mean tends to 0 when compared with the median.

Mean surprisingly loses its efficiency even in rather smooth toy situations. Suppose that a sample from i.i.d. standard normal distribution is contaminated with 5 % of i.i.d. normals with mean 0 and variance 10. The result is shown on Fig. 2. Mean's closeness compared to median drops to 0.3. Even more surprisingly, if we start with a sample of i.i.d. normals with mean 0 and variance 100 and contaminate this sample with just 5 % of i.i.d. normals with mean 0 and small variance 1, the drop in mean's closeness compared to median is even worse. Figure 3 shows that the relative closeness of mean drops to 0.1.

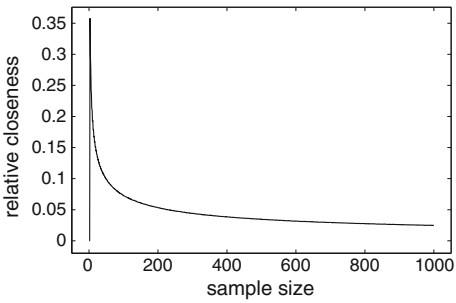


Fig. 1 Median grossly outperforms mean for heavy-tailed distributions

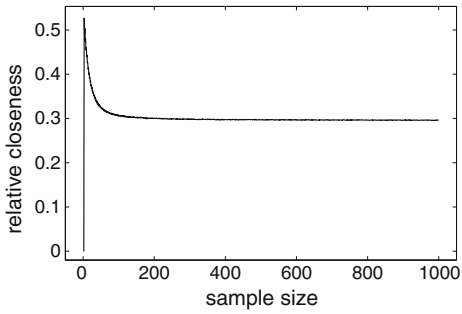


Fig. 2 Median can outperform mean for mixtures of normal distributions

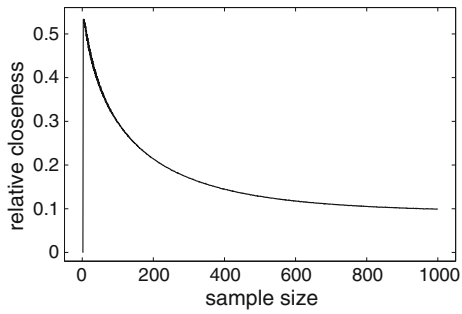


Fig. 3 Median can outperform mean for mixtures of normal distributions with small errors

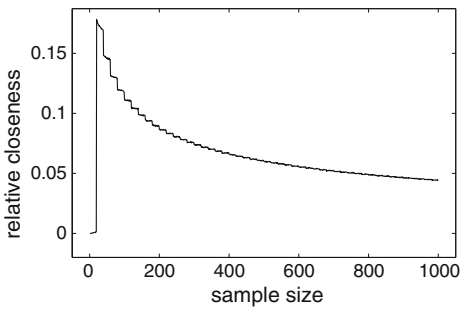


Fig. 4 Trimmed mean totally dominates mean for cauchy distributions

6.4 Truncated Mean and Mean

If our data are generated by sufficiently nice distribution such as, say, a normal distribution, then the sample mean is an efficient estimator. In those cases, it can be rigorously proven that Mean is better than Trimmed Mean in the sense of both Pitman closeness, as well as asymptotic relative efficiency.

The picture can be reversed when our data are allowed to contain outliers or when the data can be, at least partially, generated by a heavy-tailed distribution (which is the case when large values of measurement errors are possible, as is the case for BRDF measurements of specular peaks). We give here a toy example with a Cauchy distribution. Figure 4 illustrates the relative efficiency of mean compared to the trimmed mean with 10 % of the extremes in data being discarded. The unusual shape of the relative closeness curve has no explanation at the moment.

Here the mean is an inconsistent estimator of the median of the distribution, while the truncated mean is not only a consistent estimator of the median, but, with a proper choice of the truncation point, is capable of outperforming the sample median in estimating the median [43]! One needs to drop out about 76 % of the data, though. In fact, even more efficient estimators exist [44], but they require to drop out almost all of the data, and we would not advise to use them for estimation in BRDF models or for any work with moderate sample sizes.

7 Parameter Estimation for Generalized Lambertian Models

For each $\omega_i^{(p)}$ from the set of incoming angles Ω_{inc}, let $\rho^{(p)}$ denote the Lambertian radius of the BRDF's layer

$$\{f(\theta_i^{(p)}, \varphi_i^{(p)}, \theta_r^{(q)}, \varphi_r^{(q)}) \,|\, (\theta_r^{(q)}, \varphi_r^{(q)}) \in \Omega_{refl}(p)\}, \tag{8}$$

where $\Omega_{refl}(p)$ is defined by (3). Thus, we are estimating the P_{inc}-dimensional parameter vector

$$\{\rho^{(p)}\}_{p=1}^{P_{inc}}. \tag{9}$$

For $1 \le p \le P_{inc}$, let

$$\{f_{(i)}^{(p)}\}_{i=1}^{P_{refl}(p)} \tag{10}$$

be the non-decreasing sequence of order statistics of the subsample (8). Then the *sample median estimator* of the parameter vector (9) is defined as

$$\{\widehat{smed^{(p)}}\}_{p=1}^{P_{inc}}, \tag{11}$$

where

$$\widehat{smed^{(p)}}(f) = \begin{cases} f^{(p)}_{((P_{refl}(p)+1)/2)}, & P_{refl}(p) \text{ is odd;} \\ \frac{1}{2}(f^{(p)}_{(P_{refl}(p)/2)} + f^{(p)}_{(P_{refl}(p)/2+1)}), & P_{refl}(p) \text{ is even.} \end{cases}$$

Let $0 \leq \alpha < 1/2$ be a number, and let $[\cdot]$ denote the integer part of a real number. Then the *sample trimmed mean* estimator of the parameter vector (9) is defined as

$$\{\widehat{tm_{\alpha}^{(p)}}\}_{p=1}^{P_{inc}}, \tag{12}$$

where

$$\begin{aligned} \widehat{tm_{\alpha}^{(p)}}(f) = & \frac{1}{P_{refl}(p)(1-2\alpha)} \\ & \times \{([P_{refl}(p)\alpha] + 1 - P_{refl}(p)\alpha)(f^{(p)}_{([P_{refl}(p)\alpha]+1)} \\ & + f^{(p)}_{(P_{refl}(p)-[P_{refl}(p)\alpha])}) + \sum_{i=[P_{refl}(p)\alpha]+2}^{P_{refl}(p)-[P_{refl}(p)\alpha]-1} f^{(p)}_{(i)}\}. \end{aligned}$$

8 Hypothesis Testing for Generalized Diffuse Reflection Models

It is rather straightforward to build a test for checking whether any particular material is perfectly diffuse. Indeed, the corresponding null hypothesis can be tested via a t-statistic on the basis of the observed set of BRDF values. However, as we noted above, testing this hypothesis is not very informative as this null hypothesis will be rejected even for those materials that serve as diffuse reflectance standards.

Therefore, it makes more sense to test a hypothesis that a material has diffuse reflection in general, even though not perfectly diffuse with the same level of reflection for each incoming angle. This amounts to building a multiple testing procedure for testing the joint hypothesis $H_0 = \bigcap_{1 \leq p \leq P_{inc}} H_p$, where H_p is the p-th null hypothesis stating that the p-th layer (8) is laying on a sphere.

As an application of the above estimators, we propose now a class of tests for the compound hypothesis H_0. Consider any sequence of test statistics $\{MT_p\}_{1 \leq p \leq P_{inc}}$, where MT_p is used for testing the corresponding hypothesis H_p. For a given sample of points from the BRDF, let us apply the test based on MT_p for testing the

hypothesis H_p for all p. Denote the corresponding resulting p-values by $PV_1, \ldots, PV_{P_{inc}}$, and let $PV_{(1)} \leq \ldots \leq PV_{(P_{inc})}$ be the ordered set of these p-values. Then one could suggest to reject H_0 if $PV_{(p)} \leq p\alpha/P_{inc}$ for at least one p.

Under certain conditions, this multiple testing procedure is asymptotically consistent and more powerful than the procedure based on the Bonferroni principle applied to the same sequence of test statistics $\{MT_p\}_{1 \leq p \leq P_{inc}}$, which is often assumed to be the default way of testing several hypothesis simultaneously. Our procedure capitalizes on the physical fact that, as the incoming light angle grows, deviations from diffuse reflection can only grow as well. Therefore, in mathematical terms, the test statistics $\{MT_p\}_{1 \leq p \leq P_{inc}}$ would be highly positively correlated for any reasonable choice of these statistics. See [45] for details related to rigorous analysis of this type of multiple testing methods. A specific example of test statistics $\{MT_p\}_{1 \leq p \leq P_{inc}}$ was considered in [14] and [15].

Note that it is crucial to take into account the multiplicity of tests. Otherwise, irrespectively of what kind of test statistics we use, if the decisions about each of the basic hypothesis $H_0, \ldots, H_{P_{inc}}$ are made on the basis of the unadjusted marginal $p-$values, then the probability to reject some true null hypothesis will be too large and the test will not be reliable. Unfortunately, this mistake is commonly made in applications of multiple testing.

Acknowledgments The author would like to thank Gerd Wübbeler, Clemens Elster and Franko Schmähling for useful discussions and Anna Langovaya for her suggestions that led to improvement of results in this paper.

This work has been carried out within EMRP project IND 52 'Multidimensional reflectometry for industry'. The EMRP is jointly funded by the EMRP participating countries within EURAMET and the European Union.

References

1. Oren M, Nayar SK (1995) Generalization of the lambertian model and implications for machine vision. Int J Comput Vis 14(3):227–251
2. Cook RL, Torrance KE (1981) A reflectance model for computer graphics. ACM Siggraph Comput Graph 15(3):307–316
3. He XD, Torrance KE, Sillion FX, Greenberg DP (1991) A comprehensive physical model for light reflection. In: ACM SIGGRAPH computer graphics, vol 25, no 4. 1. em plus 0.5em minus 0.4em. ACM, pp 175–186
4. Lambert J, Anding E, (1982) Lamberts Photometrie: (Photometria, sive De mensura et gradibus luminis, colorum et umbrae) (1760), ser. In: Engelmann W (ed) Ostwalds Klassiker der exakten Wissenschaften, vol 1–2. [Online]. Available http://books.google.de/books?id=Fq4RAAAAYAAJ
5. Phong BT (1975) Illumination for computer generated pictures. Commun ACM 18(6):311–317. [Online]. Available http://doi.acm.org/10.1145/360825.360839
6. Blinn JF (1977) Models of light reflection for computer synthesized pictures. In: Proceedings of the 4th annual conference on computer graphics and interactive techniques, ser. SIGGRAPH '77. ACM, New York, pp 192–198

7. Lafortune EP, Foo S-C, Torrance KE, Greenberg DP (1997) Non-linear approximation of reflectance functions. In: Proceedings of the 24th annual conference on computer graphics and interactive techniques, ACM Press/Addison-Wesley Publishing Co., pp 117–126
8. Matusik W (2003) A data-driven reflectance model. Ph.D. dissertation, Citeseer
9. Matusik W, Pfister H, Brand M, McMillan L (2003) A data-driven reflectance model. In ACM transactions on graphics (TOG), vol. 22, no. 3. 1. em plus 0.5em minus 0.4em. ACM, pp 759–769
10. Brady A, Lawrence J, Peers P, Weimer W (2014) Genbrdf: discovering new analytic brdfs with genetic programming. ACM Trans Graph 33(4):114:1–114:11. [Online]. Available http://doi.acm.org/10.1145/2601097.2601193
11. Brochu E, Freitas ND, Ghosh A (2008) Active preference learning with discrete choice data. In: Advances in neural information processing systems, pp 409–416
12. Ashikhmin M, Shirley P (2000) An anisotropic phong brdf model. J Graph Tools 5(2):25–32
13. Höpe A, Koo A, Forthmann C, Verdu F, Manoocheri F, Leloup F, Obein G, Wübbeler G, Ged G, Campos J et al (2014) xd-reflect-" multidimensional reflectometry for industry" a research project of the european metrology research program (emrp). In: 12th International conference on new developments and applications in optical radiometry (NEWRAD 2014)
14. Langovoy M, Wübbeler G, Elster C (2014) Statistical analysis of brdf data for computer graphics and metrology. In: 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 785–790
15. Langovoy M (2015) Machine learning and statistical analysis for brdf data from computer graphics and multidimensional reflectometry. IAENG Int J Comput Sci 42(1):23–30
16. Subr K, Arvo J (2007) Statistical hypothesis testing for assessing monte carlo estimators: applications to image synthesis. In: 15th Pacific conference on computer graphics and applications, 2007. PG '07, pp. 106–115
17. Ferrero A, Rabal AM, Campos J, Pons A, Hernanz ML (2012) Spectral and geometrical variation of the bidirectional reflectance distribution function of diffuse reflectance standards. Appl Opt 51(36):8535–8540
18. Pinto CT, Ponzoni FJ, de Castro RM, Griffith DJ (2012) Spectral uniformity evaluation of reference surfaces for airborne and orbital sensors absolute calibration. Rev Bras Geofis 30
19. Greenberg DP, Torrance KE, Shirley P, Arvo J, Lafortune E, Ferwerda JA, Walter B, Trumbore B, Pattanaik S, Foo S-C (1997) A framework for realistic image synthesis. In: Proceedings of the 24th annual conference on computer graphics and interactive techniques, ACM Press/Addison-Wesley Publishing Co., pp 477–494
20. Arvo J, Torrance K, Smits B (1994) A framework for the analysis of error in global illumination algorithms. In: Proceedings of the 21st annual conference on computer graphics and interactive techniques, ACM, pp 75–84
21. Langovoy M, Wübbeler G, Elster C (2014) Novel metric for analysis, interpretation and visualization of BRDF data. Submitted
22. Nicodemus FE (1965) Directional reflectance and emissivity of an opaque surface. Appl Opt 4:767–775
23. Tieu K, Miller EG (2002) Unsupervised color constancy. In: Advances in neural information processing systems, pp 1303–1310
24. Rosenberg C, Ladsariya A, Minka T (2003) Bayesian color constancy with non-gaussian models. In: Advances in neural information processing systems
25. Wang S, Schwing A, Urtasun R (2014) Efficient inference of continuous markov random fields with polynomial potentials. In: Advances in neural information processing systems, pp 936–944
26. Freeman WT, Viola PA (1998) Bayesian model of surface perception. Adv Neural Inf Process Syst, vol 10
27. Li Y-M, Chen J, Qing L-Y, Yin B-C, Gao W (2004) Face detection under variable lighting based on resample by face relighting. In: Proceedings of 2004 international conference on machine learning and cybernetics, vol 6. IEEE, pp 3775–3780

28. Höpe A, Hauer K-O (2010) Three-dimensional appearance characterization of diffuse standard reflection materials. Metrologia 47(3):295. [Online]. Available http://stacks.iop.org/0026-1394/47/i=3/a=021
29. Simonot L (2009) Photometric model of diffuse surfaces described as a distribution of interfaced lambertian facets. Appl Opt 48(30):5793–5801
30. Fairbairn MB (2005) Planetary photometry: the Lommel-Seeliger law. J Roy Astron Soc Can 99:92
31. Meintanis S (1998) Moment-type estimation for positive stable laws with applications. IAENG Int J Appl Math 38:26–29
32. El Emary IM, Ramakrishnan S (2010) A critical review of statistical modeling of digital images. IAENG Int J Comput Sci 37(1):99–109
33. Ouarets S, Leroux T, Rougie B, Razet A, Obein G (2013) A high resolution set up devoted to the measurement of the bidirectional reflectance distribution function around the specular peak, at lne-cnam. In: 16th International Congress of Metrology. EDP Sciences, p 14008
34. Obein G, Ouarets S, Ged G (2014) Evaluation of the shape of the specular peak for high glossy surfaces. In: IS&T/SPIE electronic imaging, International Society for Optics and Photonics, p 901805
35. Rukhin AL (1996) On the pitman closeness criterion from the decision-theoretic point of view. Stat Decisions-Int J Stochast Methods Models 14(3):253–274
36. Robert CP, Hwang JG, Strawderman WE (1993) Is pitman closeness a reasonable criterion? J Am Stat Assoc 88(421):57–63
37. Sen PK, Kubokawa T, Saleh AME et al (1989) The stein paradox in the sense of the pitman measure of closeness. Ann Stat 17(3):1375–1386
38. Kourouklis S (1996) Improved estimation under pitman's measure of closeness. Ann Inst Stat Math 48(3):509–518
39. Keating JP, Mason RL, Sen PK (1993) Pitman's measure of closeness: a comparison of statistical estimators, vol 37. SIAM, Philadelphia
40. Savage CL (1972) The foundations of statistics. Dover Publications [Online]. Available http://books.google.de/books?id=UW5dAAAAIAAJ
41. Ibragimov IA, Khasminskii RZ (1981) Statistical estimation–asymptotic theory. Springer, New York
42. Borokov A (1999) Mathematical statistics. Taylor & Francis, [Online]. Available http://books.google.de/books?id=2CoUP7qRUxwC
43. Rothenberg TJ, Fisher FM, Tilanus CB (1964) A note on estimation from a cauchy sample. J Am Stat Assoc 59(306):460–463
44. Bloch D (1966) A note on the estimation of the location parameter of the cauchy distribution. J Am Stat Assoc 61(315):852–855
45. Sarkar SK (1998) Some probability inequalities for ordered mtp2 random variables: a proof of the simes conjecture. Ann Stat 26(2):494–504

On the Efficiency of Second-Order *d*-Dimensional Product Kernels

F.O. Oyegue, S.M. Ogbonmwan and V.U. Ekhosuehi

Abstract This study considers the efficiency of second-order product kernels. The univariate form wherein the product kernels are derived is the symmetric beta kernels. We develop a formula for the efficiency using the Epanechnikov kernel as the optimum kernel based on the fundamentals of the Asymptotic Mean Integrated Square Error (AMISE). The results reveal that the efficiency of the product kernels decreases as their dimension increases and that the product form of the univariate biweight kernel has the highest efficiency value among the beta kernels.

Keywords Asymptotic mean integrated square error · Bandwidth · Bias · Density estimation · *d*–dimensional space · Efficiency · Multivariate distribution · Product kernels

1 Introduction

This study is concerned with deriving, using the fundamentals of the Asymptotic Mean Integrated Squared Error (AMISE), a formula for the efficiency of product kernels. The AMISE is a measure of discrepancy between the estimated and the true density in kernel density estimation. This measure is used to quantify the performance of the estimator. It is necessary to study the efficiency of kernels for it enables one to choose an appropriate kernel, especially in the multivariate setting. We obtain

F.O. Oyegue (✉) · S.M. Ogbonmwan · V.U. Ekhosuehi
University of Benin, Benin City, Nigeria
e-mail: fooyegue@yahoo.com

S.M. Ogbonmwan
e-mail: ogbonmwasmaltra@yahoo.co.uk

V.U. Ekhosuehi
e-mail: virtue.ekhosuehi@uniben.edu

H.K. Kim et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-94-017-7236-5_3

a formula for the efficiency by taking the ratio of the product (multivariate) kernels to the Epanechnikov kernel. The Epanechnikov kernel form the basis for the optimum kernel. The product kernels are obtained from the symmetric beta kernels.

Density estimation is simply the construction of an estimate, $\hat{f}$, of an underlying density function, f, for a random variable, X, drawn from an observed data set. To estimate an unknown density, we use either the parametric or the nonparametric methods. The parametric methods such as the maximum likelihood method require the imposition of a functional form on an unknown density. This leads to the problem of the estimation of the parameters. Sometimes, when the density estimation is unknown and no additional information about the distribution is given, then the nonparametric density estimation, like the histogram or the kernel estimator, is applied. This approach allows the data to speak for itself. Instead of the imposition of restrictive parametric assumptions about the underlying distribution, the nonparametric methods allow one to directly approximate the $d-$dimensional density that describes how variables interact [12]. The nonparametric methods are flexible and computationally intensive. The trauma associated with the tedious computations in the nonparametric approach has been considerably reduced via the advent of easily fast computing power in the twentieth century [6]. In this work, we concentrate on one class of nonparametric density estimators, namely, the kernel density estimator. The kernel density estimator is a more reliable statistical technique that deals with some of the problems associated with the histogram which are discussed in [2, 10, 21]. In recent times, kernel density estimation has found relevance in huge computational requirement for large-scale analysis [15, 24], and in the area of human motion tracking or pattern recognition [3, 19].

A common term in kernel density estimation is the bandwidth or window width which is analogous to the bin width in histogram. The bandwidth determines how much smoothing is done. Generally, a narrow bandwidth implies that more points are allowed and this lead to a better density estimate. The technique for determining a bandwith, often called the Parzen density estimation, abounds in the literature [1, 7, 17, 18]. For a d-variate random variable, $X_1, \ldots, X_d$, drawn from a density, f, the generalized kernel estimation is given as [5]:

$$\hat{f}_{\mathrm{H}}(\mathbf{x}) = \frac{1}{n \det H} \sum_{i=1}^{n} K(H^{-1}(\mathbf{x} - \mathrm{X}_i)), \tag{1}$$

where $\mathrm{x} = (x_1, \ldots, x_d)^T$, $\mathrm{X}_i = (x_{i1}, \ldots, x_{id})^T$, $i = 1, \ldots, n$, and $K(\bullet)$ is a d dimensional kernel. This kernel is assumed to be a product (multiplicative) symmetric probability density function. The scope of this study is limited to the multivariate d dimensional kernels that are independent and supported on a region R^d. H is a symmetric and positive definite bandwidth matrix. The scaled and unscaled kernels are related by $K_H(\mathrm{x}) = |H|^{-\frac{1}{2}} K(H^{-\frac{1}{2}}\mathrm{x})$ [22].

An equal bandwidth h in all directions as in Eq. (1) corresponds to $H = h^2 I_d$, where I_d is the $d \times d$ identity matrix [6]. This leads to the expression [4, 14]

$$\hat{f}(\mathbf{x}) = \frac{1}{nh^d} \sum_{i=1}^{n} \prod_{j=1}^{d} K\left(\frac{x - x_{ij}}{h}\right), \tag{2}$$

where $\prod_{j=1}^{d} K\left(\frac{x-x_{ij}}{h}\right)$ is the product kernel of the same univariate kernels. To use the parameterization $H = h^2 I_d$ effectively, the components of the data vector should be commensurate. This can be achieved by using appropriate transformation in the data set [6, 21, 23]. The transformation involves either pre-scaling each axis (that is, normalize to unit variance, for instance) or pre-whitening the data (that is, linearly transform to have unit covariance matrix). A detailed study of this can be found in [8]. The transformation guarantees the use of the form involving single bandwidth as in Eq. (2). Many of the studies in density estimation have been centered on the univariate kernel density estimators [21]. However, this paper focuses on the multivariate settings with emphasis on the efficiency of some classical product (multivariate) kernels. The concept of efficiency is used in kernel density estimation to analyze the effect of second-order multivariate kernels so that an appropriate kernel can be chosen. The basic motivation for considering Eq. (2) is that it enables one to obtain closed form expressions for the optimal bandwidth and the asymptotic mean integrated squared error (AMISE). Thus we derive the the generalized expression for the efficiency of second-order multivariate symmetric kernel. Throughout this paper, $\int$ is the shorthand for $\int_{R^d}$. The global accuracy used in measuring Eq. (2) is the mean integrated square error (MISE). The expression for the MISE is

$$MISE(h) = E \int \left(\hat{f}(\mathbf{x}) - f(\mathbf{x})\right)^2 d\mathbf{x}. \tag{3}$$

Thus, from [22], the expression (3) can be written as a sum of integrated square bias and integrated variance of $\hat{f}_H(x)$. That is,

$$MISE\left\{\hat{f}_H(\mathbf{x})\right\} = \int \left(E\hat{f}_H(\mathbf{x}) - f(\mathbf{x})\right)^2 d\mathbf{x} + \int var\,\hat{f}_H(\mathbf{x}) d\mathbf{x}. \tag{4}$$

The concept of efficiency for univariate kernels was popularized by Silverman [21], and this was followed by the work of Wand and Jones [22], who gave an insight into the efficiency of the second-order multivariate kernels. The approach in [22] was based on taking the ratio of the spherically symmetric kernel relative to the product kernel. This study develops a method that is different from the approach adopted by [22], even though our method is motivated by the earlier works [21, 22].

The remainder of this paper is as follows. In Sect. 2, we cover the necessary background materials on the asymptotic mean integrated square error (AMISE). In Sect. 3, the generalized expression for the efficiency of second-order multivariate kernels is derived. In Sect. 4, we compare the efficiencies of multivariate kernels for the cases $d = 2, 3, 4, 5$. Section 5 concludes the paper.

2 The AMISE for the Multivariate Kernel Density Estimator

The asymptotic mean integrated square error (AMISE) is one of the most important parts in bandwidth selection. By using symmetric kernel functions, the AMISE and the optimal bandwidth for the multivariate kernel density estimator are derived. The kernel determines the slope of the estimator, while the amount of smoothing is determine by the bandwidth h. In particular, $\hat{f}_H(\mathbf{x})$ as define in (1) is a density function provided $K(\mathbf{w}) \geq 0$ and $\int K(\mathbf{w})d\mathbf{w} = 1$, where $\mathbf{w} = H^{-1}(\mathbf{x} - \mathbf{y})$ [3]. For the case $d > 1$ the most often used choice is a density function, which is symmetric about zero, and such that [20]:

$$\begin{aligned} \int \mathbf{w}K(\mathbf{w})d\mathbf{w} &= \mathbf{0}_d, \\ \int \mathbf{w}\mathbf{w}^T K(\mathbf{w})d\mathbf{w} &= \mu_2 \mathbf{I}_d, \end{aligned} \tag{5}$$

where $\mathbf{0}_d$ is a $d \times 1$ vector of zeros and μ_2 is the second moment of the random variable. The usual criterion for the optimal bandwidth is the asymptotic version of the MISE in (3) [5, 9, 11, 13, 21]. To find the AMISE, one needs to find the bias and variance of $\hat{f}_H(\mathbf{x})$. Consider

$$E\big(\hat{f}_H(\mathbf{x})\big) = \frac{1}{|H|}\int K_H(\mathbf{x} - \mathbf{y})f(\mathbf{y})d\mathbf{y}.$$

$E\big(\hat{f}_H(\mathbf{x})\big)$ is evaluated using the Taylor series expansion as [22]:

$$E\big(\hat{f}_H(\mathbf{x})\big) = \int K(\mathbf{w})\Big(f(\mathbf{x}) - tr\Big(H^{\frac{1}{2}}Df(\mathbf{x})\mathbf{w}^T + tr\Big(H^{\frac{1}{2}}D^2 f(\mathbf{x})f(\mathbf{x})H^{\frac{1}{2}}\mathbf{w}\mathbf{w}^T\Big)\Big)\Big). \tag{6}$$

The expansion in (6) is to the second-order. Imposing the conditions (5) and $\int K(\mathbf{w})d\mathbf{w} = 1$ on (6), results in

$$E\big(\hat{f}_H(\mathbf{x})\big) = f(\mathbf{x}) + \frac{1}{2!}\mu_2(K)tr\Big(H^{\frac{1}{2}}D^2 f(\mathbf{x})H^{\frac{1}{2}}\Big),$$

where

$$\mu_2(K) = \int \mathrm{x}_i^2 K^p(\mathbf{x})d\mathbf{x}, \;\; D^2 f(\mathbf{x}) = \sum_{i=1}^{d} \frac{\partial^2 f(\mathbf{x})}{\partial x_i^2}.$$

The bias term is

$$\text{Bias}\left(\hat{f}_H(\mathbf{x})\right) = \frac{1}{2!}\mu_2(K)tr\left(H^{\frac{1}{2}}D^2f(\mathbf{x})H^{\frac{1}{2}}\right),$$

and the asymptotic integrated square bias (AISB) becomes

$$\text{AISB}\left(\hat{f}_H(\mathbf{x})\right) \cong \int \text{Bias}^2\left(\hat{f}_H(\mathbf{x})\right)d\mathbf{x} \cong \frac{1}{(2!)^2}\mu_2(K)^2\int tr\left(H^{\frac{1}{2}}D^2f(\mathbf{x})H^{\frac{1}{2}}\right)^2 d\mathbf{x}. \quad (7)$$

The variance is given as

$$\text{var}\left(\hat{f}_H(\mathbf{x})\right) = \frac{f(\mathbf{x})}{n|H|^{\frac{1}{2}}}\int K(\mathbf{w})^2 d\mathbf{w},$$

and the asymptotic integrated variance (AIV) is

$$\text{AIV}\left(\hat{f}_H(\mathbf{x})\right) = \frac{R(K)}{n|H|^{\frac{1}{2}}}, \quad (8)$$

where

$$R(K) = \int K(\mathbf{w})^2 d\text{w}.$$

Combining (7) and (8) yield

$$\text{AMISE}\left(\hat{f}_H(\mathbf{x})\right) = \frac{R(K)}{n|H|^{\frac{1}{2}}} + \frac{1}{(2!)^2}\mu_2(K)^2\int tr\left(H^{\frac{1}{2}}D^2f(\mathbf{x})H^{\frac{1}{2}}\right)^2 d\mathbf{x}.$$

Since $H = h^2I_d$, it results to

$$\text{AMISE}\left(\hat{f}_H(\mathbf{x})\right) = \frac{R(K)}{nh^d} + \frac{1}{(2!)^2}\mu_2(K)^2h^4\int\left(\nabla^2f(\mathbf{x})\right)^2 d\mathbf{x}. \quad (9)$$

Minimizing (9) with respect to h leads to the formula for the optimal bandwidth of the form

$$h_{opt} = \left(\frac{dn^{-1}R(K)}{\mu_2(K)R(\nabla^2f(\mathbf{x}))}\right)^{\frac{1}{d+4}}, \quad (10)$$

where

$$R(\nabla^2f(\mathbf{x})) = \int(\nabla^2f(\mathbf{x}))^2 d\mathbf{x} < \infty.$$

Putting (10) into (9), the minimum AMISE is obtained as:

$$\text{AMISE}\left(\hat{f}_H(\mathbf{x})\right) = \left(\frac{d+4}{4d}\right)\left[\mu_2(K)^{2d}(dR(K)^4)\left(\int \left(\nabla^2 f(\mathrm{x})\right)^2 d\mathbf{x}\right)^d n^{-4}\right]^{\frac{1}{d+4}}. \quad (11)$$

Equation (10) is a closed form solution for the bandwidth vector which minimizes the expression for the AMISE in (11). Moreover, the optimal bandwidth is of order $n^{-\frac{1}{d+4}}$ and the optimal AMISE is of order $n^{-\frac{4}{d+4}}$.

3 Efficiency of the Second-Order Multivariate Kernels

In this section, the AMISE expression derived in Sect. 2 is used to develop the generalized expression for the efficiency of second order multivariate kernels. We adopt the multivariate product kernel form in [22] given as:

$$K^p(\mathbf{x}) = \prod_{i=1}^{d} K(x_i), \quad (12)$$

where $K(x_i)$ is the univariate symmetric kernel. The efficiency of the univariate symmetric kernel defined by [21] is

$$\text{Eff}\,(K) = \left(\frac{C(K_e)}{C(K)}\right)^{\frac{5}{4}}, \quad (13)$$

where

$$C(K) = \left(\int x^2 K(x)dx\right)^{\frac{2}{5}}\left(\int K(x)^2 dx\right)^{\frac{4}{5}}$$

is any given kernel constant under discussion and

$$C(K_e) = \left(\int x^2 K_e(x)dx\right)^{\frac{2}{5}}\left(\int K_e(x)^2 dx\right)^{\frac{4}{5}}$$

is the Epanechnikov kernel constant. We adopt the definition in [21]. Thus we define the general expression for the efficiency of multivariate kernels based on the product kernel as [16]

$$\text{Eff}\,(K^p(\mathbf{x})) = \left(\frac{C_d^2(K_e^p)}{C_d^2(K^p)}\right)^{\frac{d+4}{4}}, \quad (14)$$

where

$$C_d^2(K^p) = \left(\int x_1^2 K^p(\mathbf{x})d\mathbf{x}\right)^{\frac{2d}{d+4}}\left(\int K^p(\mathbf{x})^2 d\mathbf{x}\right)^{\frac{4}{d+4}}$$

is the $d-$dimensional product form of any given second order kernel constant and

$$C_d^2(K_e^p) = \left(\int x_1^2 K_e^p(\mathbf{x})d\mathbf{x}\right)^{\frac{2d}{d+4}}\left(\int K_e^p(\mathbf{x})^2 d\mathbf{x}\right)^{\frac{4}{d+4}}$$

is the $d-$dimensional product form of the Epanechnikov kernel constant.

Theorem 1 *Given the Epanechnikov kernel of the form*

$$K(x_i) = \frac{3}{4\sqrt{5}}\left(1 - \frac{x_i^2}{5}\right), \quad -\sqrt{5} \le x_i \le \sqrt{5},$$

then the efficiency for the second-order $d-$ dimensional kernel is

$$Eff\,(K^p(x)) = \left(\frac{3}{5\sqrt{5}}\right)^d \left(\int x_1^2 K^p(x)dx\right)^{-\frac{d}{2}}\left(\int K^p(x)^2 dx\right)^{-1}.$$

Proof By the definition of the product kernel (12), the multivariate version of the Epanechnikov is

$$K_e^p(\mathbf{x}) = \prod_{i=1}^{d} \frac{3}{4\sqrt{5}}\left(1 - \frac{x_i^2}{5}\right).$$

From (11), set

$$C_d^2(K) = \mu_2(K)^{\frac{2d}{d+4}} R(K)^{\frac{4}{d+4}} = \left(\int x_1^2 K(\mathbf{x})d\mathbf{x}\right)^{\frac{2d}{d+4}}\left(\int K(\mathbf{x})^2 d\mathbf{x}\right)^{\frac{4}{d+4}}. \tag{15}$$

Re-writing Eq. (15) to reflect (14), we obtain

$$C_d^2(K_e^p) = \left(\int x_1^2 K_e^p(\mathbf{x})d\mathbf{x}\right)^{\frac{2d}{d+4}}\left(\int K_e^p(\mathbf{x})^2 d\mathbf{x}\right)^{\frac{4}{d+4}}$$

and

$$C_d^2(K^p) = \left(\int x_1^2 K^p(\mathbf{x})d\mathbf{x}\right)^{\frac{2d}{d+4}}\left(\int K^p(\mathbf{x})^2 d\mathbf{x}\right)^{\frac{4}{d+4}}.$$

Thus

$$\mu_2(K_e^p) = \int x_1^2 K_e^p(\mathbf{x})d\mathbf{x} = \int x_1^2 \left(\prod_{i=1}^{d} \frac{3(5-x_i^2)}{20\sqrt{5}} \right) d\mathbf{x} = 1.$$

and

$$R(K_e^p) = \int K_e^p(\mathbf{x})^2 d\mathbf{x} = \int \left(\prod_{i=1}^{d} \frac{3(5-x_i^2)}{20\sqrt{5}} \right)^2 d\mathbf{x} = \left(\frac{3}{5\sqrt{5}} \right)^d.$$

Putting the values of $\mu_2(K_e^p)$ and $R(K_e^p)$ into (14) we get

$$\text{Eff}\,(K^p(\mathbf{x})) = \left(\frac{\left[\left(\frac{3}{5\sqrt{5}} \right)^d \right]^{\frac{4}{d+4}}}{\left[\left(\int x_1^2 K^p(\mathbf{x})d\mathbf{x} \right)^{2d} \left(\int K^p(\mathbf{x})^2 d\mathbf{x} \right) \right]^{\frac{1}{d+4}}} \right)^{\frac{d+4}{4}}.$$

Hence,

$$\text{Eff}\,(K^p(\mathbf{x})) = \left(\frac{3}{5\sqrt{5}} \right)^d \left(\int x_1^2 K^p(\mathbf{x})d\mathbf{x} \right)^{-\frac{d}{2}} \left(\int K^p(\mathbf{x})^2 d\mathbf{x} \right)^{-1}. \tag{16}$$

□

Equation (16) is the closed form expression of the efficiency for the second-order $d-$dimensional kernel.

4 Illustrative Example

We illustrate the use of the efficiency formula in Eq. (16) for some product kernels derived from the following univariate beta kernels: the uniform, the biweight, the triweight, and the Gaussian kernel. We consider the case where $d = 2, 3, 4, 5$. The Wolfram mathematica 6.0 is used for the computation of the efficiency values for each d. The results are presented in Table 1. The following can be inferred from Table 1. For $d = 2$, there is a 14 % loss in efficiency for the uniform kernel, the Gaussian lost about 10 %, the biweight and the triweight lost about 1 and 3 %, respectively. For $d = 3, 4, 5$, the uniform kernel lost about 20, 25, 31 % in efficiency, respectively; the biweight and the triweight respectively lost approximately 2, 2, 3 and 4, 5, 6 %. There is a loss of about 14, 18 and 22 % in the case of the Guassian kernel for $d = 3, 4, 5$. Furthermore, a comparison of the dimensions of the four beta kernels shows that the biweight and the triweight kernels give relatively better efficiency values than the uniform and the Gaussian kernels. This is visible in

Table 1 Efficiency of some d-dimensional product kernels

d	Uniform	Biweight	Triweight	Gaussian
2	0.8640	0.9878	0.9735	0.9048
3	0.8031	0.9818	0.9606	0.8606
4	0.7465	0.9758	0.9478	0.8186
5	0.6939	0.9699	0.9352	0.7787

Table 1 as there is a slight drop in the efficiency values of both the biweight and the triweight kernels. Nonetheless, the efficiency values of the biweight kernels are higher than that of the triweight. Thus, we conclude that the second-order product kernel derived from the univariate biweight kernel is the best choice among the product kernels.

5 Concluding Remarks

In this paper, a computational approach has been developed for the efficiency of multivariate product kernels. The Epanechnikov kernel was used as a theoretical underpinning for deriving the efficiency formula. The efficiency formula was experimented with four of the beta kernels, viz.: the Gaussian, the uniform, the biweight, and the triweight kernels. Findings revealed that the product form of the biweight and the triweight kernels have relatively high efficiency values. By this, we infer that they are better density estimators than the Gaussian and the uniform kernels form of the multivariate product kernels. Nevertheless, it is premature to conclude that the biweight and the triweight kernels are the most suitable multivariate product kernels. This is because the spherical aspects of the multivariate kernels have not been considered. We therefore suggest the development of a theoretical framework for the efficiency of multivariate kernels using the spherical methods as a grey area for future research.

References

1. Akaiki BA (1954) An approximation to the density function. Ann Inst Stat Math 6:127–132
2. Bowman AW, Azzalini A (1997) Applied smoothing technique for data analysis: the kernel approach with s-plus illustration. Oxford University Press, Oxford
3. Brox T, Rosenhahn B, Kersting U, Cremers D (2006) Nonparametric density estimation for human pose tracking. Lect Notes Comput Sci Pattern Recognit 4174:546–555
4. Cacoullos T (1966) Estimation of a multivariate density. Ann Inst Stat Math 18:178–189
5. Devroye L, Gjorfi L (1984) Nonparametric density estimation: the L1 view. Wiley, New York
6. Duong T, Hazeltom ML (2005) Convergence rates for unconstrained bandwidth matrix selectors in multivariate kernel density estimator. J Multivar Anal 93:417–433
7. Fix E, Hodges JL (1951) Discriminatory analysis–nonparametric discrimination: consistency properties. USAF School of Aviation Medicine, Randolf Field, Texas, Report Number 4,

Project Number 21-49-004. http://www.dtic.mil/dtic/tr/fulltext/u2/a800276.pdf. Cited Feb 1951
8. Fukunaga K (1990) Introduction to statistical pattern recognition. Academic Press, New York
9. Hall P, Marron JS (1987) Choice of kernel order in density estimation. Ann Stat 16:161–173
10. Hardle W, Muller M, Sperlich S, Werwatz A (2004) Nonparamtric and semiparametric models. Springer, Berlin
11. Jarnicka J (2009) Multivariate kernel density estimation with a parametric support. Opuscula Math 29(1):41–55
12. Lambert CG, Harrington SE, Harvey CE, Glodjo A (1999) Efficient online non-parametric kernel density estimation. Algorithmica 25:37–57
13. Marron JS, Wand MP (1992) Exact mean integrated squared error. Ann Stat 20(2):712–736
14. Martinez WL, Martinez AR (2002) Computational statistics handbook with MATLAB. Chapman & Hall/CRC, Florida
15. Michailidis PD, Margritis KG (2013) Accelerating kernel density estimation on the GPU using the CUDA framework. Appl Math Sci 7(30):1447–1476
16. Oyegue FO, Ogbonmwan SM (2014) The efficiency of product multivariate kernels. Lecture notes in engineering and computer science. In: Proceedings of the world congress on engineering and computer sciences, WCECS 2014. San Francisco, pp 830–834 22–24 Oct 2014
17. Parzen E (1962) On the estimation of a probability density function and the mode. Ann Math Stat 33(3):1056–1076
18. Rosenblatt M (1956) Remarks on some nonparametric estimate of a density function. Ann Math Stat 27(3):832–837
19. Rosehahn B, Brox T, Weikert J (2007) Three dimensional shape knowledge for joint image segmentation and pose tracking. Int J Comput Vis 73(3):243–262
20. Scott DW (1992) Multivariate density estimation: theory practice and visualization. Wiley, New York
21. Silverman BW (1986) Density estimation for statistics and data analysis. Chapman & Hall, London
22. Wand MP, Jones MC (1995) Kernel smoothing. Chapman & Hall, London
23. Wu TJ, Chen CF, Chen HYA (2007) Variable bandwidth selector in multivariate kernel density estimation. Stat Probab Lett 77(4):462–467
24. Zheng Y, Jestes J, Phillips JM, Li F (2013) Quality and efficiency in kernel density estimation for large data. In: Proceedings of the ACM SIGMOD international conference on management of data. ACM, New York, pp 433–444

Comparing the Markov Order Estimators AIC, BIC and EDC

Chang C.Y. Dorea, Paulo A.A. Resende and Catia R. Gonçalves

Abstract In the framework of nested hypotheses testing, several alternatives for estimating the order of a Markov chain have been proposed. The AIC, Akaike's entropy-based information criterion, constitutes the best known tool for model identification and has had a fundamental impact in statistical model selection. In spite of the AIC's relevance, several authors have pointed out its inconsistency that may lead to overestimation of the true order. To overcome this inconsistency, the Bayesian information criterion, BIC, was proposed by introducing in the penalty term the sample size that led to a consistent estimator for large samples. A more general approach is exhibited by the EDC, efficient determination criterion, that encompass both AIC and BIC estimates. Under proper setting, the EDC, besides being a strongly consistent estimate, is an optimal estimator. These approaches are briefly presented and compared by numerical simulation. The presented results may support decisions related to estimator's choice.

Keywords AIC · BIC · EDC · Markov chain order · Optimal EDC · Penalty term

1 Introduction

Higher order Markov chains, by its very definition, is the most flexible model for finitely dependent sequences of random variables. In practical settings, estimation of the dependency order is needed to identify other model parameters. Based on the

C.C.Y. Dorea (✉) · C.R. Gonçalves
Universidade de Brasilia, Brasilia, DF 70910-900, Brazil
e-mail: changdorea@gmail.com; changdorea@unb.br

C.R. Gonçalves
e-mail: catia.unb@gmail.com; catiarg@unb.br

P.A.A. Resende
Presidencia da Republica, Brasilia, DF 70150-900, Brazil
e-mail: pa@pauloangelo.com

H.K. Kim et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-94-017-7236-5_4

penalized log-likelihood function and within nested hypotheses testing framework, several estimation alternatives have been proposed. The Akaike's [1] entropy-based information criterion, AIC, constitutes a very useful tool for model identification and has had a fundamental impact in statistical model selection. The AIC was designed to be an approximately unbiased estimate of the Kullback-Leibler divergence between the fitted model relative to the true model. The fact that when mean log-likelihood ratio is used to estimate the divergence quantity, the bias introduced by the maximum likelihood estimate of the parameters needs to be corrected. For the AIC procedure the correction (penalty) term is taken to be the number of independent parameters of the model. In the case of a multiple Markov chain $X = \{X_n\}_{n \geq 1}$ of unknown order $r < K < \infty$ with finite sate space E we have

$$\mathrm{AIC}(k) = -2 \log \hat{L}(k) + 2m^k(m-1), 0 \leq k \leq K$$

with the AIC estimator given by

$$\hat{r}_{\mathrm{AIC}} = \underset{0 \leq k \leq K}{\arg\min}\ \mathrm{AIC}(k) \tag{1}$$

where $\hat{L}(k)$ is the maximum likelihood estimate assuming that k is the true order and m denotes the cardinality of the set $E = \{1, \ldots, m\}$. In spite of the AIC's relevance, there was no rigorous analysis about its behaviour and showed a tendency of overestimating the true order. In fact, Katz [7] formally derived the asymptotic distribution of AIC estimator and proved its inconsistency for the Markov Chain case, no matter how large the sample size is taken. To overcome this inconsistency, the Bayesian information criterion, BIC, was proposed by Schwarz [10]. The BIC procedure introduces in the penalty term the sample size and it is a consistent estimate

$$\mathrm{BIC}(k) = -2 \log \hat{L}(k) + m^k(m-1) \log n, \quad 0 \leq k \leq K$$

where n is the sample size and define

$$\hat{r}_{\mathrm{BIC}} = \underset{0 \leq k \leq K}{\arg\min}\ \mathrm{BIC}(k). \tag{2}$$

More recently, Csiszár and Shields [3] established the strong consistency for BIC, which strengthens earlier results by removing the assumption that sets up an a priori bound K on the order. Also, Zhao et al. [11] introduced the EDC, efficient determination criterion, that encompass both the AIC and the BIC criteria

$$\mathrm{EDC}(k) = -2 \log \hat{L}(k) + \gamma(k) c_n \tag{3}$$

where $\gamma(\cdot)$ is a positive and strictly increasing function and $c_n \geq 0$. It is shown that $\hat{r}_{\mathrm{EDC}} = \underset{0 \leq k \leq K}{\arg\min}\ \mathrm{EDC}(k)$ is strongly consistent provided $c_n > 0$ satisfies

$$\frac{c_n}{\log\log n} \xrightarrow[n]{[}]\infty \quad \text{and} \quad \frac{c_n}{n} \xrightarrow[n]{[}]0. \tag{4}$$

The rates of convergence from Dorea and Zhao [5] and the results from Dorea [4] indicate that with a proper choice of $\gamma(\cdot)$ and c_n one should expect that

$$\hat{r}_{\text{BIC}} \leq \hat{r}_{\text{EDC}} \leq \hat{r}_{\text{AIC}}.$$

Moreover, it is shown that under regularity conditions the optimal choice is given by

$$\text{EDC}_{\text{opt}}(k) = -2\log\hat{L}(k) + 2m^{k+1}\log\log n. \tag{5}$$

and we redefine $\hat{r}_{\text{EDC}}$ as

$$\hat{r}_{\text{EDC}} = \underset{k\geq 0}{\arg\min}\ \text{EDC}_{\text{opt}}(k). \tag{6}$$

With more than one alternative to estimate r it is natural to seek comparison among them. Katz [7] presented some modest numerical simulation, supported by computational resources available at the time, to compare $\hat{r}_{\text{AIC}}$ and $\hat{r}_{\text{BIC}}$. In this paper, we analyse the comparative performance of $\hat{r}_{\text{AIC}}$, $\hat{r}_{\text{BIC}}$ and $\hat{r}_{\text{opt}}$. Altogether 63 cases were studied by ranging $m = 2, \ldots, 10$ and $r = 0, \ldots, 6$. For each case 1,000 models were generated and for each model, 349 samples. Since both $\hat{r}_{\text{BIC}}$ and $\hat{r}_{\text{opt}}$ possess the same asymptotic behavior, the sample size n also played an important role in our analysis. The sample sizes were taken from $n = 10$ up to $n = 10^8$. Our findings show that, in general, $\hat{r}_{\text{opt}}$ outperforms $\hat{r}_{\text{BIC}}$. For small samples, all considered estimators have a tendency to underestimate the true order of the chain. Contrary to what Katz implicitly suggested in his numerical simulations, the probability of overestimation for $\hat{r}_{\text{AIC}}$ can be negligible in the case of complex models. These results may support the choice of which estimator to use in real situations.

In Sect. 2 some core known results on these estimators are presented. The simulation methodology and results are detailed in Sect. 3. And, based on the empirical evidences, Sects. 3.3 and 3.4 gathers the concluding remarks as well as some theoretical motivation of the estimators behavior for small and relatively large samples.

2 Background Theory

Let $X_1^n = (X_1, \ldots, X_n)$ be a sample from a multiple Markov chain $X = \{X_n\}_{n\geq 1}$ of unknown order r. Assume that X takes value on a finite state space $E = \{1, \ldots, m\}$ and that the transition matrix P has probabilities given by

$$p(a_{r+1}|a_1^r) = P(X_{n+1} = a_{r+1}|X_{n-r+1}^n = a_1^r) \tag{7}$$

where $a_1^r = a_1^k a_{k+1}^r = (a_1, \ldots, a_r) \in E^r$.

For estimating the true order r, in the context of nested hypotheses testing, we consider a sequence of statistical models $\{\mathscr{M}_k\}_{k\in N}$ with $.\Theta_k \subset \Theta_{k+1}, \forall k \in \mathscr{N}$, where Θ_k is the parameter set for $\mathscr{M}_k$ and so on. For a Markov chain of order k the model $\mathscr{M}_k$ can be determined by $\eta(k) = m^{k-1}(m-1)$ parameters, precisely, the transition probabilities $\{p(a_{k+1}|a_1^k)\,|\,a_1^{k+1} \in E^{k+1}\}$ with the constraint $\sum_{a^{k+1}} p(a_{k+1}|a_1^k) = 1$. Therefore, we can take $\Theta_k \in R^{\eta(k)}$ and $\mathrm{M}_k = \{P_{\theta_k}|\theta_k \in \Theta_k\}$ where P_{θ_k} is a possible transition matrix for a chain of order k. Since a chain of order $s<k$ can always be modelled as a chain of order k our problem reduces to find the smallest k for which the true transition matrix P satisfies $P \in \mathrm{M}_k$. And this can be accomplished by estimating the Kullback-Leibler divergence between the assumed model and the true model. An unbiased estimate can be derived from the log-likelihood ratio corrected by a penalty term. In practical setting, given the observation X_1^n from a chain of order k we have the maximum likelihood

$$\hat{L}(k) = \prod_{a^{k+1}} \hat{p}^{N(a_1^{k+1})}(a_{k+1}|a_1^k) \text{ and } \hat{p}(a_{k+1}|a_1^k) = \frac{N(a_1^{k+1})}{N(a_1^k)}, \tag{8}$$

where $N(a_1^k)$ is the number of occurrences of a_1^k in X_1^n

$$N(a_1^k) = \sum_{j=1}^{n-k+1} 1(X_j = a_1, \ldots, X_{j+k-1} = a_k) \tag{9}$$

and the sums are taken over positive terms $N(a_1^{k+1}) > 0$, or else, we convention $0/0 = 0 \cdot \infty = 0^0 = 0$.

From (3) and assuming $\gamma(k) = m^k(m-1)$ we can derive $\mathrm{AIC}(k)$, $\mathrm{BIC}(k)$ and $\mathrm{EDC}_{\mathrm{opt}}(k)$ by taking $c_n = 2$, $c_n = \log n$ and $c_n = \frac{2m}{m-1}\log\log n$, respectively. It can be shown that the performance of $\hat{r}_{\mathrm{AIC}}$, $\hat{r}_{\mathrm{BIC}}$ and $\hat{r}_{\mathrm{EDC}}$ relies essentially on the behavior of the quantities

$$\frac{1}{\log\log n}\log\frac{\hat{L}(r)}{\hat{L}(k)} \quad \text{and} \quad \frac{1}{n}\log\frac{\hat{L}(r)}{\hat{L}(k)}$$

and the validity of the Law of the Iterated Logarithm for Markov chains. A sharp result can be stated.

Theorem 1 ([4]). *Assume that $m \geq 2$, the associated first order chain $\{Y_n^{(r)} = (X_n, \ldots, X_{n+r-1})\}_{n\geq 1}$ is ergodic and c_n satisfies*

$$\liminf_{n\to\infty} \frac{c_n}{\log\log n} \geq \frac{2m}{m-1} \quad \textit{and} \quad \limsup_{n\to\infty} \frac{c_n}{n} = 0. \tag{10}$$

Then for $\gamma_m(k) = m^k(m-1)$ *we have*

(i) If $k \geq r$ then

$$\lim_{n\to\infty} sup \frac{\log \hat{L}(k) - \log L(k)}{\log\log n} = \gamma_m(r) - \gamma_m(k) \; a.s. \tag{11}$$

(ii) If $0 \leq k < r$ *then there exists* $\delta(k) > 0,\ \delta(k) \geq \delta(k+1)$ *such that*

$$\lim_{n\to\infty} \frac{\log \hat{L}(r) - \log \hat{L}(k)}{n} = \delta(k) \; a.s. \tag{12}$$

(a.s. stands for almost sure convergence).

From the above specific asymptotic behavior of $\log \frac{\hat{L}(r)}{\hat{L}(k)}$: if c_n is bounded then the corresponding estimator is not consistent, that is the case of $\hat{r}_{\mathrm{AIC}}$. Also, over-estimation may occur since for $k > r$ we might have

$$\frac{\mathrm{AIC}(k) - \mathrm{AIC}(r)}{\log\log n} \leq -2\gamma_m(k) + 2\gamma_m(r) < 0. \tag{13}$$

If c_n satisfies (4) then we have a strongly consistent estimator, that is the case of $\hat{r}_{\mathrm{BIC}}$ and $\hat{r}_{\mathrm{EDC}}$. If hypotheses of Theorem 1 are satisfied then we can adjust properly $c_n\gamma_m(k)$ to stay asymptotically between the AIC and BIC cases, leading to the optimal estimator $\hat{r}_{\mathrm{EDC}}$ given by (6). Finally we observe that if $p(a_{r+1}|a_1^r) > 0$ for all $a_1^{r+1} \in E^{r+1}$ then the associated chain $\{Y_n^{(r)}\}$ is ergodic. In the next section we gather simulation results, some of them presented at WCECS2014 (see [6]).

3 Simulation Results

Though the asymptotic behaviors of $\hat{r}_{\mathrm{AIC}}$, $\hat{r}_{\mathrm{BIC}}$ and $\hat{r}_{\mathrm{EDC}}$ are well-known in theory, it is practically impossible to derive the behavior for small samples. Analytical expressions for exact distributions (as a function of the sample size n) are not available. It should drive to huge calculations depending on unknown probabilities that couldn't be approximated by asymptotic distributions.

We considered 63 cases of Markov chains, varying $m = 2, \ldots, 10$ and $r = 0, \ldots, 6$. For each case we randomly generated 1000 transition matrices and for each matrix, one large sample of length 100,000,000 and 349 "sub-samples" by considering the fragmentation from 0 to a properly chosen sample sizes. Using this

technique it is possible to reuse the partial sums $N(a_1^k)$ and achieve a considerable computational gain. From the theoretical point of view, this is a reasonable approximation. The cases were chosen according to the available computational resources. The sizes of "sub-samples" where chosen empirically to properly compare the estimators. Although such numbers do not appears large, the most complex considered case, $m = 10$ and $r = 6$, has $9,000,000 = 10^6 \times (6-1)$ parameters, and the estimators couldn't fit the true order, even for samples of length 100 m. To generate the the transition matrices we implemented the algorithm using the model proposed by Raftery [9]. In this approach, the transition probabilities are defined by

$$p(a_{r+1}|a_1^r) = \sum_{i=1}^{r} \lambda_i Q(a_i, a_{r+1})$$

where Q is a stochastic matrix and $\lambda = (\lambda_1, \ldots, \lambda_r) \in (0,1)^r$. Doing so, we concentrate all dependency information at λ and can produce a better dependency variability. Also, we used the random algorithm suggested by Park and Miller [8].

3.1 EDC Versus BIC

For EDC we will be considering EDC_{opt} and the corresponding estimator $\hat{r}_{\text{EDC}}$ given by (6). For small complexity $\gamma_m(r) = m^r(m-1)$ (number of free parameters), the sample sizes n can be small too. Tables 1 and 2 provide simulation results for the cases $m = 4,\ r = 1$ and $m = 10,\ r = 1$, respectively. The column n is the sample size, the columns "$<$", "$=$" and "$>$" represent respectively the rates of underestimation, fitness and overestimation for each n. For example, consider $\hat{r}_{\text{EDC}}$ and assume that the true order is $r = 1$, then "$<$", "$=$" and "$>$" refers for the obtained rates of the events $\{\hat{r}_{EDC} = 0\}$, $\{\hat{r}_{\text{EDC}} = 1\}$ and $\{\hat{r}_{\text{EDC}} \in \{2,3,4,5,6\}\}$.

As complexity grows, the sample sizes need to be enlarged. Tables 3 and 4 give the cases $m = 4,\ r = 3$ and $m = 5,\ r = 4$.

In all cases, EDC exhibits a better performance than BIC. In the smaller complexity case ($m = 4,\ r = 1$) the fitness rates are similar. However, for more complex cases, EDC needed nearly half steps n, as compared to BIC, to achieve 50 % of fitness. This difference becomes bigger as larger complexities are considered. To access this outperformance, "how much EDC is better than BIC", in Table 5 we provide the needed sample size for EDC and BIC to determine correctly the order for at least 50% of simulated sequences. The last column shows the ratio $\frac{n \text{ for BIC to hit } 50\%}{n \text{ for EDC to hit } 50\%}$. For the not mentioned cases, estimators did not achieve 50 %. Indeed, the differences increase as the complexity grows. It happens because complex models require larger sample sizes that will result in larger differences between the penalty terms. Table 1 shows that the fitness of EDC and BIC are quite similar. In fact, for the very simple and atypical cases, such as $m = 2,\ r \leq 2$ or

Table 1 Distribution of hits ($m = 4$ and $r = 1$)

n	EDC (%)			BIC (%)		
	$<$	$=$	$>$	$<$	$=$	$>$
10	98.7	1.3	0.0	99.1	0.9	0.0
25	90.2	9.8	0.0	91.4	8.6	0.0
68	50.6	49.4	0.0	60.3	39.7	0.0
775	0.0	100.0	0.0	0.1	99.9	0.0
900	0.0	100.0	0.0	0.0	100.0	0.0

Table 2 Distribution of hits ($m = 10$ and $r = 1$)

n	EDC (%)			BIC (%)		
	$<$	$=$	$>$	$<$	$=$	$>$
218	99.8	0.2	0.0	100.0	0.0	0.0
425	40.9	59.1	0.0	100.0	0.0	0.0
450	28.9	71.1	0.0	99.9	0.1	0.0
600	3.1	96.9	0.0	91.1	8.9	0.0
775	0.1	99.9	0.0	48.2	51.8	0.0
950	0.0	100.0	0.0	15.4	84.6	0.0
1812	0.0	100.0	0.0	0.0	100.0	0.0

Table 3 Distribution of hits ($m = 4$ and $r = 3$)

n	EDC (%)			BIC (%)		
	$<$	$=$	$>$	$<$	$=$	$>$
1562	99.9	0.1	0.0	100.0	0.0	0.0
2375	98.8	1.2	0.0	99.9	0.1	0.0
23,125	50.2	49.8	0.0	65.1	34.9	0.0
9,375,000	0.0	100.0	0.0	0.6	99.4	0.0
23,750,000	0.0	100.0	0.0	0.0	100.0	0.0

Table 4 Distribution of hits ($m = 5$ and $r = 4$)

n	EDC (%)			BIC (%)		
	$<$	$=$	$>$	$<$	$=$	$>$
6500	100.0	0.0	0.0	100.0	0.0	0.0
32,500	99.8	0.2	0.0	100.0	0.0	0.0
68,750	93.6	6.4	0.0	99.7	0.3	0.0
600,000	49.8	50.2	0.0	66.4	33.6	0.0
1,437,500	33.5	66.5	0.0	49.9	50.1	0.0
16,875,000	7.0	93.0	0.0	13.8	86.2	0.0
100,000,000	0.0	100.0	0.0	3.4	96.6	0.0

$m = 3$, $r = 1$, our simulations show that BIC performs better than EDC. It occurs because, for not too large sample size the penalty term for BIC is smaller than that for EDC.

Table 5 50 % Fitness for EDC versus BIC

r	m	n_{EDC}	n_{BIC}	Ratio
1	2	76	50	0.65
	3	53	50	0.94
	4	72	85	1.18
	5	100	150	1.50
	6	143	225	1.57
	7	200	337	1.68
	8	250	475	1.90
	9	337	625	1.85
	10	425	775	1.82
2	2	1125	925	0.82
	3	1125	1500	1.33
	4	2000	3250	1.62
	5	3125	5750	1.84
	6	5750	11,250	1.95
	7	8000	16,875	2.10
	8	10,625	23,750	2.23
	9	16,875	40,000	2.37
	10	23,750	62,500	2.63
3	2	10,625	11,250	1.05
	3	10,000	15,625	1.56
	4	23,750	45,000	1.89
	5	47,500	100,000	2.10
	6	93,750	212,500	2.26
	7	162,500	400,000	2.46
	8	293,750	775,000	2.63
	9	525,000	1,437,500	2.73
	10	637,500	1,812,500	2.84
4	2	32,500	37,500	1.15
	3	81,250	137,500	1.69
	4	225,000	475,000	2.11
	5	600,000	1,437,500	2.39
	6	1,562,500	4,250,000	2.72
	7	2,875,000	8,125,000	2.82
	8	4,750,000	13,750,000	2.89
5	2	143,750	175,000	1.21
	3	337,500	650,000	1.92
	4	1,687,500	4,000,000	2.37
	5	5,625,000	15,000,000	2.66
6	2	400,000	525,000	1.31
	3	2,000,000	4,000,000	2.00
	4	11,875,000	28,750,000	2.42

In fact, Table 5 shows that BIC has a higher tendency to underestimate the order. This can be justified by observing that if the hypotheses of Theorem 1 are satisfied then from (12) we have for $k<r$,

$$\frac{\mathrm{BIC}(k)-\mathrm{BIC}(r)}{n}\approx 2\delta(k)+[\gamma_m(k)-\gamma_m(r)]\frac{\log n}{n}. \tag{14}$$

Since for $k<r$ we have $\gamma_m(k)-\gamma_m(r)<0$, it follows that if n is not large enough we would have $\mathrm{BIC}(k)<\mathrm{BIC}(r)$. In (14) if we replace BIC by $\mathrm{EDC}=\mathrm{EDC}_{\mathrm{opt}}$ then the term $\frac{\log n}{n}$ becomes $\frac{2m\ \log\log n}{n(m-1)}$. Underestimation may occur if

$$\frac{2\delta(k)}{\gamma_m(r)-\gamma_m(k)}<\frac{2m\ \log\ \log\ n}{n(m-1)}.$$

Clearly, if $m\geq 3$ and $n\geq 100$ we always have $\frac{2m\log\log n}{m-1}<\log n$. And this gives a theoretical justification for Table 5. From (11) we conclude that BIC and EDC never overestimates the true order.

3.2 AIC's Performance

Despite the inconsistency of AIC and the existence of strong consistent alternatives, this estimator have been widely used. Some comments from Katz [7] deserve a better understanding ("…except for $n=50$, the AIC procedure virtually never underfits…"; "…AIC always has a significant probability to overestimate…"). Thus, we performed some numerical simulation with the aim to analyse AIC's behavior and access its overestimation probabilities. Table 6 presents rates of underfitting, hits and overfitting of hits for EDC, BIC and AIC.

From Table 6 simulation results, we can identify few characteristics for AIC's hit rates: underestimation for tiny samples; a better performance for small samples; and at stable optimal rate for very large n. The sample sizes for this behavior depends heavily on the complexities of the considered cases. More specifically: (i) for small complexity models ($\gamma_m(r)\leq 10$) AIC tends to overestimate r with small probability and performs better than EDC if the sample size is small ($n<200$); and for large sample size ($n>1,000$), EDC performs better than AIC; (ii) for medium or large complexity models ($\gamma_m(r)>20$) overestimation occurs with negligible probability and AIC performs better than EDC up to medium sized sample. For instance, $\gamma_m(r)=100$ and $n<5,000$; $\gamma_m(r)=800$ and $n<500,000$.

Following Katz [7] arguments, we can actually compute the overestimation probability. Assume $r\leq K=7$ then, asymptotically,

Table 6 Distribution of hits for EDC, BIC and AIC (in %)

r	m	n	EDC (%)			BIC (%)			AIC (%)		
			$<$	$=$	$>$	$<$	$=$	$>$	$<$	$=$	$>$
1	3	10	80.30	19.70	0.00	73.90	26.10	0.00	63.10	36.90	0.00
		22	72.20	27.80	0.00	67.40	32.60	0.00	39.80	59.90	0.30
		168	15.00	85.00	0.00	15.80	84.20	0.00	3.30	93.50	3.20
		1375	0.60	99.40	0.00	0.80	99.20	0.00	0.10	96.20	3.70
		3125	0.00	100.00	0.00	0.10	99.90	0.00	0.00	96.20	3.80
		5000	0.00	100.00	0.00	0.00	100.00	0.00	0.00	97.10	2.90
1	4	10	98.70	1.30	0.00	99.10	0.90	0.00	96.10	3.90	0.00
		131	18.40	81.60	0.00	27.50	72.50	0.00	1.60	98.40	0.00
		212	7.30	92.70	0.00	12.20	87.80	0.00	0.20	99.70	0.10
		975	0.00	100.00	0.00	0.00	100.00	0.00	0.00	99.90	0.10
2	3	17	100.00	0.00	0.00	100.00	0.00	0.00	99.90	0.10	0.00
		137	96.50	3.50	0.00	97.30	2.70	0.00	58.59	41.30	0.10
		175,000	1.70	98.30	0.00	3.50	96.50	0.00	0.10	99.80	0.10
		4,750,000	0.00	100.00	0.00	0.00	100.00	0.00	0.00	99.90	0.10
2	5	137	100.00	0.00	0.00	100.00	0.00	0.00	99.70	0.30	0.00
		400	99.90	0.10	0.00	100.00	0.00	0.00	67.00	33.00	0.00
		650	99.00	1.00	0.00	100.00	0.00	0.00	50.10	49.90	0.00
		750	97.00	3.00	0.00	99.90	0.10	0.00	47.00	53.00	0.00
		3125	49.90	50.10	0.00	68.10	31.90	0.00	20.90	79.10	0.00
		6000	35.80	64.20	0.00	49.00	51.00	0.00	13.70	86.30	0.00
		106,250	5.40	94.60	0.00	10.60	89.40	0.00	0.50	99.50	0.00
		187,500	4.10	95.90	0.00	6.40	93.60	0.00	0.00	100.00	0.00
		837,500	0.00	100.00	0.00	1.50	98.50	0.00	0.00	100.00	0.00
		2,000,000	0.00	100.00	0.00	0.00	100.00	0.00	0.00	100.00	0.00
3	10	17,500	100.00	0.00	0.00	100.00	0.00	0.00	99.39	0.61	0.00
		81,250	99.39	0.61	0.00	100.00	0.00	0.00	61.26	38.74	0.00
		131,250	91.49	8.51	0.00	100.00	0.00	0.00	49.24	50.76	0.00
		637,500	49.94	50.06	0.00	77.07	22.93	0.00	20.62	79.38	0.00
		1,812,500	30.63	69.37	0.00	49.74	50.26	0.00	12.31	87.69	0.00
		11,250,000	10.31	89.69	0.00	19.51	80.49	0.00	0.60	99.40	0.00
		13,750,000	7.90	92.10	0.00	18.21	81.79	0.00	0.00	100.00	0.00
		100,000,000	0.00	100.00	0.00	6.41	93.59	0.00	0.00	99.20	0.80
4	4	2000	100.00	0.00	0.00	100.00	0.00	0.00	99.80	0.20	0.00
		9000	99.90	0.10	0.00	100.00	0.00	0.00	81.90	18.10	0.00
		15,625	98.40	1.60	0.00	99.90	0.10	0.00	68.70	31.30	0.00
		37,500	88.60	11.40	0.00	96.90	3.10	0.00	49.70	50.30	0.00
		225,000	49.30	50.70	0.00	64.90	35.10	0.00	20.90	79.10	0.00
		475,000	36.60	63.40	0.00	49.40	50.60	0.00	13.30	86.70	0.00
		16,250,000	2.80	97.20	0.00	7.60	92.40	0.00	0.00	100.00	0.00
		100,000,000	0.10	99.90	0.00	0.40	99.60	0.00	0.00	100.00	0.00

$$
\begin{aligned}
P(\hat{r}_{\mathrm{AIC}} > r) &= \sum_{i=r}^{K} P(\hat{r}_{\mathrm{AIC}} = i) \\
&\leq \sum_{i=r}^{K} P\big(2[\log \hat{L}(i+1) - \log \hat{L}(i)] > 2(\gamma_m(i+1) - \gamma_m(i))\big) \\
&\cong \sum_{i=r}^{K} P\big(\chi^2[\gamma_m(i+1) - \gamma_m(i)] > 2(\gamma_m(i+1) - \gamma_m(i))\big).
\end{aligned}
$$

where we have used the fact that $2[\log \hat{L}(l) - \log \hat{L}(r)] \sim \chi^2(\gamma(l) - \gamma(r))$ and $\mathrm{AIC}(l) - \mathrm{AIC}(r) \sim -\chi^2(\gamma_m(l) - \gamma_m(r)) + 2(\gamma_m(l) - \gamma_m(r))$ for $l > r$ (see [2]). For the simplest case $m = 2, r = 1$, the asymptotic bounds for overestimation probability of $\hat{r}_{\mathrm{AIC}}$ is

$$
\begin{aligned}
\sum_{i=2}^{7} P(\hat{r}_{\mathrm{AIC}} = i) &> P\big(2[\log \hat{L}(2) - \log \hat{L}(1)] > 2[\gamma_m(2) - \gamma_m(1)]\big) \\
&= P\big(\chi^2(\gamma_m(2) - \gamma_m(1)) > 2[\gamma_m(2) - \gamma_m(1)]\big) \\
&\cong 0.13
\end{aligned}
$$

and

$$
\begin{aligned}
\sum_{i=2}^{7} P(\hat{r}_{\mathrm{AIC}} = i) &< \sum_{i=2}^{7} P\big(2[\log \hat{L}(i) - \log \hat{L}(i-1)] > 2[\gamma_m(i) - \gamma_m(i-1)]\big) \\
&= \sum_{i=2}^{7} P\big(\chi^2(\gamma_m(i) - \gamma_m(i-1)) > 2[\gamma_m(i) - \gamma_m(i-1)]\big) \\
&\cong 0.13 + 0.09 + 0.05 = 0.27.
\end{aligned}
$$

By increasing the complexity we get for the case $m = 6,\ r = 1$,

$$
\begin{aligned}
\sum_{i=2}^{7} P(\hat{r}_{\mathrm{AIC}} = i) &< \sum_{i=2}^{7} P\big(2[\log \hat{L}(i) - \log \hat{L}(i-1)] > 2[\gamma_m(i) - \gamma_m(i-1)]\big) \\
&= \sum_{i=2}^{7} P\big(\chi^2(\gamma_m(i) - \gamma_m(i-1)) > 2[\gamma_m(i) - \gamma_m(i-1)]\big) \\
&\cong 6 \cdot 10^{-10}.
\end{aligned}
$$

Table 7 shows some of the computed probabilities for different complexities.

Table 7 Computed probabilities for χ^2 distribution

m	l	$\gamma_m(l) - \gamma_m(l-1)$	$P(\chi^2(\gamma_m(l) - \gamma_m(l-1)) > 2(\gamma_m(l) - \gamma_m(l-1)))$	probability
$m = 2$	2	2	$P(\chi^2(\gamma(2) - \gamma(1)) > 2(\gamma(2) - \gamma(1)))$	0.135335
	3	4	$P(\chi^2(\gamma(3) - \gamma(2)) > 2(\gamma(3) - \gamma(2)))$	0.0915782
	4	8	$P(\chi^2(\gamma(4) - \gamma(3)) > 2(\gamma(4) - \gamma(3)))$	0.0423801
	5	16	$P(\chi^2(\gamma(5) - \gamma(4)) > 2(\gamma(5) - \gamma(4)))$	0.00999978
	6	32	$P(\chi^2(\gamma(6) - \gamma(5)) > 2(\gamma(6) - \gamma(5)))$	0.000659928
	7	64	$P(\chi^2(\gamma(7) - \gamma(6)) > 2(\gamma(7) - \gamma(6)))$	0.00000361702
$m = 3$	2	12	$P(\chi^2(\gamma(2) - \gamma(1)) > 2(\gamma(2) - \gamma(1)))$	0.0203410
	3	36	$P(\chi^2(\gamma(3) - \gamma(2)) > 2(\gamma(3) - \gamma(2)))$	0.000340357
	4	108	$P(\chi^2(\gamma(4) - \gamma(3)) > 2(\gamma(4) - \gamma(3)))$	0.00000000333
	5	324	$P(\chi^2(\gamma(5) - \gamma(4)) > 2(\gamma(5) - \gamma(4)))$	$<10^{-10}$
	6	972	$P(\chi^2(\gamma(6) - \gamma(5)) > 2(\gamma(6) - \gamma(5)))$	$<10^{-10}$
	7	2916	$P(\chi^2(\gamma(7) - \gamma(6)) > 2(\gamma(7) - \gamma(6)))$	$<10^{-10}$
$m \geq 6$	2	150	$P(\chi^2(\gamma(2) - \gamma(1)) > 2(\gamma(2) - \gamma(1)))$	$<10^{-10}$
	3	900	$P(\chi^2(\gamma(3) - \gamma(2)) > 2(\gamma(3) - \gamma(2)))$	$<10^{-10}$
	4	5400	$P(\chi^2(\gamma(4) - \gamma(3)) > 2(\gamma(4) - \gamma(3)))$	$<10^{-10}$
	5	32400	$P(\chi^2(\gamma(5) - \gamma(4)) > 2(\gamma(5) - \gamma(4)))$	$<10^{-10}$
	6	194400	$P(\chi^2(\gamma(6) - \gamma(5)) > 2(\gamma(6) - \gamma(5)))$	$<10^{-10}$
	7	1166400	$P(\chi^2(\gamma(7) - \gamma(6)) > 2(\gamma(7) - \gamma(6)))$	$<10^{-10}$

3.3 Concluding Remarks

The results show that "small" penalty terms should imply to a tendency of overestimate the true order, likewise "large" penalty terms implies underestimation. Thus, in general,

$$\hat{r}_{\mathrm{AIC}} \leq \hat{r}_{\mathrm{BIC}} \leq \hat{r}_{\mathrm{EDC}}.$$

Besides that,

(i) Both $\hat{r}_{\mathrm{BIC}}$ and $\hat{r}_{\mathrm{EDC}}$ never overestimate the true order.
(ii) $\hat{r}_{\mathrm{BIC}}$ has a higher tendency to underestimate the order.
(iii) $\hat{r}_{\mathrm{EDC}}$ outperforms $\hat{r}_{\mathrm{BIC}}$; it is the most efficient consistent estimator.
(iv) For small complexity models $(\gamma_m(r) < 10)$ and small sample size $(n \leq 200)$, $\hat{r}_{\mathrm{AIC}}$ performs better than $\hat{r}_{\mathrm{EDC}}$, but may overestimate the true order.
(v) For medium or large complexity models $(\gamma_m(r) > 20)$, overestimation by $\hat{r}_{\mathrm{AIC}}$ is negligible and efficiency of $\hat{r}_{\mathrm{AIC}}$ and $\hat{r}_{\mathrm{EDC}}$ is comparable.
(vi) We observed that $\delta(k)$ is significant on the behavior of estimators based on penalized log-likelihood. By a simple computation, we verify that $\delta(k) \leq \log(m)$, but a better bound must be investigated to achieve better order estimators.

3.4 Computational Environment Details

The simulation algorithm uses the libc6 2.7 library and was compiled using GCC 4.3.1. The processed data was persisted in a PostgreSQL database. For reports building, was used AWK. The computation were done using a 4 servers cluster equipped with AMD Turing 64 processors.

Acknowledgments This work was supported in part by CNPq-Brazil, CAPES-Brazil and FAPDF-Brasilia.

References

1. Akaike H (1974) A new look at the statistical model identification. IEER Trans Autom Control 19(6):716–723
2. Billingsley P (1961) Statistical methods in Markov chains. Ann Math Stat 32(1):12–40
3. Csiszár I, Shields PC (2000) The consistency of the BIC Markov order estimator. Ann Stat 28(6):1601–1619
4. Dorea CCY (2008) Optimal penalty term for EDC Markov chain estimator. Ann de l'Inst Stat l'Univ de Paris 52:15–26

5. Dorea CCY, Zhao LC (2006) Bounds for the probability of wrong determination of the order of a Markov chain by using the EDC criterion. J Stat Planning Infer 136:3689–3697
6. Dorea CCY, Goncalves CR, Resende PAA (2014) Simulation results for markov model selection : AIC, BIC and EDC. Lecture notes in engineering and computer science. In: Proceedings of the world congress on engineering and computer science 2014, WCECS2014, 22–24 Oct 2014. San Francisco, USA, pp 899–901
7. Katz RW (1981) On some criteria for estimating the order of a Markov chain. Technometrics 23(3):243–249
8. Park SK, Miller KW (1988) Random number generators : good ones are bad to find. Commun ACM 31(10):1192–1201
9. Raftery AE (1985) A model for high-order Markov chains. J Roy Stat Soc B 47(3):528–539
10. Schwarz G (1978) Estimating the dimension of a model. Ann Stat 6(2):461–464
11. Zhao LC, Dorea CCY, Gonçalves CR (2001) On determination of the order of a Markov chain. Stat Infer Stochast Process 4(3):273–282

Neural Network Ensemble Based QSAR Model for the BBB Challenge: A Review

Mati Golani and Idit. I. Golani

Abstract The blood-brain barrier (BBB) presents a real challenge to the pharmaceutical industry. The BBB is a very effective screener of diverse kinds of bacterial infections. Unfortunately, this functionality prevents from many drugs to penetrate it. In order to improve drug development process an assessment model is required. Effective assessment models can drastically reduce development times, by cutting off drugs with low success rates. It also saves financial resources since clinical trials will focus mainly on drugs with higher likelihood of permeation. This work addresses the challenge by means of artificial neural net (ANN) based assessment tool. Embedding 'wisdom of experts' approach, the presented assessment tool is combined of a neural net ensemble, a group of trained neural nets that correspond to an input value set with a prediction of the barrier permeation. The returned output is the median of the ensemble's members output. The input set is composed of drug physicochemical properties such: Lipophilicity, Molecular Size (depends on Molecular Mass/Weight), Plasma Protein Binding, PSA—Polar Surface Area of a molecule, and Vd—Volume of Distribution, and Plasma Half Life (t ½). Challenged with the relatively small learning data-set, leave one out (LOO) which is a special case of k-fold cross validation is conducted. Although the training effort for building ANNs is much higher, in small data-sets ANNs yield much better model fitting and prediction results than the logistic regression.

Keywords BBB · Brain to plasma ratio · Ensemble · Neural net · Pharmacokinetics · QSAR

M. Golani (✉)
Department of Software Engineering, ORT Braude College, Snunit 51,
P.O.Box 78, Karmiel 21982, Israel
e-mail: matig@braude.ac.il

I.I. Golani
Department of Biotechnology Engineering, ORT Braude College, Snunit 51,
P.O.Box 78, Karmiel 21982, Israel
e-mail: igolani@braude.ac.il

H.K. Kim et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-94-017-7236-5_5

1 Introduction

The blood-brain barrier (BBB) is very effective screener of diverse kinds of bacterial infections. This screening feature of the BBB presents a real challenge to the pharmaceutical industry. Unfortunately, this functionality in addition prevents many drugs from penetrating it. In order to make the drug development process an efficient one, an assessment model is required. Effective assessment model can drastically reduce development times, by cutting off drugs with low success rates [1]. It also saves financial resources since clinical trials will focus mainly on drugs which are more likely to succeed on their task.

1.1 Pharmacology Perspective

The Blood Brain Barrier (BBB) consists of a monolayer of brain micro vascular endothelial cells (BMVEC), which are joined together by tight junctions and form a cellular membrane [2, 3]. BMVECs surrounded by a basement membrane, together with other components: pericytes, astrocytes and microglia, compose a neurovascular unit [3].

The BBB has a carrier function which is responsible for the transport of nutrients into the brain and removal of metabolites from it. While small lipid-soluble molecules (e.g. ethanol) diffuse passively through the BBB, other essential polar nutrients (glucose, amino acids) require some specific transporters. The BBB has also a barrier function that restricts the transport of potentially toxic substances through the BBB. This is achieved by a para-cellular barrier (tight endothelial junctions); trans-cellular barrier (endocytosis and trans-cytosis); enzymatic barrier (proteins with enzymatic activities) and efflux transporters. The specific barrier function of the BBB is important for preventing Central Nervous System (CNS) from harmful xenobiotics, but at the same time, prevents or limits the penetration of many drugs to the CNS [4].

The ability of these drugs to penetrate the BBB or be transported across the BBB is mainly dependent on their physiochemical properties and their affinity to a specific transport system [5].

1.2 Common Descriptors

Drug distribution into the CNS depends on the physicochemical properties of the compound, including: lipophilicity (log P), molecular weight (MW), and PK parameters such as: protein binding, volume of distribution (V_d), half-life etc. [6].

- Lipophilicity—Compound lipophilicity plays an important role in the absorption, distribution, metabolism, and excretion (ADME) of therapeutic drugs. Lipophilicity represents the affinity between a molecule and its lipophilic environment. It is commonly measured by its distribution behavior in a biphasic system, either liquid-liquid (e.g., partition coefficient in 1-octanol/water) or solid liquid (retention on reversed-phase high-performance liquid chromatography (RPHPLC) or thin-layer chromatography (TLC) system). Lipophilicity refers to the ability of a chemical compound to dissolve in fats, oils, lipids, and non-polar solvents such as hexane or toluene. Lipophilicity is often expressed as Log P, logarithm of partition coefficient P between lipophilic organic phase (1-octanol) and polar aqueous phase. While high degree of lipid solubility favors crossing the BBB by transmembrane diffusion, it also favors uptake by the peripheral tissues, thus it can lower the amount of the drug presented to the BBB [7]. In many situations lipophilicity is a good predictor of BBB penetration [8].
- Molecular weight—The diffusion coefficient in liquids is inversely proportional to the hydrodynamic radius of the compound, and the hydrodynamic radius is approximately proportional to the square root of the molecular mass; i.e., the entry of a compound into the CSF depends on the square root of the molecular mass. Although the penetration of large hydrophilic compounds into the CSF is low, there is no absolute cutoff. Molecules as large as 1 gM are present in normal CSF at approximately 1/1000 of their serum concentration. The optimal molecular mass for passage into the brain lies in the region of 300–400 Da [9–12]. The best approximation of molecular size influence on BBB penetration is that it is inversely related to the square route of a molecular weight [13].
  A limited number of drugs with high lipophilicity and low molecular size can penetrate to the brain mainly by passive diffusion.
- Polar Surface Area (PSA)—PSA is defined as the sum of polar atoms surface (oxygen, nitrogen and attached hydrogen) in a molecule. PSA is a commonly used metric for the optimization of a drug's ability to permeate cells. Molecules with a polar surface area of greater than 140 angstroms perform poorly with regard to cell membranes permeation. This parameter has been shown to correlate very well with BBB penetration [14, 15]. BBB permeation decreases 100-fold as the surface area of the drug is increased by 2-fold (from 52 angstroms to 105 angstroms) [9].
- Protein binding—The extent of drug distribution into tissues, including the CNS, depends on the degree of plasma protein binding (albumin, α1-acid glycoprotein, and lipoproteins). Drugs which are highly bound confined to the vascular system and as result have a relatively low *Volume of distribution*. Drugs with lower binding degree are more available for distribution to other organs and tissues, including the CNS. Only unbound drug is available for passive diffusion through the BBB and for pharmacologic effect. The penetration rate into the brain is slow for highly protein-bound drugs [16].

- Volume of distribution (V_d)—is a proportionality factor that relates to the amount of a drug to its measured concentration. The apparent volume of distribution is a theoretical volume of fluid into which the total drug administered would have to be diluted to produce the concentration in plasma. Some drugs distribute mostly into fat, others remain in extracellular fluid, while the rest are bound extensively to specific tissues. For a drug that is highly tissue-bound, very little drug remains in the circulation, thus plasma concentration is low and volume of distribution is high [17].

$$V_D = V_P + V_T\left(\frac{f_u}{f_{u_t}}\right) \tag{1}$$

where:
V_p Plasma volume
V_T Apparent tissue volume
f_u Fraction unbound in plasma
f_{ut} Fraction unbound in tissue

- Plasma Half Life ($T_{½}$)—The biological half-life or elimination half-life of a substance is defined by the time it takes for a substance to lose half of its pharmacologic, physiologic, or radiologic activity. In a medical context, it is usually considered as the time it takes for the blood plasma concentration of a substance to be reduced by one half ("plasma half-life"). The relationship between the biological and plasma half-lives of a substance can be complex depending on the substance in question, due to factors including accumulation in tissues, active metabolites, and receptor interactions.
- Brain/Plasma ratio (Permeation measure)—The most common method to study brain penetration in vivo is the determination of the brain/plasma ratio in rodents. For that, the test compound is dosed and both plasma and brain are sampled. The logBB describes the ratio between brain and blood (or plasma) concentrations and provides a measure of the extent of drug permeation through the BBB

$$\text{LogBB} = \frac{AUC_{\text{tot.brain}}}{AUC_{\text{tot.blood}}} \tag{2}$$

Another in vivo measurement of CNS permeation is the log of the permeability-surface area coefficient (log PS) which is considered to be the most appropriate in vivo measurement [18, 19]. However, this is a resource-intensive measure that requires microsurgical expertise. This method's advantage is by eliminating drug's serum binding. Nevertheless, by using log BB together with plasma protein binding, one can produce same or even better results.

During drug development, in vitro, ex vivo and in vivo models have been developed in order to examine the mechanisms by which different drugs penetrate the BBB.

Tissue distribution studies are commonly conducted by a traditional method using radiolabeled compounds. Brain tissue is homogenized and precipitated, and the total brain concentration of the radioactive compound is determined using liquid scintillation counting and related to its concentration in plasma.

An alternative method is quantitative microdialysis, a widely used technique that permits quantifications of drug transport to the brain. Drug concentrations measured by microdialysis are influenced by properties of the probe and perfusion solution, by the post-surgery interval in relation to surgical trauma, and tissue integrity properties [20].

All the mentioned methods for drugs permeation to the BBB are labor intensive, demand expensive compounds and equipment and use many animals. Rapid screening methods can speed up discovery and minimize the number of drug candidates for further detailed studies.

Such computational models which exist since the 1980s are based on drug's lipophilicity, hydrogen-bond potential, pKa/charge and molecular size [21–25]. However, in these models, other factors that can determine drug's concentration at the brain capillary surface, are not included. Factors such plasma protein binding or volume of distribution (V_d), which are present in the presented model.

The rest of the paper is organized as follows: Sect. 2 introduces known structured and non-structured based modeling methods in pharmacokinetics, and present the neural-net based model. In Sects. 3, 4 and 5 we present data pre-processing, training, and testing results, respectively. Finally, in Sect. 6, we introduce the architecture for implementation, and show results.

2 Modeling Techniques

Diverse modeling techniques such multi-linear regression, clustering, Neural nets, Bayesian neural nets [26], and decision-trees [27] where introduced with regard to pharmacokinetics modeling.

A fashionable classification of BBB permeation as appear in some published papers is to classify the BBB permeation measure into two classes: "good" (CNSp+), and "bad" (CNSp−) [28]. While the measure is indeed qualitative, a finer resolution classification may provide better comparable order between candidate drugs performance.

A neural network (ANN) is a mathematical model which is based on the biological brain structure. Interconnected processing units that form a network.

2.1 ANN for Pharmacokinetic Modeling

Neural network based approach is well known in the pharmacokinetic domain [28]. In comparison with multi-linear regression, ANNs are more flexible, robust, and better at prediction [29]. Furthermore, multi-linear regression is more sensitive to the relationship between the number of patterns and number of variables, thus it needs to be monitored in order to avoid chance effects [30]. Another disadvantage is that drug data often contains correlated or skewed information. This can then lead to the construction of poor regression models [29]. Usage of ANN ensembles that aggregate the descriptors' output has been also reviewed [31], where it was concluded that an ensemble improves generalization performance.

A distinguishing feature of neural networks is that knowledge is distributed throughout the network itself rather than being explicitly written into the program. The network then learns through exposure to diverse input set with known output.

3 Data Pre-processing

Relevant data of 47 drugs was collected from the literature. Only drugs for which all required metrics were available, were collected.

3.1 Consideration

Neural network training can be executed in a more efficient manner if certain preprocessing steps on the network inputs and targets are performed. Prior the network design process, the data is collected and prepared. It is generally difficult to incorporate prior knowledge into a neural network; therefore the network can only be as accurate as the data that is used for training it. After the data has been collected, there are two crucial steps to be performed before training starts: the data is uniformly distributed and then normalized.

3.2 Data Distribution

Some properties of the collected drugs have poor distribution. ANN prediction results tend to be more promising when the data is properly distributed. In order to improve the data distribution log operator was applied on V_d, Half Life and Brain to Plasma Ratio properties. V_d values are presented in Fig. 1. It can be concluded from Fig. 1 that input data has better distribution using a log operation with regard to V_d.

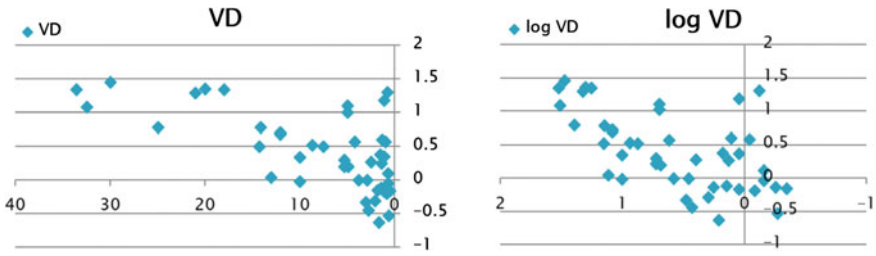


Fig. 1 Better uniformity in distribution with Log VD

3.3 Normalization

Normalized data has a common base, which means that every member is evaluated for each metric with respect to other members metric in the group on a scale range of [−1, 1]. ANN's perform much better on normalized data sets. The normalization step was applied on the input and the target vectors of the data set (BBB permeability).

3.4 Measures and Precision Concern

As mentioned in Sect. 2 the permeability measure should be qualitative rather than quantitative. This is due to the fuzzy nature of measurement and lack of persistence. For example, plasma to brain ratio of Clozapine (Clozaril) appears as 24 [32], or 4.1 [33], which is a large gap. Therefore it was decided that the drugs will be divided into four permeability groups in the range 1–4 where 1 is the least permeable drug, and 4 is the most. A deviation value of 1 is acceptable, and considered a success (e.g. a drug that was detected as group 3 and is actually a member of group 4, will be considered as true positive).

4 Training

4.1 Network Modeling

A feed-forward back propagation network was chosen. In such topology, each input measure is connected to an input neuron; there may be one or more hidden layer neurons, and an output neuron that provides the output measure (permeability). The hyperbolic tangent transfer function was used for the hidden layer neurons, and a logarithmic transfer function was used for the output layer. Learning rate was set to 0.05 with a momentum of 0.1.

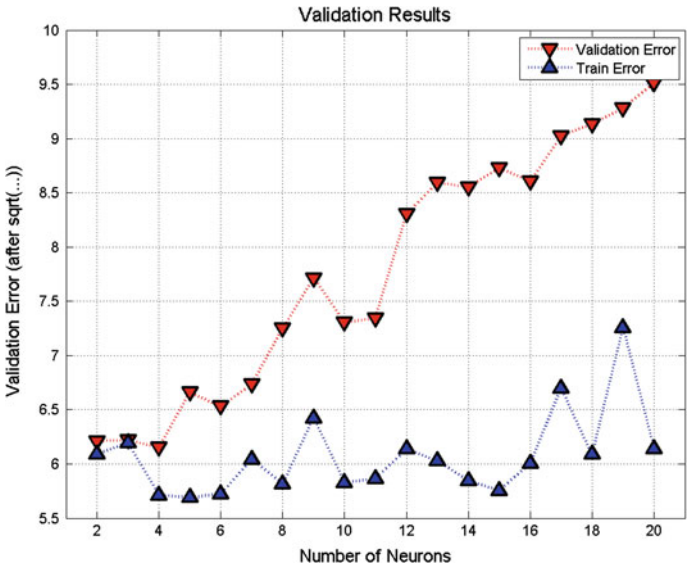


Fig. 2 Validation/training error versus hidden layer size

4.2 Hidden Layer Size

One should take into consideration when comparing networks with relatively similar accuracy, that the smaller the network, the more general it is in terms of model. When the network size increases, it may just encapsulate the specific data set instead of the general model. In order to determine the proper hidden layer size, an initial training phase was conducted on networks with variable hidden layer size.

As reflected in Fig. 2 Validation/training error versus hidden layer size, it infers that a hidden layer of 2–3 neurons provides best results. Bigger layers maybe provide better results with respect to the training error, but this result is actually misleading since it is a symptom of over-fitting. As shown in Fig. 2 Validation/training error versus hidden layer size, beyond 3 neurons in the hidden layer, the validation error actually increases and those networks reduce generalization.

4.3 Early Stopping

In machine learning, early stopping is a known method for improving generalization. The data is divided into a training-set and a validation-set.

The training set is used for computing the gradient and updating the network weights and biases. The validation set is used for monitoring. The error on the validation set is monitored during the training process. The validation error normally decreases during the initial phase of training, as does the training set error. However,

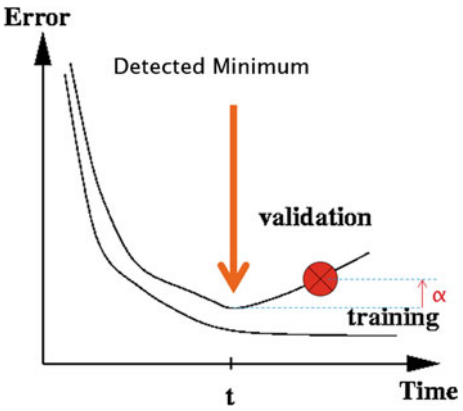


Fig. 3 Early stopping training

when the network begins to over-fit the data, the error on the validation set typically begins to rise. When the validation error increases for a specified number of iterations, or beyond a predefined threshold −Alpha, the training is stopped. Early stopping is effectively limiting the used weights in the network and thus imposes regularization. This procedure is illustrated in Fig. 3 early stopping training.

4.4 Cross Validation

In small data sets leave-one-out (LOO) cross-validation is normally applied. This is a special case of k-fold cross validation [28, 34]. With a very small sample size (18 bankrupt and 18 non-bankrupt firms), Fletcher and Goss employ an 18-fold cross-validation method for model selection. Although the training effort for building ANNs is much higher, ANNs yield much better model fitting and prediction results than the logistic regression [35].

4.5 Net Tournament

During the cross validation, and for each fold, a tournament between 100 networks was conducted. Only the winner network with best results during this fold (minimal error) was retained, as illustrated in Fig. 4. The algorithm described above is elaborated in Algorithm 1.

4.6 ANN Ensemble

Since a network training tournament is performed for each fold, the outcome is a group of winning ANN's, one for each sample. Most often one would pick the best

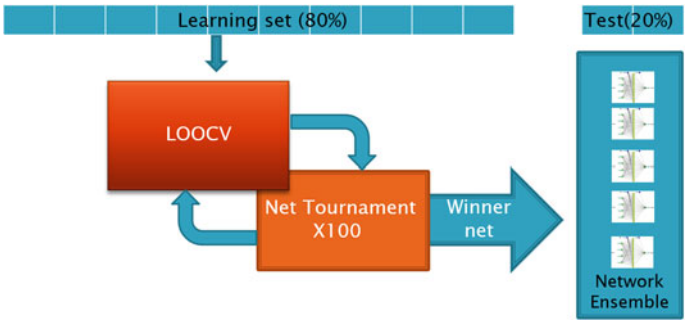


Fig. 4 Design structure of the learning phase and ensemble generation

performing network. Nevertheless, here we suggest a different approach, i.e. a neural net ensemble. A neural net ensemble is a group of ANN's that provides a single output to a given input. This output can be the average of the ensemble members output (most common), a quorum based result, median, etc. In this work we have used the median of ensemble members output as the ensemble's output.

Algorithm 1

Part 1 – Preprocessing

- *Distribute data.*
- *Normalize data.*
- *Predefine the quantity of neurons in hidden layer. Parameter named* ($\boldsymbol{n_q}$).

Part 2 – Build the ensemble

1. *Randomly divide the data into training (80%) and testing (20%) sets (ts). Testing set consists of drugs that contain no extreme values (1or -1).*
2. *For every fold (drug) f∈ training_set, use f as the validation_set and training_set\f as learning_set (LOOCV):*
   $\boldsymbol{v_g}$ *is a validation group, which* $|\boldsymbol{v_g}|=1$.
   $\boldsymbol{t_g}$ *is a training group.* $\boldsymbol{t_g}$ = training_set\ $\boldsymbol{v_g}$.
   2.1. *For j = 1 to 100*
      2.1.1. *Create NN* ($\boldsymbol{n_j}$) *with* $\boldsymbol{n_q}$ *neurons in hidden layer.*
      2.1.2. *Train* $\boldsymbol{n_j}$ *using* $\boldsymbol{t_g}$ *according to "Early-Stop Strategy".*
      2.1.3. *Validate* $\boldsymbol{n_j}$ *using* $\boldsymbol{v_g}$.
      2.1.4. *Save* $\boldsymbol{n_j}$ *with its error in array nj_array.*
   2.2. *Select the best NN in* ***nj_array*** *array –* NN_x *with minimal training error.*
   2.3. *Insert* NN_x *to ensemble (****nne****)*

4.7 Testing

The preprocessed data was repetitively (several hundred test cycles), and randomly divided into a Training (80 %) and a Testing (20 %) group

- The training set was utilized, and the ANN ensemble was generated
- Drug data from the test group was presented to the ensemble, and its output was compared to its known permeability measure—in terms of permeation group membership
- Results were grouped with accordance to the delta between the predicted and actual permeability group (see Sect. 3.4).

5 Architecture

Following is a description of the implementation sequence:

- A first analysis for determining the hidden layer size was conducted both on Matlab and the Encog framework.
- Next phases as described in Sect. 4, were implemented in C# using the Encog . NET package. Data was saved on a server database. The application was designed as a client/server 3-Tier architecture application:

  – *Presentation tier*—This is the topmost level of the application. The presentation tier displays a user interface, which enables a convenient (a user friendly) access to the model. It communicates with other tiers by outputting results to the application tier.
  – *Application tier* (*business logic*)—The logic controls the application‘s functionality by performing detailed processing.
  – *Data tier*—A delegator to the database server. Here information is stored and retrieved. This tier keeps the data in neutral and independent form with regards to the application servers or business logic. Giving data its own tier also improves scalability and performance.

Figure 5 provides a graphical representation of the results. The presented results are the average of several hundred runs. In addition, diverse early stopping settings were investigated in terms of maximal number of epochs [15–100 K], and alpha [5–20 %] values. Most combinations provided a similar and satisfactory result of 89 % success rate. Maximal accuracy was reached with 15 K epochs limit and an Alpha of 5 %.

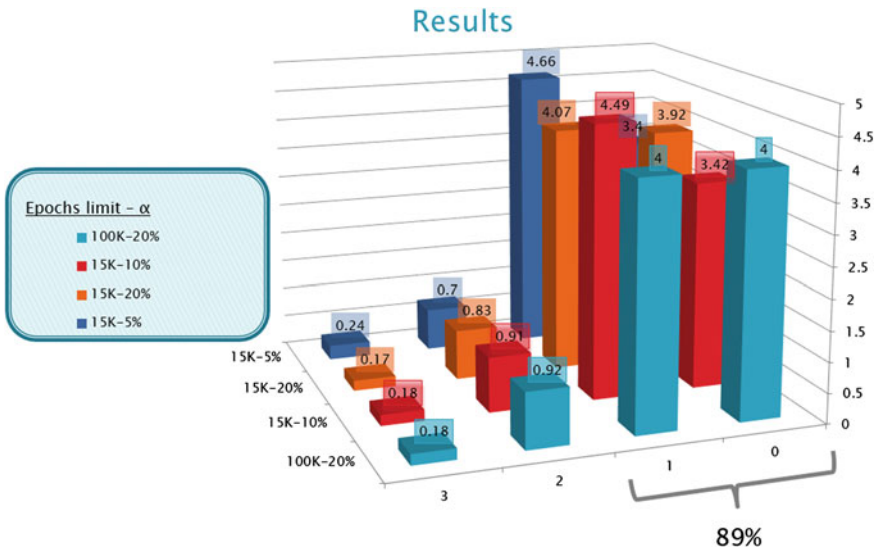


Fig. 5 Permeability group classification precision

6 Conclusion and Future Work

This chapter, proposes a new ensemble neural net based mechanism as an evaluator for drug-BBB permeation. A design time evaluator for drug development—in particular, by providing finer permeation classification, with relatively high success rates. Due to the small given data set, a leave one out cross validation technique was performed.

Our goal is to develop an approach that allows an interactive drug design that is less labor intensive, or demand expensive with respect to compounds/equipment, thus uses less animals in pre-clinical stages. This can be achieved by using such a mechanism for scoring candidates, while excluding inferior candidates, and performing the more expensive pre-clinical stages on the provided best candidates. Our specific contribution in this work is threefold. We have incorporated plasma protein binding as an input parameter of the model; We have extended the fissionable binary classification CNSp+, CNSp− to a wider four values set; We also propose a new modeling mechanism using neural-net ensemble—grouped of tournaments winners—that provides finer permeation resolution while coping with relatively smaller data sets which is extremely challenging. The benefits of this approach have been discussed in this work.

A possible and promising track for a future research involves larger dataset incorporation. We also intend to extend our model with metrics such as: Hydrogen-bonding (Hydrogen bond acceptor/donor), and drug's affinity to efflux

transporters such as P-glycoprotein (P-gp). Since there is a correlation between the neural net input layer size and the recommended dataset size, such an extension would also require an essential extension of the dataset size.

References

1. Golani M, Golani I (2014) Neural net ensemble based QSAR modeler for drug blood brain barrier permeation. In: Lecture notes in engineering and computer science: proceedings of the world congress on engineering and computer science 2014, WCECS, San Francisco, USA, pp 818–823, 22–24 Oct 2014
2. Abott NJ (2005) Physiology of the blood-brain barrier and its consequences for drug transport to the brain. Int Congr Ser 1277:3–18
3. Cardoso FL, Brites D, Brito MA (2010) Looking at the blood-brain barrier: molecular anatomy and possible investigation approaches. Brain Res Rev 64:328–363
4. Loscher W, Postschka H (2005) Role of drug efflux transporters in the brain for drug disposition and treatment of brain diseases. Prog Neurobiol 76:22–76
5. Begley DJ (2004) ABC transporters and the blood-brain barrier. Curr Pharm Des 10 (12):1295–1312
6. Schmidt S, Gonzalez D, Derendorf H (2010) Significance of protein binding in pharmacokinetics and pharmacodynamics. J Pharm Sci 99(3):1107–1122
7. Greig NH, Brossi A, Pei XF, Ingram DK, Soncrant TT (1995) Designing drugs for optimal nervous system activity. In: Greenwood J, Begley DJ, Segal MB (eds) New concepts of a blood-brain barrier. Plenum Press, New York, pp 251–264
8. Waterhouse RN (2003) Determination of lipophilicity and its use as a predictor of blood-brain barrier penetration of molecular imaging agents. Mol Imaging Biol 5(6):376–389
9. Fischer H, Gottschlich R, Seelig A (1998) Blood-brain barrier permeation molecular parameters governing passive diffusion. J Membr Biol 165:201–211
10. van de Waterbeemd H, Camenisch G, Folkers G, Chretien JR, Raevsky OA (1998) Estimation of blood-brain barrier crossing of drugs using molecular size and shape, and H-bonding descriptors. J Drug Target 6:151–165
11. Pardridge WM (2001) Brain drug targeting: the future of brain drug development. Cambridge University Press, Cambridge
12. Levin VA (1980) Relationship of octanol/water partition coefficient and molecular weight to rat brain capillary permeability. J Med Chem 23:682–684
13. Felgenhauer K (1980) Protein filtration and secretion at human body fluid barriers. Pflugers Arch 384:9–17
14. Shityakov S, Neuhaus W, Dandekar T, Förster C (2013) Analysing molecular polar surface descriptors to predict blood-brain barrier permeation. Int J Comput Biol Drug Des 6(1–2):146–156
15. Kelder J, Grootenhuis PD, Bayada DM, Delbressine LP, Ploemen JP (1999) Polar molecular surface as a dominating determinant for oral absorption and brain penetration of drugs. Pharm Res 16(10):1514–1519
16. Nau R, Sorgel F, Prange HW (1994) Lipophilicity at pH 7.4 and molecular size govern the entry of the free serum fraction of drugs into the cerebrospinal fluid in humans with uninflamed meninges. J Neurol Sci 122:61–65
17. van de Waterbeemd H (2005) Which in vitro screens guide the prediction of oral absorption and volume of distribution? Basic Clin Pharmacol Toxicol 96:162–166
18. Martin I (2004) Prediction of blood-brain barrier penetration: are we missing the point? Drug Discov Today 9:161–162

19. Pardridge WM (2004) Log(BB), PS products and in silico models of drug brain penetration. Drug Discov Today 9(9):392–393
20. De Lange EC, Danhof M (2002) Considerations in the use of cerebrospinal fluid pharmacokinetics to predict brain target concentrations in the clinical setting: implications of the barriers between blood and brain. Clin Pharmacokinet 41:691–703
21. Young RC et al (1988) Development of a new physicochemical model for brain penetration and its application to the design of centrally acting H2 receptor histamine antagonists. J Med Chem 31(3):656–671
22. Chikhale EG, Ng KY, Burton PS, Borchardt RT (1994) Hydrogen bonding potential as a determinant of the in vitro and in situ blood-brain. Pharm Res 11(3):412–419
23. Abraham MH (2004) The factors that influence permeation across the blood-brain barrier. Eur J Med Chem 39(3):235–240
24. Jezequel SG (1992) Central nervous system penetration of drugs: importance of physicochemical properties. Progr Drug Metab 13:141–178
25. Atkinson F, Cole S, Green C, van de Waterbeemd H (2002) Lipophilicity and other parameters affecting brain penetration. Curr Med Chem CNS Agents 2(3):229–240
26. Goodwin JT, Clark DE (2005) In silico predictions of blood-brain barrier penetration: considerations to "keep in mind". J Pharmacol Exp Ther 315(2):477–483
27. Suenderhauf C, Hammann F, Huwyler J (2012) Computational prediction of blood-brain barrier permeability using decision tree induction. Molecules 17(9):10429–10445
28. Turner JV, Maddalena DJ, Cutler DJ (2004) Pharmacokinetic parameter prediction from drug structure using artificial neural networks. Int J Pharm 270(1–2):209–219
29. Butina D, Segall MD, Frankcombe K (2002) Predicting ADME properties in silico: methods and models. Drug Discovery Today 7:S83–S88
30. Topliss JG, Edwards RP (1979) Chance factors in studies of quantitative structure-activity relationships. J Med Chem 22:1238–1244
31. Agrafiotis DK, Cedeño W, Lobanov VS (2002) On the use of neural network ensembles in QSAR and QSPR. J Chem Inf Comput Sci 42:903–911
32. Zhang G, Terry A Jr, Bartlett MG (2007) Sensitive liquid chromatography/tandem mass spectrometry method for the simultaneous determination of olanzapine, risperidone, 9-hydroxyrisperidone, clozapine, haloperidol and ziprasidone in rat brain tissue. J Chromatogr B 858(1):276–281
33. Maurer TS, DeBartolo DB, Tess DA, Scott DO (2005) Relationship between exposure and nonsprcific binding of thirty-three central nervous system drugs in mice. Drug Metab Dispos 33(1):175–181
34. Tetko IV, Livingstone DJ, Luik AI (1995) Neural network studies. 1. Comparison of overfitting and overtraining. J Chem Info Comp Sci 35:826–833
35. Fletcher D, Goss E (1993) Forecasting with neural networks: an application using bankruptcy data. Inf Manag 24:159–167

Building Heating Feed-Forward Control Method and Its Application in South Ural State University Academic Building

Dmitry A. Shnayder, Vildan V. Abdullin and Aleksandr A. Basalaev

Abstract This paper proposes a new structure of heating feed-forward control system for multi-storey buildings based on indoor air temperature inverse dynamics model. The suggested model enables real-time assessment of the impact of perturbing factors that cannot be directly measured on indoor air temperature with due allowance for lag and nonlinear nature of thermohydraulic processes involved in the building heating cycle. To measure current indoor air temperature values, the authors used a distributed field-level sensor network. The approaches to estimate heat consumption when heat meter is not installed, are proposed. This paper also contains the results of identification for the formulated model, as well as a calculation of energy savings after deployment of heating control system in the academic building of South Ural State University in accordance with International Performance Measurement and Verification Protocol (IPMVP).

Keywords Automated heat station · Building thermal performance simulation · Feed-forward control · Heat consumption estimation · Heating of buildings · International performance measurement and verification protocol · Inverse dynamics model

D.A. Shnayder · V.V. Abdullin (✉) · A.A. Basalaev
Automatics and Control Department, South Ural State University,
Lenina pr., 76, Chelyabinsk 454080, Russia
e-mail: vildan@ait.susu.ac.ru

D.A. Shnayder
e-mail: shnayder@ait.susu.ac.ru

A.A. Basalaev
e-mail: alexander-basalaev@mail.ru

H.K. Kim et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-94-017-7236-5_6

1 Introduction

Making building heating systems more efficient is one of the key tasks of energy and resource conservation. In Russia and some other northern countries, where heating season takes up most of the year, heating costs account for the majority of spendings on energy resources consumed by residential and office buildings. For example, the share of heating energy in overall energy consumption of South Ural State University main campus totals 36.7 %, which ultimately corresponds to 30,100 Gcal (based on the 2012 energy audit).

Energy efficiency of heating systems used in buildings could not possibly be upgraded without automatic control systems that implement a variety of control algorithms [1–3]. Baseline control principle used in these systems is control of the indoor temperature by reference to the primary perturbing factor, i.e. outdoor temperature T_{out}. This approach appears viable, as it guarantees adequate quality, while implementing simple control algorithms and using data that is easy to measure. However, the key factor that determines the quality of heating system performance is indoor air temperature T_{ind}. This is why it is important to design heating control systems that take into account the actual T_{ind} values. Control principle based on T_{ind} is described in a number of papers [4, 5], but there are certain challenges that make practical implementation of this approach rather problematic:

- air temperature varies in different rooms of a multi-storey building;
- building heating system is highly inertial and performs as a nonlinear distributed system, which renders control by indoor air temperature rather difficult in a situation when common engineering methods are employed;
- the building is exposed to numerous perturbing factors (Fig. 1) that are hard to measure or evaluate in real terms.

The indoor air temperature T_{ind} of a building depends on its volume, building envelope type, the quantity of applied heating energy Q_{source}, inner and external perturbing factors, such as the outdoor air temperature T_{out}, solar radiation J_{rad}, wind V_{wind}, internal heat release Q_{int}, and the building's accumulated internal heating energy Q_{acc} [8]. However, the signals T_{ind}, Q_{source}, and T_{out} presented in Fig. 1 can be measured quite easily in practice, while direct measurement of J_{rad}, V_{wind}, Q_{int} and Q_{acc} that affect the temperature T_{ind} is actually problematic.

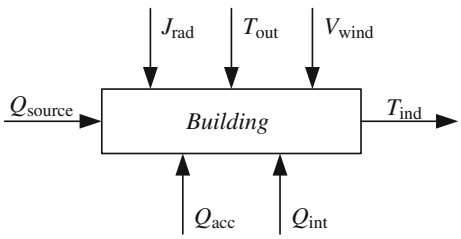


Fig. 1 Factors affecting the indoor air temperature

These are the reasons why most heating control systems for multi-storey buildings that exist on the market and are widely deployed in real life either completely disregard the indoor air temperature, or refer to its value only to monitor the quality of control measures, without actually using it to adjust the control signal.

To solve above-mentioned problems, complex objects of this sort are now commonly designed in accordance with model-predictive control methods [6, 7] that are based on mathematical modeling of the object. This paper describes a new approach to heating control in a multi-storey building that is based on estimation of perturbations affecting on the indoor air temperature using the building thermal performance inverse dynamics model.

2 Control Approach Based on Inverse Dynamics Model

2.1 Feed-Forward Control System Structure

To take into account the unmeasured factors that affect temperature $T_{\text{ind}}(t)$, we referred to the approach based on the concept of generalized temperature perturbation $T_{\text{z}}(t)$ [8] characterizing the effect of the factors mentioned above on the indoor air temperature.

Generic structure of the proposed feed-forward control system is described in Fig. 2. As we see in the figure, baseline control of heat supply for building heating purposes follows a standard pattern with the use of automated building heat station that controls heating power $Q_{\text{h}}^{T\text{out}}$ depending on the key perturbing factor—outdoor air temperature. The structure shown in Fig. 2 is augmented by a feed-forward control loop used to adjust $Q_{\text{h}}^{T\text{out}}$ depending on estimated value T'_{z} of general temperature perturbation T_{z}. Thus, the adjusted heating power value $Q_{\text{h}}^{T\text{ind}}$ fed into the building is calculated as follows:

$$Q_h^{T\text{ind}} = Q_h^{T\text{out}} + \Delta Q_h^{T\text{ind}}, \tag{1}$$

where $\Delta Q_{\text{h}}^{T\text{ind}}$—adjusting value of heating power produced by feed-forward temperature controller.

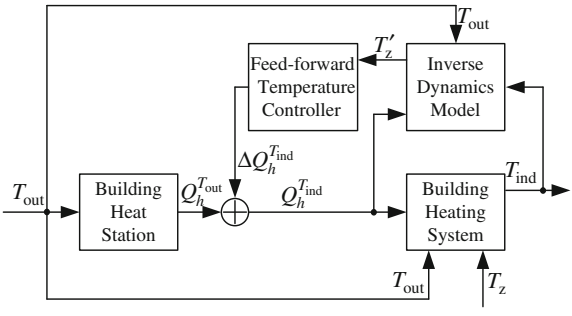


Fig. 2 Generic structure of the building heating control system with feed-forward temperature controller

Let us consider the inverse dynamics model. In accordance with [8], the heat balance equation takes the following form:

$$T'_{\text{ind}}(t) = Q_{\text{h}}(t)/(q_{\text{h}} \cdot V) + T_{\text{ind}}(t) - T_{\text{z}}(t), \tag{2}$$

where $T'_{\text{ind}}(t)$ stands for the predicted value of indoor air temperature [the prediction horizon is determined by the fluctuation of indoor air temperature as a result of the perturbing factors (Fig. 1)]; $T_{\text{out}}(t)$ is outdoor air temperature; $Q_{\text{h}}(t)$ stands for heating power applied to the heating system from the heating radiator; q_{h} represents specific heat loss (per cubic meter); and V stands for external volume of the building.

A block diagram of building thermal performance dynamics model composed in accordance with (2) is presented in Fig. 3. The key input signal in this model is represented by heating power $Q_{\text{h}}(t)$ delivered by heating radiators and generated by the heat station. The dynamics operator $F_0\{\bullet\}$ describes building's heat exchange process dynamics [8].

According to the model shown in Fig. 3, the feed-forward value of indoor air temperature can be determined by the following equation:

$$T'_{\text{ind}}(t) = F_0^{-1}\{T_{\text{ind}}(t)\}, \tag{3}$$

where $F_0^{-1}\{\bullet\}$ stands for the inverse dynamics operator calculated using the exponential filtration method [8, 9]. Consequently, the estimated value of general temperature perturbation T'_{z} can be counted as follows:

$$T'_{\text{z}}(t) = Q_{\text{h}}(t)/(q_{\text{h}} \cdot V) + T_{\text{out}}(t) - F_0^{-1}\{T_{\text{ind}}(t)\}, \tag{4}$$

that is presented on Fig. 4.

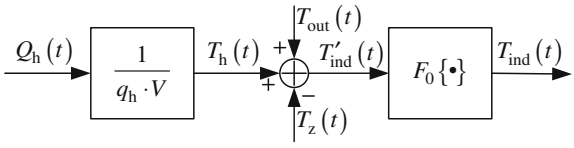


Fig. 3 Block diagram of the building thermal performance dynamics model

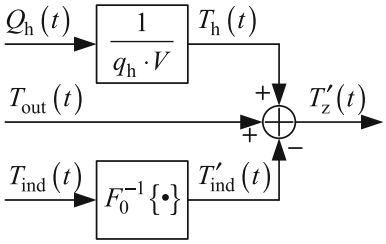


Fig. 4 Block diagram of the building thermal performance inverse dynamics model

The main difference of this approach from other works like [10] is that this is an out-of-the-box solution that includes a way to estimate disturbances.

2.2 Modeling Thermohydraulic Performance of the Building's Heat Station

In addition to building heat exchange processes, the structure and performance of equipment installed in the building's heat station have a significant impact on the heating process.

Heat station is a fairly complex engineering facility. It contains control devices (valves, pumps) with nonlinear properties, as well as process controllers that implement certain control algorithms (typically, a PID controller with control signal depending on outdoor air temperature in accordance with heating curve). A heat station of this sort is deployed in the academic building, which was used for the research [11].

Heat station of the studied building has a standard design with a pump group on the supply pipeline, control valve on the return pipeline, and a displacement bypass with a check valve to prevent direct flow from supply pipe to return pipe. The building's heating system has a vertical design with down-feed risers and single-pipe radiator connections.

The model represents thermohydraulic processes modelled in *VisSim* visual modeling software on the basis of the model proposed in [12]. The model provides numerical solution of a system of algebraic equations that describe hydraulic processes in the heat medium and transfer of heating energy between the heat medium and heat load. For convenience, the model is represented as an array of independent functional blocks that describe standard elements of the heating system: pipe, control valve, check valve, pump, heat load, etc. Each functional block contains all the equations that describe the processes inherent to the block with a desired approximation. The blocks are tied by unidirectional connections, and the direction of connection corresponds to the positive direction of flow (flow rate $G > 0$). Each connection is a vector of three elements, including absolute pressure P (Pa), mass flow rate G (kg/s), and temperature T (°C), which characterizes parameters of heat medium in the corresponding pipe section. Therefore, connections between the blocks can be associated with physical connection of elements.

The key input signal in this model is the position of stem in control valve Y, 0…1, assigned by the process controller installed in the heat station. The model's output value is heating power Q_h emitted into the rooms of the building by heating radiators.

The simplified Piping and instrumentation diagram (P&ID) of heat station and building's heating system is shown in Fig. 5. As we can see from [11], this is a nonlinear model.

The model includes distributed elements (pipelines and heating radiators). Models of key functional blocks built in accordance with the above approach are described in [11].

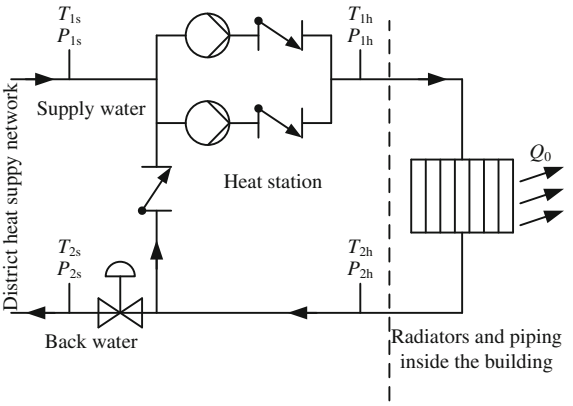


Fig. 5 Block diagram of the building thermal performance dynamics model

2.3 Heat Consumption Estimation Using Reduced Number of Measured Data

When heating meter is not installed, it is impossible to measure heating energy directly. In this case heating energy consumption can be estimated using various indirect methods. Two methods came out to be viable.

Indirect flow calculation.

Heating energy is often calculated by heating meters as follows (Fig. 5):

$$Q = c \cdot G_s \cdot (T_{1s} - T_{2s}), \tag{5}$$

where c stands for thermal capacitance of heat medium, G_s is flow rate, T_{1s} and T_{2s} are heat medium temperatures in the supply and return pipelines, respectively.

In case heating meter is not installed, there is also no flow meter. However, most heating stations are equipped with pressure sensors. In this case, flow rate can be determined in accordance with the following formula:

$$G_h = \sqrt{(P_{1h} - P_{2h})/S_h}, \tag{6}$$

where P_{1h} and P_{2h} are values of pressure in the supply and return pipelines of heating system, respectively, S_h stands for pressure loss (hydraulic resistance) of heating system. The value of S_h is constant for heating systems without individual room control and it can be determined once using portable flow meter:

$$S_h = (P_{1h} - P_{2h})/G_h^2. \tag{7}$$

Therefore, heating energy, consumed by building heating system, can be calculated as follows (Fig. 5):

$$Q = c \cdot \sqrt{(P_{1h} - P_{2h})/S_h} \cdot (T_{1h} - T_{2h}). \quad (8)$$

Heating energy estimation using indoor temperature.

Let us determine heating system of a building as an equivalent heating radiator. Therefore, heat consumption Q can be estimated as follows:

$$Q = k_h \cdot F_h \cdot (T_h - T_{ind}), \quad (9)$$

where k_h is a heat transfer coefficient of radiator, F_h is a heating area of radiator, T_h is an average temperature of radiator which can be estimated as follows:

$$T_h = (T_{1h} + T_{2h})/2. \quad (10)$$

Here, the values of k_h and F_h are constant and can be determined once using Eq. (5) ≡ (9) by substituting (10) in (9) and (6) in (5):

$$k_h \cdot F_h = c \cdot G_h \cdot (T_{1h} - T_{2h})/((T_{1h} + T_{2h})/2 - T_{ind}). \quad (11)$$

where G_h can be measured using portable flow meter.

In order to achieve higher accuracy, it is recommended to perform series of experiments, estimating $k_h \times F_h$ as an arithmetic average value.

When $k_h \times F_h$ is determined, the heat consumption can be estimated using formula (9).

Compared to Indirect flow calculation, this approach achieves lower accuracy, however it can be applied to the systems with variable hydraulic parametres, e.g. systems with individual room control.

3 Application of Proposed Heating Control Solutions for South Ural State University Academic Building

3.1 Approach to Measuring Indoor Air Temperature Using Field-Level Sensor Network

Theoretical studies and experiments show that air temperature in different spaces of the same building may vary significantly as a result of exposure to various perturbing factors described above. This is why building heating control based on indoor air temperature in a single reference area proved to be practically unviable [13]. To solve this problem, we deployed a distributed network of *Dallas 18B20* digital temperature sensors measuring temperature in different premises of the building [14]. In addition, we conducted an experimental study of wireless data transfer from sensors to *RFM XDM2510HP* embedded communication modules integrated into a WirelessHART wireless sensor network [11, 15]. This technology

showed its viability during the case study and will be used in our further research on the subject matter of this study.

The indoor air temperature T_{ind} of a building, which is the average value of indoor temperatures in each room, accounting for differences in area, is calculated as follows:

$$T_{\text{ind}}(t) = \left(\sum_i S_i \cdot T_{\text{ind}i}(t)\right) \Big/ \left(\sum_i S_i\right), \tag{12}$$

where S_i, $T_{\text{ind}i}$ stand for the area and temperature of the i-th room, respectively, and t is time. Using the average temperature T_{ind} permits us to estimate relatively fast perturbations, such as wind, solar radiation, or local heat sources, which affect thermal performance of some rooms, e.g., the rooms of one side of the building—for the entire building.

3.2 Identification of Model Parameters Based on Academic Building Data

To produce a model suitable for a real asset, this study included identification of parameters of heating system deployed in one of the academic buildings of South Ural State University. Building and heat station model parameters were determined in the course of experiment and borrowed from the archival data of existing SCADA system. The following basic model parameters were obtained during the identification process:

- time constant of the building thermal performance model $T_{\text{bld}} = 20$ h;
- net lag of building heat performance model $\tau_{\text{bld}} = 1$ h;
- average specific heat loss of the building $q_{\text{h}} = 0.09$ W/(m^3 °C).

Parameters of the control valve and pumps installed in the heat station were defined using original datasheets. Pipeline lengths and diameters used in heat station model are based on the data specified in design and operating documents for the building heating system.

Feed-forward temperature controller was implemented as a PID controller with gains: $K_{\text{p}} = 0.1$, $K_{\text{i}} = 1/9000$, $K_{\text{d}} = 0$.

Verification of the resulting model revealed modeling error for indoor air temperature in the range of ±1 °C, and the value of root mean square modeling error (RMSE) is 0.263 °C. Indoor air temperature chart is shown in Fig. 6, and air temperature modeling error is presented in Fig. 7.

The value of mean absolute percentage error (MAPE) for heating power Q_{h} is 5.7 %. The chart of heating power applied to the heating system is shown in Fig. 8.

Energy saving effect of the proposed approach is in the stable control of indoor air temperature at a comfort level, along with a significant reduction of impact of

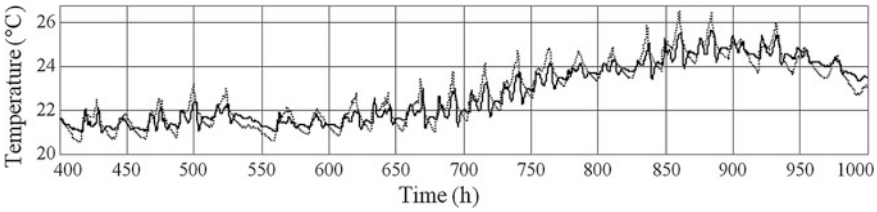


Fig. 6 Indoor air temperature. *Dotted line* stands for actual value; *solid line* stands for modelling value

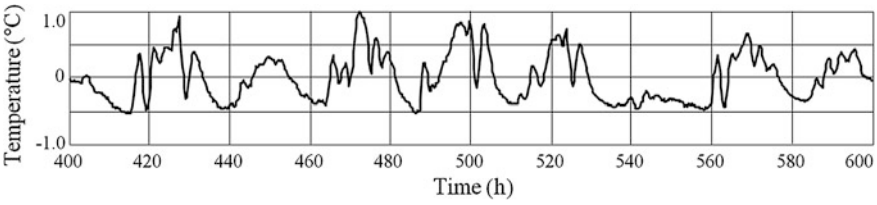


Fig. 7 Indoor air temperature modeling error

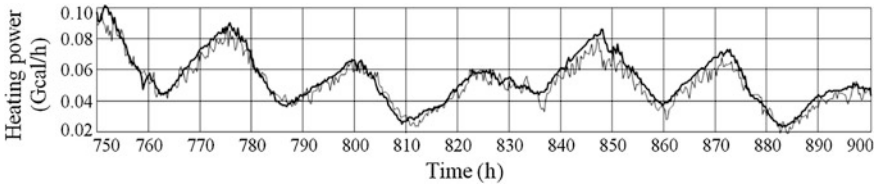


Fig. 8 Heating power applied to the heating system. *Thick solid line* stands for actual value; *thin solid line* stands for modelling value

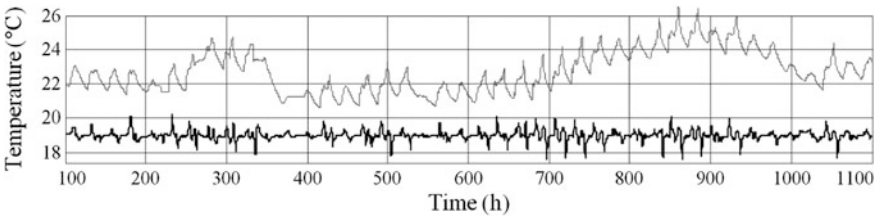


Fig. 9 Indoor air temperature. *Thin solid line* stands for actual value in case of control by outdoor temperature; *thick solid line* stands for expected value in case of feed-forward control deployment

perturbing factors and automatic compensation of statistical control error caused either by inaccuracy of temperature chart, or by the effects of structural changes in the building's thermal performance. Figure 9 shows a chart of actual indoor air temperature fluctuations for the baseline control option and the expected reduction of fluctuations after deployment of feed-forward control.

3.3 Savings Determination

Below is the analysis of system deployment benefit based on the International Performance Measurement and Verification Protocol (IPMVP).

3.3.1 Stage 1

The first stage of the building's heating system automation process was intended to deploy a standard automatic control system that adjusts consumption of heat medium depending on the outdoor air temperature.

According to IPMVP, a statistical model is required to benchmark system efficiency before and after completion of energy conservation measures (ECM) in the reporting period under comparable conditions. A linear regression model is used to build the statistical model for the reporting period. Average daily outdoor temperature and heat medium temperature at building inlet are considered for factors of the model.

The baseline period was determined as the period from March 20, 2013 to April 17, 2013, which corresponds to the reporting period from March 20, 2014 to April 17, 2014. March 20 is the date of ECM completion and system commissioning in 2014. April 17 is the end date of the heating season in 2013, which ended earlier than in 2014.

We considered two factor combinations to build the statistical model for the reporting period.

The first combination of factors includes outdoor temperature and heat medium temperature at building inlet.

Heat energy consumption E_D (Gcal) can be estimated as follows:

$$E_D = G \cdot (T_{1h} - T_{2h}), \tag{13}$$

where G is a heat medium flow rate, T_{1h} and T_{2h} stand for heat medium temperatures at building inlet and outlet respectively. For radiators heat consumption E_R can be estimated using (9) where $E_R = Q$. Heat losses E_L of a building can be estimated as follows:

$$E_L = q \cdot (T_{ind} - T_{out}), \tag{14}$$

where q is a heat transfer coefficient of a building, T_{ind} is an indoor temperature, T_{out} is an outdoor temperature.

Substituting (10), (13), (14) in (9), we get the relationship:

$$E_R = k_h \cdot F_h \cdot (T_{1h} - E_D/(2 \cdot G) - E_L/q - T_{out}), \tag{15}$$

According to heat balance ($E_D = E_R = E_L$), let us determine E_D using (15):

$$E_D = (T_{1h} - T_{out})/(1/(k_h \cdot F_h) + 1/(2 \cdot G) + 1/q). \quad (16)$$

Parameters k_h, F_h and q were permanent because there was no ECM in the baseline period. And there was no automatic heat control system in the building. For this reason heat medium flow rate G was supposed to have no significant changes in the baseline period. T_{1h} could be changed by district heating plant. So T_{1h} and T_{out} are variables. So daily heat energy consumption E_D is calculated in the first model as follows:

$$E_D = a_1 \cdot (T_{1h} - T_{out}), \quad (17)$$

where (T_{1h}–T_{out}) is a variable and a_1 is a coefficient determined as follows:

$$a_1 = 1/(1/(k \cdot F) + 1/(2 \cdot G) + 1/q). \quad (18)$$

As the result of the regression analysis $a_1 = 0.15$. The coefficient of determination R_2 equals 0.95 with confidence level of 0.95 ($n_1 = 28$ days). The value is higher than 0.75, that proves high statistical significance of the model in accordance with IPMVP. The selected factor passes the Student's t-test ($t_1 = 2.06 < t_{estim} = 22.68$), and the P-value ($P = 4 \times 10^{-19}$) is much lower than the significance point of 0.05, which indicates high statistical significance of this factor. Standard error SE_1 of daily heat energy consumption in the first model is 2.08 Gcal.

The second combination of factors includes indoor and outdoor temperatures. In this case daily heat energy consumption E_D is calculated using only (14) where difference (T_{ind}–T_{out}) is a daily degree-days factor D_d. The degree-day values for each day were calculated by subtracting the values of outdoor air temperature from the nominal value of indoor temperature (20 °C). So E_D is calculated in the second model as follows:

$$E_D = a_1 \cdot D_d, \quad (19)$$

where $a_1 = q$, which had no changes in the baseline period.

As the result of the regression analysis $a_1 = 0.43$. For the period under consideration, the coefficient of determination R^2 equals 0.97 with confidence level of 0.95. The value is also higher than 0.75. The degree-days factor passes the Student's t-test ($t_1 = 2.06 < t_{estim} = 28.96$), and the P-value ($P = 7 \times 10^{-22}$) is also much lower than the significance point of 0.05. Standard error SE_1 of daily heat energy consumption in the second model is 1.65 Gcal.

A comparison between the results of two models shows that the second model is better than the first model. It can be explained by lesser sensitivity of the second model to unaccounted factors. Figure 10 represents heat energy consumption in the baseline period. It should be noticed that there were significant changes in outdoor temperature and heat consumption in the period from April 1 to April 3. These great

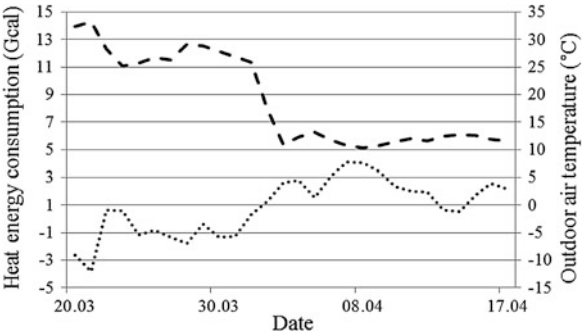


Fig. 10 Heat energy consumption in the baseline period. *Dotted line* stands for outdoor air temperature; *dashed line* stands for heat energy consumption

changes of heat consumption also could be caused by manual correction of heat medium flow rate. But this factor cannot be taken into account in the model because information about values of manual corrections will not be available after implementation of the automatic heat control system. So the degree-days factor model is chosen due to its high robustness to unaccounted factors.

Then adjusted-baseline heat energy consumption was calculated for the reporting period using (19). Figure 11 represents actual and adjusted-baseline heat energy consumption in the reporting period.

Standard error of daily heat energy consumption in the second model is 1.65 Gcal. Standard error of adjusted-baseline heat energy consumption for the entire baseline period SE_{ABL1} is calculated as follows [16]:

$$SE_{\mathrm{ABL1}} = \sqrt{n \cdot SE_1^2}. \tag{20}$$

Calculated standard error for the entire reporting period is 8.72 Gcal.

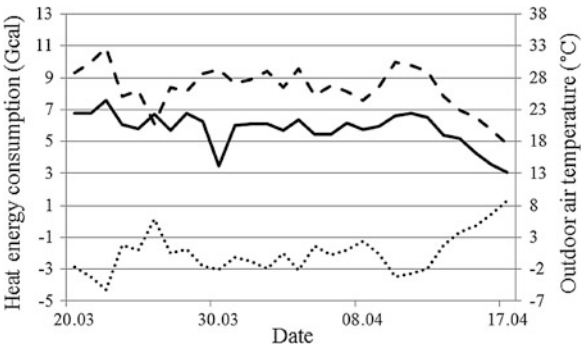


Fig. 11 Adjusted-baseline and reporting-period heat energy consumption in the reporting period. *Dotted line* stands for outdoor air temperature; *dashed line* stands for actual heat energy consumption; *solid line* stands for adjusted-baseline heat energy consumption

Absolute error of heat energy consumption for the entire period AE_{ABL1} in the model with account for the t-value (t_1 = 2.06) is calculated as follows:

$$AE_{\mathrm{ABL1}} = SE_{\mathrm{ABL1}} \cdot t. \tag{21}$$

Calculated absolute error for the entire reporting period is 17.93 Gcal.

Estimated adjusted-baseline heat energy consumption in the reporting period E^{*}_{ABL1} under comparable conditions without account for the error is 234.06 Gcal. Actual value of consumed heat energy E_{RP1} is 162.14 Gcal. The value of savings E^{*}_{ECO1} without account for error is $E^{*}_{\mathrm{ECO1}} = E^{*}_{\mathrm{ABL1}} - E_{\mathrm{RP1}} = 71.92$ Gcal.

The value of savings without account for error in the reporting period is more than four times higher than the standard error. This meets the condition of acceptable uncertainty. According to IPMVP [16], savings must be at least twice higher than standard error.

Thus, savings E_{ECO1} with account for error AE_{ABL1} total 71.92 ± 17.93 Gcal.

The value of minimum relative savings with account for error AE_{ABL1} is calculated using the following formula:

$$E^{\min}_{\mathrm{REL\,ECO1}} = \frac{(E^{*}_{\mathrm{ABL1}} - AE_{\mathrm{ABL1}}) - E_{\mathrm{RP1}}}{E^{*}_{\mathrm{ABL1}} - AE_{\mathrm{ABL1}}} \cdot 100\,\%. \tag{22}$$

Minimum relative savings in the reporting period total 25.0 %.

3.3.2 Stage 2

Stage 2 is intended for practical deployment of feed-forward control approach proposed in this study based on the inverse dynamics model. Source model of baseline consumption was calibrated in accordance with the building's heat meters in the period from March 20, 2014 to April 17, 2014, which corresponded to the reporting period of the previous stage. The model for the upgraded system was built by introducing a feedback control module.

Figure 12 represents modeling results for one month of the heating season.

Model calibration standard error of hourly heat energy consumption after outlier filtering is as follows: SE_2 = 0.007 Gcal. Model calibration standard error for the entire period SE_{C} is calculated using the following formula:

$$SE_{\mathrm{C}} = \sqrt{n_2 \cdot SE_2^2}, \tag{23}$$

where n_2 is a sample size (n_2 is 501 after outlier filtering). Calculated standard error of model calibration for the entire period is 0.16 Gcal.

Model calibration absolute error for the entire period AE_{C} with account for the t-value (t_2 = 1.97) is calculated using the following formula:

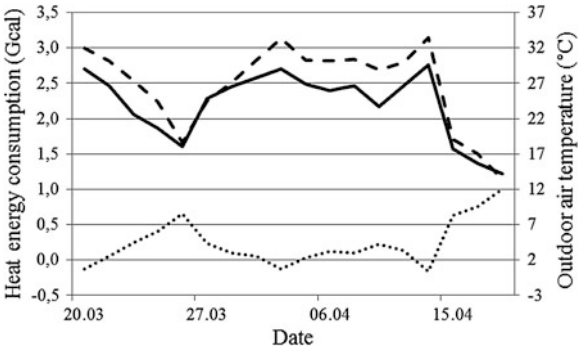


Fig. 12 Heat energy consumption before and after ECM under comparable conditions. *Dashed line* stands for baseline heat energy consumption before ECM; *solid line* stands for heat energy consumption after ECM; *dotted line* stands for outdoor air temperature chart

$$AE_{\mathrm{C}} = SE_{\mathrm{C}} \cdot t_2. \tag{24}$$

Absolute error of model calibration for the entire period AE_{C} is 0.32 Gcal.

Total heat energy consumption is calculated after model calibration and outlier filtering. Estimated total baseline heat energy consumption E^*_{BL2} without account for error is 54.08 Gcal. Estimated total report-period heat energy consumption E^*_{RP2} without account for error is 48.15 Gcal.

The value of minimum relative savings with account for error AE_{C} is calculated using the following formula:

$$\begin{aligned} E^{\min}_{\mathrm{REL\,ECO2}} &= \frac{(E^*_{\mathrm{BL2}} - AE^{\max}_{\mathrm{C}}) - (E^*_{\mathrm{RP2}} - AE^{\max}_{\mathrm{C}})}{E^*_{\mathrm{BL2}} - AE^{\max}_{\mathrm{C}}} \cdot 100\% \\ &= \left(E^*_{\mathrm{BL2}} - E^*_{\mathrm{RP2}}\right) \cdot 100\,\% \big/ \left(E^*_{\mathrm{BL2}} - AE^{\max}_{\mathrm{C}}\right). \end{aligned} \tag{25}$$

Based on the modeling results, additional relative savings due to practical deployment of the authors' proposed approach were estimated at 10.9 %.

4 Conclusion and Future Work

The results obtained demonstrate the overall viability of the proposed approach that accounts for the indoor air temperature and employs feed-forward control based on thermal performance inverse dynamics model, as well as prove high efficiency of this approach in automatic heating control systems. If deployment of baseline control by outdoor air temperature showed in practice a minimum 25 % saving of heating energy, according to the simulation results it is expected to obtain about 10 % extra saving of heating energy in case of deploying the proposed feed-forward control.

Deployment and experimental studies of the proposed heating feed-forward control in hardware in the studied academic building of South Ural State University started in January 2015 and will be completed at the end of the heating season in April 2015.

References

1. Tabunschikov YA, Brodach MM (2002) Mathematical modeling and optimization of building thermal efficiency. (Russian: Математическое моделирование и оптимизация тепловой эффективности зданий), AVOK-PRESS, Moscow
2. Salmerón JM, Álvarez S, Molina JL, Ruiz A, Sánchez FJ (2013) Tightening the energy consumptions of buildings depending on their typology and on climate severity Indexes. Energy Build 58:372–377
3. Zhou D, Park SH (2012) Simulation-assisted management and control over building energy efficiency—a case study. Energy Procedia 14:592–600
4. Shnayder DA, Shishkin MV (2000) Adaptive controller for building heating systems applying artificial neural network. (Russian: Адаптивный регулятор отопления здания на основе искусственных нейронных сетей). In: Automatics and control in technical systems, Edited book. South Ural State University press, Chelyabinsk, pp 131–134
5. Moroşan P-D (2011) A distributed MPC strategy based on Benders' decomposition applied to multi-source multi-zone temperature regulation. J Process Control 21:729–737
6. Salsbury T, Mhaskar P, Qin SJ (2013) Predictive control methods to improve energy efficiency and reduce demand in buildings. Comput Chem Eng 51:77–85
7. Žáčeková E, Prívara S, Váňa Z (2011) Model predictive control relevant identification using partial least squares for building modeling. In: Proceedings of the 2011 Australian control conference, AUCC 2011, Article number 6114301, pp 422–427
8. Abdullin VV, Shnayder DA, Kazarinov LS (2013) Method of building thermal performance identification based on exponential filtration. In: Lecture notes in engineering and computer science: proceedings of the world congress on engineering, vol III, WCE 2013, 3–5 July 2013, London, pp 2226–2230
9. Gorelik SI, Kazarinov LS (1994) Prediction of random oscillatory processes on the basis of the exponential smoothing method. (Russian: Прогнозирование случайных колебательных процессов на основе метода экспоненциального сглаживания), Avtomat i telemekh, vol 10, pp 27–34
10. Thomasa B, Soleimani-Mohsenib M, Fahlén P (2005) Feed-forward in temperature control of buildings. Energy Build 37:755–761
11. Abdullin VV, Shnayder DA, Basalaev AA (2014) Building heating feed-forward control based on indoor air temperature inverse dynamics model. In: 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, pp 886–892
12. Shishkin MV, Shnayder DA (2004) Modelling thermo-hydraulic processes using Vissim simulation environment. (Russian: Моделирование теплогидравлических систем в среде VisSim). In: Bulletin of the South Ural State University, Series Computer Technologies, Automatic Control, Radio Electronics, vol 3., 9(38):120–123
13. Panfyorov SV (2009) Adaptive control system of building thermal conditions. (Russian: Адаптивная система управления тепловым режимом зданий). In: Proceedings of the 7th

Russian national scientific and engineering conference ‘Automatic systems for research, education and industry’, Novokuznetsk, pp 224–228
14. Shishkin MV, Shnayder DA (2002) Automated process monitoring and control (SCADA) system based on MicroLan. (Russian: Автоматизированная система мониторинга и управления технологическими процессами на основе сети MicroLan). In: New software for business of Ural region, Issue 1, proceedings of the regional scientific and engineering conference, Magnitogorsk State Technical University press, Magnitogorsk, pp 84–89
15. Shnayder DA, Abdullin VV (2013) A WSN-based system for heat allocating in multiflat buildings. In: 2013 36th international conference on telecommunications and signal processing proceedings, TSP 2013, 2–4 July 2013, Rome, Article number 6613915, pp 181–185
16. International performance measurement and verification protocol: concepts and options for determining energy and water savings, vol 1. EVO 10000-1:2012, Efficiency Valuation Organization (EVO)

Comparisons of Vector Control Algorithms for Doubly-Fed Reluctance Wind Generators

Milutin Jovanović, Sul Ademi and Jude K. Obichere

Abstract This chapter presents a comparative development and thorough assessment of vector control methodologies for a promising brushless doubly-fed reluctance generator (BDFRG) technology for adjustable speed wind turbines. The BDFRG has been receiving increasing attention in research and industrial communities due to the low operation & maintenance costs afforded by the partially-rated power electronics and the high reliability of brushless construction, while offering competitive performance to its commercially popular and well-known slip-ring counterpart, the doubly-fed induction generator (DFIG). The two robust, machine parameter independent control schemes, one with flux (field) vector orientation (FOC) and the other voltage vector-oriented (VOC), have been built and their response examined by realistic simulation studies on a custom-designed BDFRG fed from a conventional 'back-to-back' IGBT converter. The high quality of the simulation results has been experimentally validated on a laboratory BDFRG test facility under VOC conditions as a preferred control option at large-scale wind power levels.

Keywords Brushless · Doubly-Fed machines · Reactive power control · Reluctance generators · Vector control · Velocity control · Wind power

M. Jovanović (✉) · J.K. Obichere
Faculty of Engineering and Environment, Northumbria University Newcastle, Newcastle upon Tyne NE1 8ST, UK
e-mail: milutin.jovanovic@northumbria.ac.uk

J.K. Obichere
e-mail: jude-kennedy.obichere@northumbria.ac.uk

S. Ademi
Institute for Energy and Environment, Department of Electronic and Electrical Engineering, University of Strathclyde, Glasgow G1 1RD, UK
e-mail: sul.ademi@strath.ac.uk

H.K. Kim et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-94-017-7236-5_7

1 Introduction

The brushless doubly-fed generator (BDFG) has been considered as a viable alternative to the traditional DFIG for wind energy conversion systems (WECS) [1–7]. In these applications, where limited variable speed ranges of 2:1 or so are required [2, 3, 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 additional cost advantages of higher reliability and maintenance-free operation by the absence of brush gear [9, 10].

The BDFG has two standard sinusoidally distributed 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. 1), the brushless doubly-fed reluctance generator (BDFRG) [1–3, 7], appears to be more attractive than its 'nested' cage rotor form, the brushless doubly-fed induction generator (BDFIG) [4–6, 11, 12]. This preference has been mainly attributed to the prospect for higher efficiency [7] and simpler control[1] associated with the cage-less reluctance rotor [16]. However, the BDFG rotor must have half the total number of stator poles to provide the rotor position dependent magnetic coupling between the stator windings required for the machine torque production [17].

An introduction of the grid codes and strict regulations for wind turbine generators to comply with under abnormal operating conditions [18], have revealed another important and attractive BDFG merit, the superior low-voltage-fault-ride-through (LVFRT) capability to the DFIG [15, 19]. It has been shown that owing to the larger leakage inductances and lower fault current levels, the LVFRT of the BDFG may be accomplished safely without crowbar circuitry in contrast to DFIG [18, 20].

Various control algorithms[2] have been developed for the BDFRG including scalar control [3, 22], primary voltage vector-oriented control (VOC) [3, 13, 14, 22], primary flux (field)-oriented control (FOC) [13, 14], direct torque and flux control [22, 23], direct torque and reactive power control [24, 25], direct power control [26], sliding mode power control [27], and even non-linear Lyapunov control theory [8]. Although a comparative qualitative analysis of some of these control methods has been partly made in [22] (and more detailed for the DFIG in [15, 21]), no similar quantitative study, other than that in [22], has been reported specifically on FOC versus VOC of large BDFRG wind turbines, and there has been very little theoretical or applied work published on true FOC for this machine in general. An exception is the latest contribution made in [13] where the two control techniques have been experimentally verified and compared on a small

[1]Field-oriented control of the primary reactive power and electromagnetic torque is inherently decoupled in both the BDFRG [13, 14] and DFIG [15], but not in the BDFIG [4, 11].

[2]A good literature review on control of the BDFIG can be found in [4–6], and of the DFIG in [15, 21].

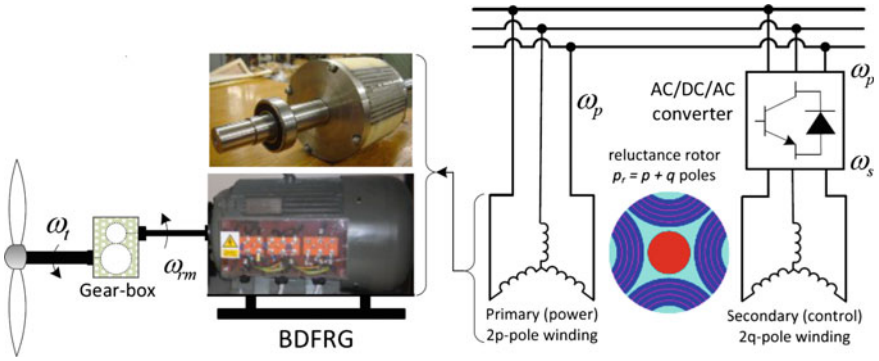


Fig. 1 A conceptual diagram of the doubly-fed reluctance wind generator (DFRWG)

BDFRG(M) prototype but for fixed loading scenarios of no immediate interest to the target applications.

It is interesting that the FOC and VOC terms have often been interchangeably used in the literature to indicate the same control approach despite their quite distinctive meanings. With a proper selection of the reference frames, they indeed become very similar in nature and response, especially with larger machines of lower resistances considered here. Nevertheless, they have clear differences and performance trade-offs to be pointed out in the following using a custom-designed 2 MW BDFRG [2] as an example. The maximum torque per inverter ampere (MTPIA) control strategy has been implemented because of the efficiency gain by minimising the secondary current magnitude for a given load, and hence reducing both the winding copper and inverter switching losses [3, 28]. Another strong reason for looking at this particular control objective is purely technical and associated with practical implications of the insufficient current rating, and other design limitations, of the proof of concept BDFRG [24–26] used to validate the VOC theory presented in the sequel. This prevented common close to unity primary power factor values to be achieved for real-time testing. Extensive realistic simulation studies taking into account the usual practical effects (e.g. transducers' DC offset, noise in measurements, and a PWM power converter model) are produced to support the discussions. These are accompanied by the experimental results generated on a scaled-down BDFRG prototype to demonstrate the controller viability and to verify the high accuracy of the simulated models under challenging speed dependent loading conditions encountered with typical horizontal-axis wind turbines, the characteristics of which have been emulated in a laboratory environment. In this sense, this chapter is a significant extension in scope to the content of both [14] and [13].

2 Basic BDFRG Theory

The BDFRM(G) dynamic model in arbitrary rotating reference frames, using standard notation and assuming motoring convention, can be represented as [1]:

$$\left.\begin{aligned} \underline{v}_{p,s} &= R_{p,s}\cdot\ \underline{i}_{p,s} + \frac{d\underline{\lambda}_{p,s}}{dt} = R_{p,s}\cdot\ \underline{i}_{p,s} + \frac{\partial\underline{\lambda}_{p,s}}{\partial t} + j\omega_{p,s}\cdot\ \underline{\lambda}_{p,s} \\ \underline{\lambda}_p &= \underbrace{L_p i_{pd} + L_{ps} i_{sd}}_{\lambda_{pd}} + j\cdot\underbrace{(L_p i_{pq} - L_{ps} i_{sq})}_{\lambda_{pq}} \\ \underline{\lambda}_s &= \lambda_{sd} + j\cdot\lambda_{sq} = \sigma L_s \underline{i}_s + \underbrace{\frac{L_{ps}}{L_p}\underline{\lambda}_p^*}_{\underline{\lambda}_{ps}} \end{aligned}\right\} \tag{1}$$

where the primary and secondary winding are denoted by the subscripts 'p' and 's' respectively, $\sigma = 1 - L_{ps}^2/(L_p L_s)$ 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, torque and mechanical power relationships for the machine with p_r rotor poles and $\omega_{p,s} = 2\pi f_{p,s}$ applied frequencies to the respective $2p$-pole and $2q$-pole windings (Fig. 1) are [1]:

$$\theta_{rm} = \frac{\theta_p + \theta_s}{p_r} \Leftrightarrow \omega_{rm} = \frac{d\theta_{rm}}{dt} = \frac{\omega_p + \omega_s}{p_r} \Leftrightarrow n_{rm} = 60\ \cdot\ \frac{f_p + f_s}{p_r} \tag{2}$$

$$T_a = J\cdot\frac{d\omega_{rm}}{dt} = \underbrace{\frac{3p_r}{2}(\lambda_{ps_d} i_{sq} - \lambda_{ps_q} i_{sd})}_{T_e} - T_L(\omega_{rm}) - F\cdot\omega_{rm} \tag{3}$$

$$P_m = T_e\cdot\omega_{rm} = \underbrace{\frac{T_e\cdot\omega_p}{p_r}}_{P_p} + \underbrace{\frac{T_e\cdot\omega_s}{p_r}}_{P_s} = P_p\cdot(1 + \frac{\omega_s}{\omega_p}) \tag{4}$$

where $\omega_s > 0$ for 'super-synchronous' operation, $\omega_s < 0$ at 'sub-synchronous' speeds (i.e. an opposite phase sequence of the secondary to the primary winding) in (2), and $\omega_{syn} = \omega_p/p_r$ is the synchronous speed (for $\omega_s = 0$ i.e. a DC secondary) as with a $2p_r$-pole wound rotor synchronous turbo-machine. Notice that all the ω_p rotating vectors in the primary voltage/flux equations in (1) are in ω_p frame, while the corresponding secondary counterparts, including the λ_{ps} components in (3), are stationary in $p_r\omega_{rm} - \omega_p = \omega_s$ frame [1]. Given that λ_p and λ_{ps} in (3) are approximately constant by the primary winding grid connection, torque control can be achieved through the secondary dq currents in the ω_s frame. The power flow in the primary winding is to the grid for the generating ($T_e < 0$) regime under consideration, while the secondary can either take or deliver real power (P_s) subject to the ω_s sign; the BDFRG would absorb (produce) $P_s > 0$ at sub (super)-synchronous speeds.

3 Controller Design

A structural diagram of the primary voltage/flux angle and frequency estimation technique in discrete form with appropriate *dq* frame alignment options for VOC/FOC blocks is shown in Fig. 2. The entire BDFRG system layout with a generic controller design is presented (Fig. 3). A standard phase-locked-loop

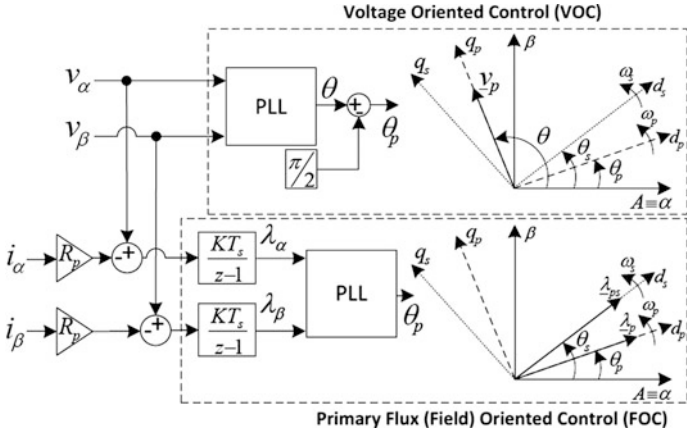


Fig. 2 Angular position estimation of primary voltage and flux vectors in a stationary $\alpha-\beta$ frame

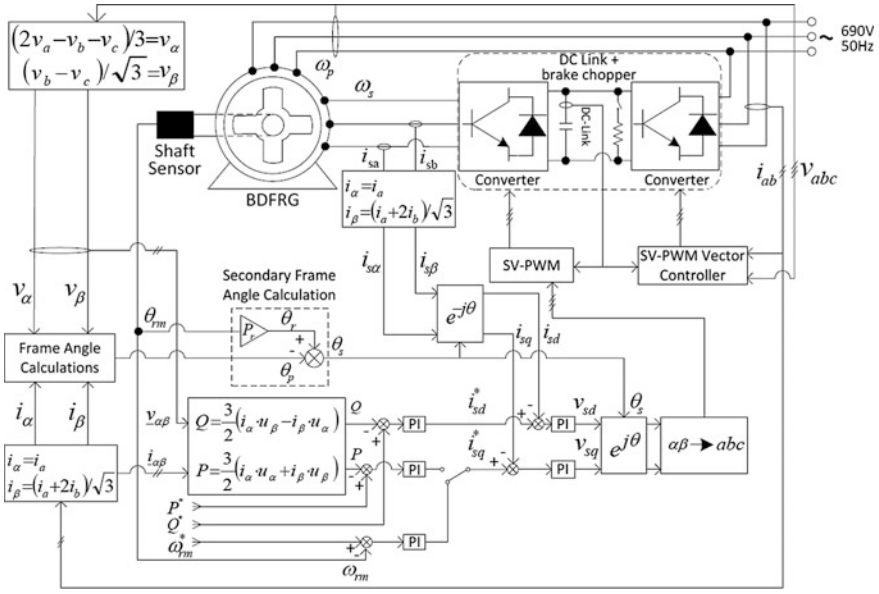


Fig. 3 A structural schematic of the BDFRG with vector control of power electronics converter

(PLL) algorithm, readily available in the *Simulink* library, has been used to retrieve θ and/or θ_p in Fig. 2 from the measured voltages and/or currents. A conventional vector controller with space-vector PWM of the active rectifier has been implemented to regulate the DC link voltage at unity line power factor [29]. The real (P) and reactive (Q) power calculations are reference frame invariant and have been done using the stationary frame voltages ($v_{\alpha\beta}$) and currents ($i_{\alpha\beta}$) as the least computationally intensive allowing the highest control rates 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 desired setting in power (P^*) or closed-loop speed mode (ω^*_{rm}). For example, either P^* or ω^*_{rm} may correspond to the Maximum Power Point Tracking (MPPT) of a wind turbine [3, 8] while Q^* may be chosen to optimize certain performance indicator of the machine like torque per secondary ampere in this chapter.

4 Voltage-Oriented Control (VOC)

The control form expressions can be derived from (1) in the natural reference frames, ω_p (e.g. $d_p - q_p$ for primary winding) and ω_s (e.g. $d_s - q_s$ for secondary winding) rotating frames (Fig. 2), where the respective vector components are DC quantities. Substituting for $\underline{i}_p$ from the $\underline{\lambda}_p$ equation of (1) into $\underline{S}_p = \frac{3}{2}\underline{v}_p\underline{i}^*_p$ would lead to the following relationships for the primary mechanical and reactive power:

$$P_{p_{voc}} = \frac{3}{2}\omega_p(\lambda_{ps_d} i_{sq} - \lambda_{ps_q} i_{sd}) = P_{p_{foc}} - \frac{3}{2}\omega_p \lambda_{ps_q} i_{sd} \tag{5}$$

$$Q_{p_{voc}} = \frac{3}{2}\omega_p\left(\frac{\lambda_p^2}{L_p} - \lambda_{ps_d} i_{sd} - \lambda_{ps_q} i_{sq}\right) = Q_{p_{foc}} - \frac{3}{2}\omega_p \lambda_{ps_q} i_{sq} \tag{6}$$

VOC of P_p and Q_p is coupled as both the i_{sd} and i_{sq} secondary current components appear in (5) and (6). The level of coupling can be reduced by aligning the q_p-axis of the reference frame to the primary voltage vector as proposed in Fig. 2. In this case, the primary flux vector ($\underline{\lambda}_p$) 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 as in Fig. 2, VOC should be similar to FOC as $\lambda_{ps_d} \gg \lambda_{ps_q}$ i.e. $\lambda_{ps_d} \approx \lambda_{ps}$ so that (5) and (6) become:

$$P_{p_{voc}} \approx P_{p_{foc}} = \frac{3}{2}\omega_p \lambda_{ps} i_{sq} = \frac{3}{2}\frac{L_{ps}}{L_p}\omega_p \lambda_p i_{sq} \tag{7}$$

$$Q_{p_{voc}} \approx Q_{p_{foc}} = \frac{3}{2}\frac{\omega_p \lambda_p^2}{L_p} - \frac{3}{2}\omega_p \lambda_{ps} i_{sd} = \frac{3}{2}\frac{\omega_p \lambda_p}{L_p}(\lambda_p - L_{ps} i_{sd}) = \frac{3}{2}\omega_p \lambda_p i_{pd} \tag{8}$$

The P_p versus i_{sq} and Q_p versus i_{sd} functions above are nearly linear, which justifies the use of PI control in Fig. 3.

5 Flux-Oriented Control (FOC)

The primary flux oriented (e.g. with the reference frame d_p-axis aligned to $\underline{\lambda}_p$ as in Fig. 2) forms of the flux equations in (1) and torque in (3) become [1, 22]:

$$\underline{\lambda}_p = \underbrace{L_p i_{pd} + L_{ps} i_{sd}}_{\lambda_{pd}=\lambda_p} + j \cdot \underbrace{(L_p i_{pq} - L_{ps} i_{sq})}_{\lambda_{pq}=0} \tag{9}$$

$$\underline{\lambda}_s = \underbrace{\sigma L_s i_{sd} + \lambda_{ps}}_{\lambda_{sd}} + j \cdot \underbrace{\sigma L_s i_{sq}}_{\lambda_{sq}} = \sigma L_s \underline{i}_s + \underbrace{\frac{L_{ps}}{L_p}\lambda_p}_{\lambda_{ps}} \tag{10}$$

$$T_e = \frac{3p_r L_{ps}}{2L_p}\lambda_p i_{sq} = \frac{3p_r}{2}\lambda_{ps} i_{sq} = \frac{3p_r}{2}\lambda_p i_{pq} \tag{11}$$

The corresponding real and reactive power are now given by (7) and (8).

The most important advantage of FOC over VOC is the inherently decoupled control of P_p (or T_e) and Q_p through i_{sq} and i_{sd} variations, respectively, which is immediately obvious from (7), (8) and (11). This fact greatly facilitates the FOC design. However, these appealing FOC properties come at the cost of the $\underline{\lambda}_p$ angle estimation (θ_p in the FOC block of Fig. 2) and difficulties with suppressing the troublesome DC offset effects on the voltage integration accuracy. In addition, the primary winding resistance (R_p) generally needs to be known, and particularly with smaller machines. As entirely parameter independent, the VOC approach does not suffer from any of these FOC constraints but has compromised load disturbance rejection abilities and inferior control quality as a trade-off.

6 Simulations of 2 MW BDFRG Turbine

The preliminary performance comparisons of the FOC/VOC schemes in Fig. 3 have been carried out using the parameters of a large custom-designed BDFRG [2] summarized in Table 1. In order to make the simulations as genuine 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 superimposed to the ideal signals to account for practical effects of

Table 1 The BDFRG design parameters and ratings

Rotor inertia [J]	3.8 kg^2	Power [P_r]	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

the measurement noise and current/voltage transducers errors; (iii) Finally, the rotor position and speed information has been provided by an incremental encoder.

In a typical WECS, the turbine output torque on the generator side of the gear-box for the MPPT in the base speed region (i.e. between the 'cut-in' and the rated wind speed), can be formulated as [3, 8]:

$$T_{mppt} = \frac{A \cdot \rho \cdot C_p(\lambda_{mppt}, \gamma) \cdot R^3}{2 \cdot g^3 \cdot \lambda_{mppt}^3} \cdot \omega_{rm}^2 = K_{mppt} \cdot \omega_{rm}^2 \tag{12}$$

where ρ is the air density, $C_p(\lambda, \gamma)$ is the power (performance) coefficient (i.e. the maximum turbine efficiency as $\lambda = \lambda_{mppt}$ in this case), $\lambda_{mppt} = R\omega_t/u$ 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 maximize C_p), R the radius of the circular swept area ($A = \pi R^2$), and $g = \omega_{rm}/\omega_t$ is the gear ratio. The shaft torque-speed profile in (3) is of the same form as (12):

$$T_L = -\frac{P_r}{\omega_r} \cdot \left(\frac{n_{rm}}{n_{\max}}\right)^2 \approx -19 \cdot \left(\frac{n_{rm}}{1000}\right)^2 \text{kN m} \tag{13}$$

The simulation results in Figs. 4, 5 and 6 have been produced by running the control algorithms in Fig. 3 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). The reference speed trajectory is set as a steep ramp signal suited for dynamically not very demanding turbines even at extremely turbulent winds.

The top plots in Fig. 4 show the precise speed tracking potential of the controller upon the BDFRG start-up, the response being virtually identical for the FOC and VOC. The primary power (P) and torque (T_e) curves accurately follow the outputs of (13) for given speed settings. Except for a difference in losses, and considering that $\omega_p \approx$ const, P and T_e are directly related according to (4) and (5) which explains a close resemblance in their shape. One can also hardly see any disparity between the FOC and VOC results under the MTPIA conditions ($i_{sd} = 0$) when (5) effectively becomes (7). The reactive power (Q) is controlled at ≈1.35 MVAr, obtained from (8) for $i_{sd} = 0$ and $\lambda_p \approx u_p/\omega_p$, to meet the MTPIA objective. Note that whilst the Q behavior with the decoupled FOC is largely unaffected by the P

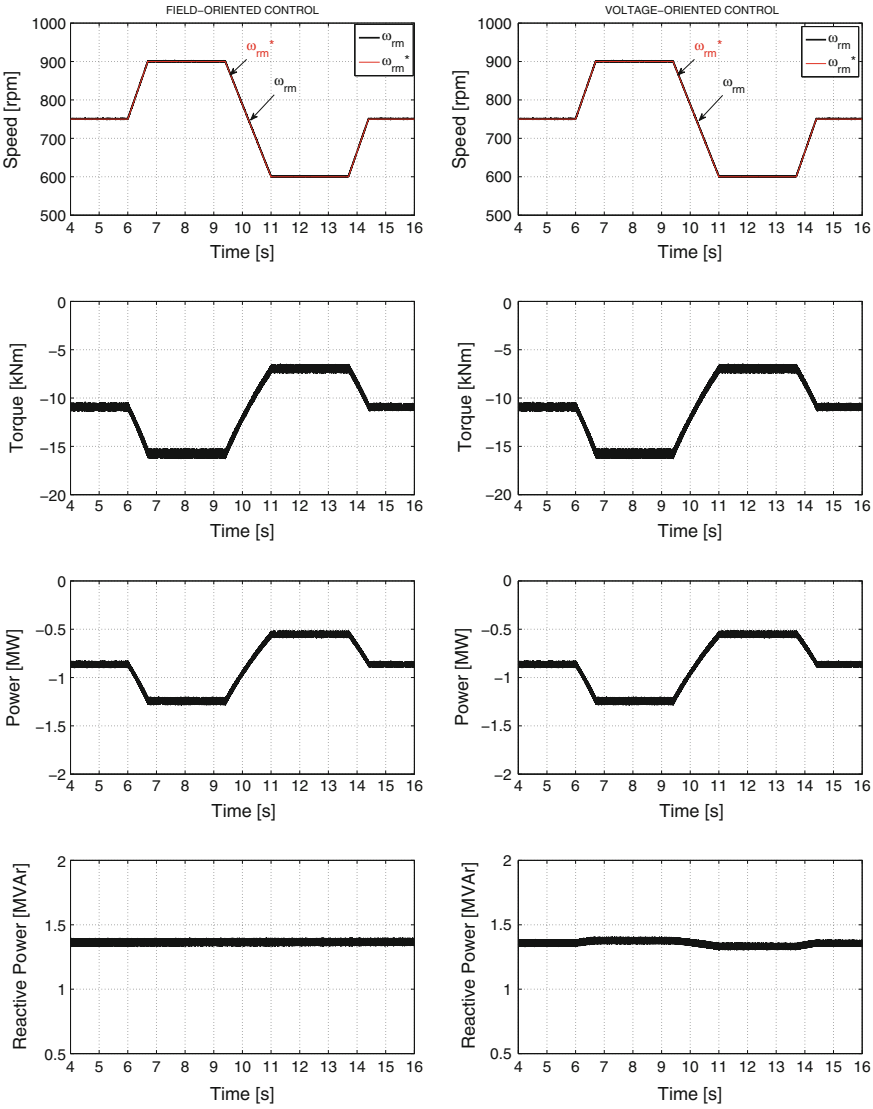


Fig. 4 Simulation of the BDFRG operation in a narrow range around synchronous speed

variations, it is rather distorted in the VOC case by the presence of the coupling i_{sq} term in (6).

The secondary ($i_{sd,q}$) and primary ($i_{pd,q}$) current waveforms in Fig. 5 are notably smooth with no transient over-currents as the PI regulators do not need to be saturated to handle 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 graphs. The coupling effects of the i_{sq}

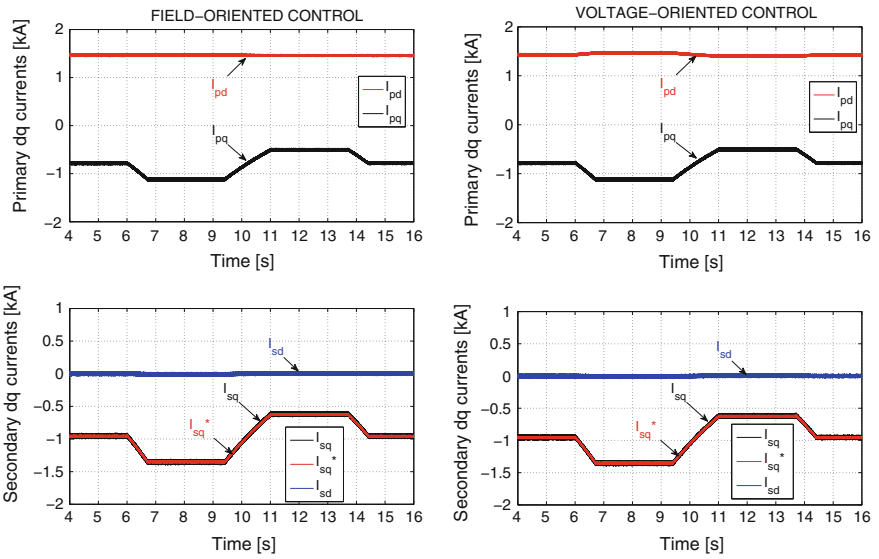


Fig. 5 Simulated MTPIA responses of the BDFRG current components in the corresponding rotating reference frames

clearly manifest themselves as speed (torque) dependent disturbance (e.g. offsets) in the non-controllable i_{pd} profiles by analogy to the VOC scenario for P and Q. The FOC i_{pd} (and Q) levels, on the other hand, are constant in average throughout.

Figure 6 illustrates the secondary currents of the BDFRG while riding through the synchronous speed (750 rev/min) from 900 to 600 rev/min. The respective PWM sector step-changes of $\underline{v}_s$ can be found in [14]. In the super-synchronous mode at 900 rev/min, the secondary vectors rotate in a positive (anti-clockwise)

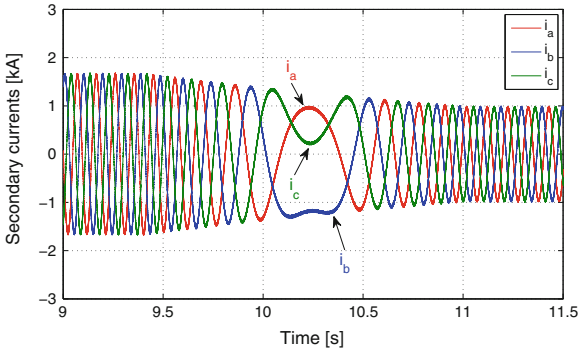


Fig. 6 Simulated BDFRG secondary current waveforms showing a phase sequence reversal from positive (‘*abc*’) to negative (‘*acb*’) during transition from super to sub-synchronous speed

direction and $\omega_s > 0$ in (2). The secondary phase sequence is reversed at sub-synchronous speeds (600 rev/min) so $\omega_s < 0$ in (2) for the clockwise rotating vectors. The latter are stationary around 750 rev/min at DC secondary currents i.e. $\omega_s = 0$ in (2).

7 Experimental Verification

The validity of the simulation model has been proven using the BDFRG test rig described in [24–26]. The prototype Y/Y connected, 6/2-pole machine is rated at 1.5 kW, 1000 rev/min, 2.5 A, 400 V, 50 Hz and has the following parameters: $J = 0.1$ kg^2, $R_p = 11.1\ \Omega$, $R_s = 13.5\ \Omega$, $L_p = 0.41$ H, $L_s = 0.57$ H, and $L_{ps} = 0.34$ H. The plots in Figs. 7 and 8 are respectively produced by simulating and executing the VOC scheme in Fig. 3 on a Simulink® compatible dSPACE® platform at 5 kHz inverter switching rate under similar operating conditions as simulated in Fig. 4. 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 (13) but with 14 N m loading at 1000 rev/min in real-time.

Figures 7 and 8 demonstrate an excellent tracking of reference speeds in a limited range around synchronous speed (750 rev/min). It is important to point out that the measured and simulated speeds are virtually identical indicating the high reliability of the developed Simulink® model. The T_e estimates in the same figure adequately reflect the BDFRG desired mechanical input (T_L) determined by (13) except during the speed transients by the presence of deceleration or acceleration torque (e.g. T_a in (3)) depending on whether the machine is to slow-down ($T_a < 0$) or speed-up ($T_a > 0$). These differences between T_e and T_L profiles do exist for the large BDFRG too but are not visible with the scaling adopted in Fig. 4.

Although a generally good agreement between the simulation and experimental results is more than obvious from Figs. 7 and 8, the test waveforms for both the T_e and P are apparently shifted away from the simulated counterparts as the PI speed regulator necessarily has to over-estimate the $i_{sq} \sim T_e$ (i.e. under-estimate $|i_{sq}|$ as $i_{sq} < 0$) reference to compensate for the rotational (mainly core) losses of the real machine not being accounted for by the torque controller. It should also be noted that unlike the large-scale case in Fig. 4 both the simulated and experimental Q responses in Figs. 7 and 8 are severely disrupted by the speed dependent P variations as a consequence of the much more pronounced cross-coupling effects of the i_{sq} term in (6) coming from the unusually high winding resistances of the test BDFRG.

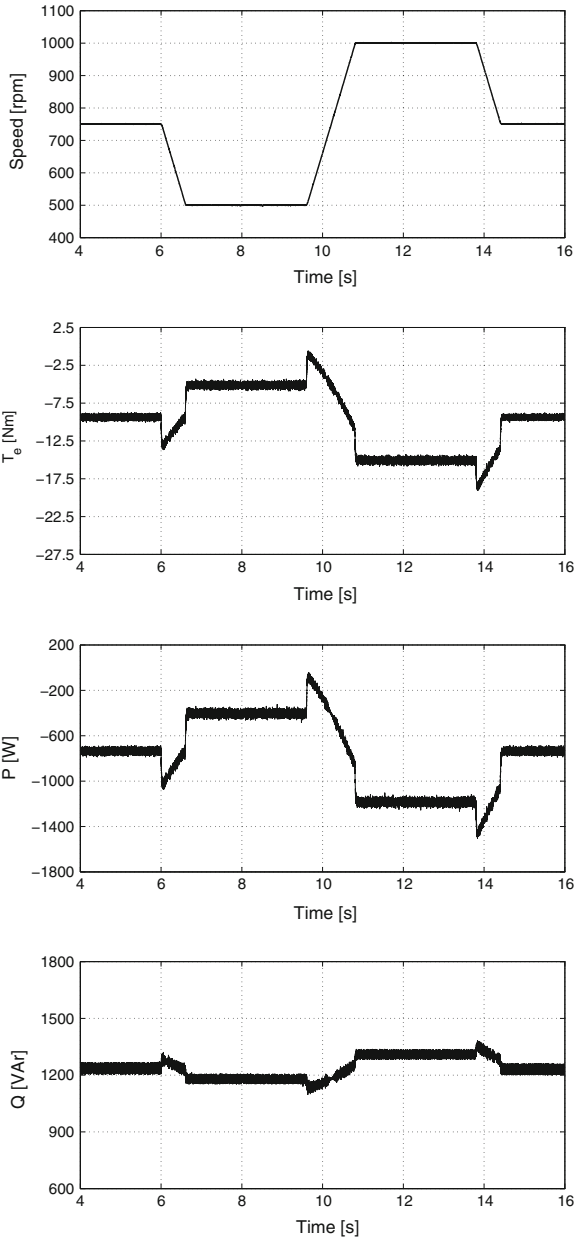


Fig. 7 Simulink® results for VOC of the BDFRG speed, torque, primary real and reactive power down to synchronous speed at $Q^* \approx 1250$ VAr in Fig. 3

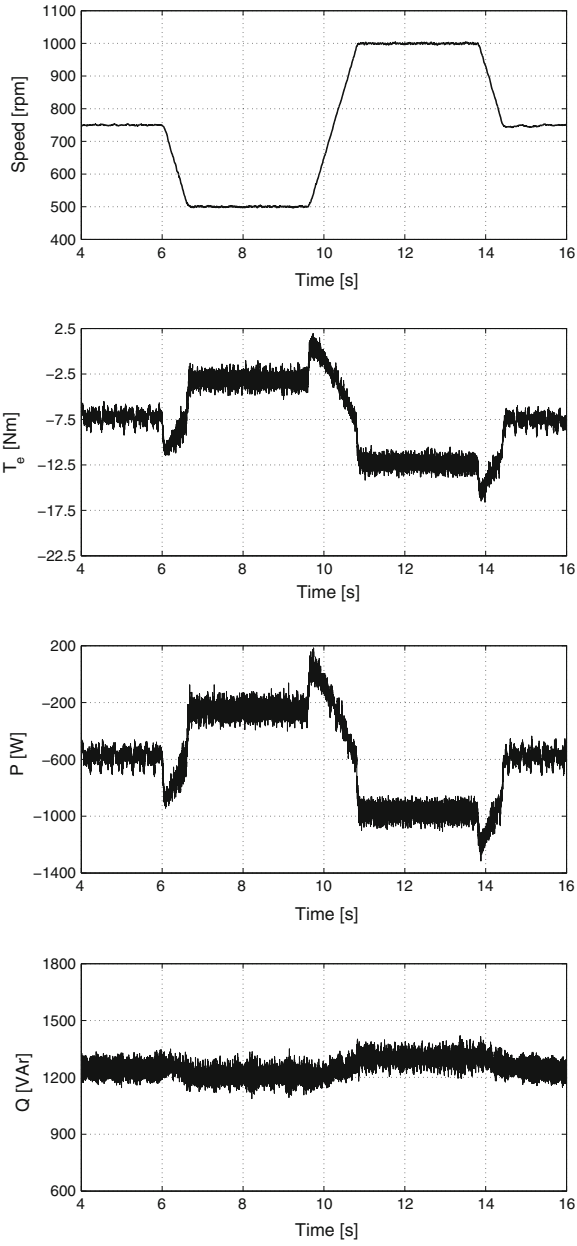


Fig. 8 Experimental validation of the simulations in Fig. 7 for the small-scale BDFRG

8 Conclusions

This chapter has concerned with theoretical and practical aspects of high performance control with voltage (VOC) and flux space-vector orientation (FOC) for the BDFRG—an emerging candidate for variable speed constant frequency applications such as wind or marine turbines (but also pump-alike drives), where the cost advantages of partially-rated power electronics and reliability of brush-less construction as well as the LVFRT prospects can be fully exploited. The truthful Simulink® studies have indicated the undoubt potential and effectiveness of the controller in both VOC and FOC modes. These salient properties have been successfully experimentally substantiated by the presented VOC results for typical variable loading conditions of a small-scale BDFRG featuring high winding resistances when the uncompensated cross-coupling effects on the reactive power response are the most detrimental. The accuracy of the developed simulation model should be continuously improving, and existing differences between the VOC and FOC responses gradually disappearing, with increasing machine sizes (and hence lowering winding resistances), which warrants its use as a promising low-cost and user-friendly computing alternative to expensive specialist software commonly used for investigating the operation of larger wind turbines of similar design to that under consideration.

Possible directions for future research may include the sensitivity analysis of the controller robustness to the BDFRG inductance variations, the MPPT implementation using natural wind profiles, and/or upgrading the machine model to incorporate iron losses. These remedial steps would further enhance the torque control standard making the entire simulation programme even more representative and versatile.

References

1. Betz RE, Jovanović MG (2003) Introduction to the space vector modelling of the brushless doubly-fed reluctance machine. Electric Power Compon Syst 31(8):729–755
2. Dorrell DG, Jovanović 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
3. 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
4. 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
5. 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
6. 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
7. 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):1237–12431

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 GJW, Koutoulakos E (2009) Reliability of wind turbine subassemblies. IET Renew Power Gener 3(4):387–401
11. 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 Industr Electron 60 (6):2477–2485
12. 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
13. Ademi S, Jovanović M (2015) Vector control methods for brushless doubly fed reluctance machines. IEEE Trans Industr Electron 62(1):96–104
14. Ademi S, Jovanovic M, Obichere JK (2014) Comparative analysis of control strategies for large doubly-fed reluctance wind generators. Lecture notes in engineering and computer science: proceedings of the world congress on engineering and computer science 2014, WCECS 2014, pp 245–250, San Francisco, USA, 22–24 Oct 2014
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 Industr Electron 60(7):2776–2798
16. Knight A, Betz R, Dorrell D (2013) Design and analysis of brushless doubly fed reluctance machines. IEEE Trans Ind Appl 49(1):50–58
17. Hsieh MF, Lin IH, Dorrell D (2013) Magnetic circuit modeling of brushless doubly-fed machines with induction and reluctance rotors. IEEE Trans Magn 49(5):2359–2362
18. Long T, Shao S, Malliband P, Abdi E, McMahon R (2013) Crowbarless fault ride-through of the brushless doubly fed induction generator in a wind turbine under symmetrical voltage dips. IEEE Trans Industr Electron 60(7):2833–2841
19. 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
20. Long T, Shao S, Abdi E, McMahon R, Liu S (2013) Asymmetrical low-voltage ride through of brushless doubly fed induction generators for the wind power generation. IEEE Trans Energy Convers 28(3):502–511
21. 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
22. Jovanović M (2009) Sensored and sensorless speed control methods for brushless doubly fed reluctance motors. IET Electr Power Appl 3(6):503–513
23. Jovanović 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
24. Chaal H, Jovanović M (2012) Practical implementation of sensorless torque and reactive power control of doubly fed machines. IEEE Trans Industr Electron 59(6):2645–2653
25. Chaal H, Jovanovic M (2012) Toward a generic torque and reactive power controller for doubly fed machines. IEEE Trans Power Electron 27(1):113–121
26. Chaal H, Jovanovic M (2012) Power control of brushless doubly-fed reluctance drive and generator systems. Renew Energy 37(1):419–425
27. Valenciaga F (2010) Second order sliding power control for a variable speed-constant frequency energy conversion system. Energy Convers Manag 52(12):3000–3008
28. Betz RE, Jovanovic MG (2002) Theoretical analysis of control properties for the brushless doubly fed reluctance machine. IEEE Trans Energy Convers 17(3):332–339
29. 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

Diagnosis of Alarm Systems: A Useful Tool to Impact in the Maximization for Operator's Effectiveness at Power Plants

Eric Zabre and Víctor Jiménez

Abstract Alarm system, alarm rationalization and interface design play a critical and important role to determine the ability and effectiveness of operators at power plants. Due to this, an adequate and proper alarm system impact to the answer and attention in the presence of abnormal situations from a control room at power plant. In industrial processes, such as petrochemical, paper, electricity, among others, it is necessary to optimize the management of resources in order to guarantee the work team, equipment and installation's safety. In the past decade, alarm management has positioned itself as the most important of priorities in term of safety aspects, but to get this an exhausted review, diagnostic, and improvement labors to alarm system have been required. In this chapter, a system prototype known as Diagnosis System ASARHE to diagnosis generating unit at power plants, is presented. It provides as results the performance operative state of alarm systems, previous to depuration these last through the application of an alarm rationalization methodology based on ANSI/ISA and EMMUA international norms. This diagnosis was fundamental to be able to improve any alarm management system that has a direct impact on the most important human factor of every processing plant: the operator.

Keywords Alarm monitoring · Alarm rationalization · Control room · Distributed control system · Generating unit · Power plant · Safety

E. Zabre (✉) · V. Jiménez
Simulation Department, Instituto de Investigaciones Eléctricas, Reforma 113,
Col. Palmira, 62490 Cuernavaca, Morelos, Mexico
e-mail: ezabre@iie.org.mx

V. Jiménez
e-mail: vmjs@iie.org.mx

H.K. Kim et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-94-017-7236-5_8

1 Introduction

In the past 1960, modernization of distributed control system (DCS) at power plants has provoked changes on the traditional way to operate, supervise, and diagnose the operation of the units at power plants. Technologists, suppliers, and experts in alarms systems administration believed to improve operation the units on having incorporated exorbitant alarms quantities into the DCS, in other words, from a conventional quantity in order of tens or hundreds (80–250) of alarms to quantities of thousands (4000–20,000). This situation of "progress" was considered a myth and this has been confirmed thanks to alarm management.

During the practice of last decades it has been necessary to put attention to situations of risk, accidents, incidents and other aspects related to the safety of the resources at power plants.

Trips quantity of the units increased in an alarming way without being provided necessary with a convincing explanation of why occurs even with better instrumentation, major quantity of signals for supervising, fastest and opportune information delivery to the operator, best improve in communications, in the operator interfaces, more modern equipment, among others. That's why the need to diagnose the performance level of the alarm systems before and after realizing a suitable alarm management.

2 Background

In the 90s, immediately after the creation of the ASM (Abnormal Situation Management) consortium, the EEMUA 191 (Engineering Equipment and Materials Users Association) guide, the NAMUR 102 (Alarm management) recommendation and the review of the already created ANSI/ISA 18.2 norm [1, 2] there has been concern to review the alarms systems for the purpose of improving the generation units.

Additionally programs and diagnosis systems that allow to experts and consultants in the topic to identify where the origin of the problems is located have been developed. Where does the problem initiate? Why the problem exist? How can trips quantity by unit be minimized? What tools can contribute to improve the operation of the generation process? and so on.

All these questions converge on the need for a necessary and urgent diagnosis that identifies elements or areas of progress that contribute substantially to a better operation of the units from the operator, this last being fundamental in the operation of control room at power plants.

On having diagnosed the operative state of units not only anomalous situations are identified, but it is necessary to understand the areas of opportunity that support the operator in a more efficient way to operate the power plant. This finally will

ultimately lead to a better use of any nature resources: human, financial/economical, equipment, raw material, etc.

Most of cases, there is a confusion of what an alarm is. It is very important that before examining alarm management best practices, to understand the complete concept of the alarm meaning. No matter what the process is concerning to, an alarm has the following purposes.

1. To alert of an abnormal change
2. To communicate the nature of the change as well as possible causes
3. To direct to take proper corrective action

> In the best practices, the most important contribution of an alarm is that it needs an action as part of the operator tasks.

As preamble, it is important to mention that the Instituto de Investigationes Eléctricas http://vmwl1.iie.org.mx/sitioIIE/site/indice.php or IIE, as part of the activities of alarm management project, initiated in October 2010, has developed software diagnosis tools for power plants to the CFE (Federal Commission of Electricity), the only company of generation, transmission and distribution of electricity in the Mexican territory. The initial labor was realized in a thermoelectric power plant, gotten the first objectives of alarms rationalization such changes were incorporated into the database of two units, the first one off line and the second one in line with the DCS executing, experiencing the first results and increasing the safety of the systems of the power plant and the operative reliability of the same plant [3]. Figure 1 shows available findings distribution on related analysis and monitoring systems that have contributed to improve alarm system on a global scale.

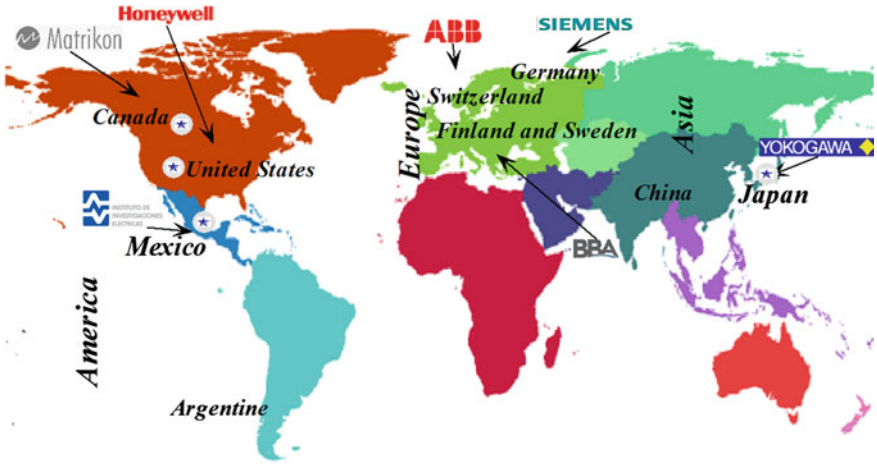


Fig. 1 Benchmark of system diagnosis around the world

3 Related Works

Later to the advent of the DCS, introduced about 1975–80 for different firms as Honeywell®, Mitsubishi®, Yokogawa®, Siemens®, ABB®, among others, and with the change of paradigm through modern computers with big databases, high resolution interfaces and fast communications, have brought a big change in the alarm management. Commercial applications exist for the diagnosis alarm systems as well in petrochemical and gas area [4] and the electrical sector. For example, for the last one, the Matrikon® Alarm Manager's Advanced Analysis [5] that automatically generates a report of unit performance in accordance with parameters established by EEMUA and ANSI/ISA norms; it identifies redundancy in the alarm configuration, chattering alarms, priority distribution, statistics of alarm occurrence, etc. Another similar system designed for identification of related problems with alarms is the Y-Plant Alert™ of Yokogawa, which detects alarms events, avalanche alarms, visualization on screen of the alarms state, events inside certain intervals that provide to the alarm manager useful information of the state of the unit. The IIE has developed a generic tool for the off line analysis of the alarm systems of any unit in a generating at power plant, which purpose is to take it to the practice and extend it like evaluation support and complement to the suggested guidelines by the international norms, for the activities of alarm system rationalization.

All these with certain automation are based on the guidelines of the reference norms as well as on the guide lineaments of PAS alarm management [6].

4 Asarhe® Diagnosis System

The IIE diagnosed the alarms systems in the electrical generation sector in six chosen pilot power plants, by technology type: hydroelectric, coal, combined cycle, diesel, geo-thermoelectric and steam conventional thermoelectric, identifying the absence of the procedures or guide for alarm documentation and classification, even for the alarm design or redesign, and maintenance of the alarm systems, based on international reference norms. Based on the obtained results, a system named ASARHE* (Analysis of Signs of Alarms based on Historical Events Records) has been prepared, of property technology to realize diagnoses of units [7, 8], in addition to preparing the alarms philosophy for every power plant that considers as specifications, the requests that establish the criteria for the management of alarm systems, and how it is applied in other areas.

Due to different commercial DCS, there are also diverse ways of presenting variables to monitor, so there in not standardization on displaying alarms on operator station. This last provoked a mix of signals related to maintenance, communications between software and hardware in order to control and maintain the unit within normal operational conditions.

Different DCS used to make a laborious data interpretation of the alarm system diagnosis.

4.1 Initial Interface

One of the most important aspects to determine the alarm system performance is precisely the quantitative analysis of alarm historical record, in which information of each alarm is stored with tag, name or description, prioritization level criteria (critical, warning or tolerance alarm), set point, and occurrence date and hour.

Figure 2 shows the main interface of the ASARHE to prepare the tasks sequence that must be carried out and parameters for the analysis of the alarm historical files until the result and its interpretation.

The tasks are as follow.

1. Cleaning of previous historical records
2. Power plant selection to be analyzed
3. Unit deployment at power plant
4. DCS type
5. Technology description of power plant
6. Selection of historical record directory
7. Deployment of historical record
8. Download information to ASARHE
9. Exit to process information and graphs deployment

Fig. 2 Initial interface of ASARHE

4.2 Process Flow of the Proposed System

Historical data is stored at the engineering station in a proper format of the DCS, for which is necessary to convert them to an understandable format for the ASARHE to process later the information and deliver the analysis results to the user. Figure 3 shows schematic representation of data conversion, data processes, and alarms distribution graphs.

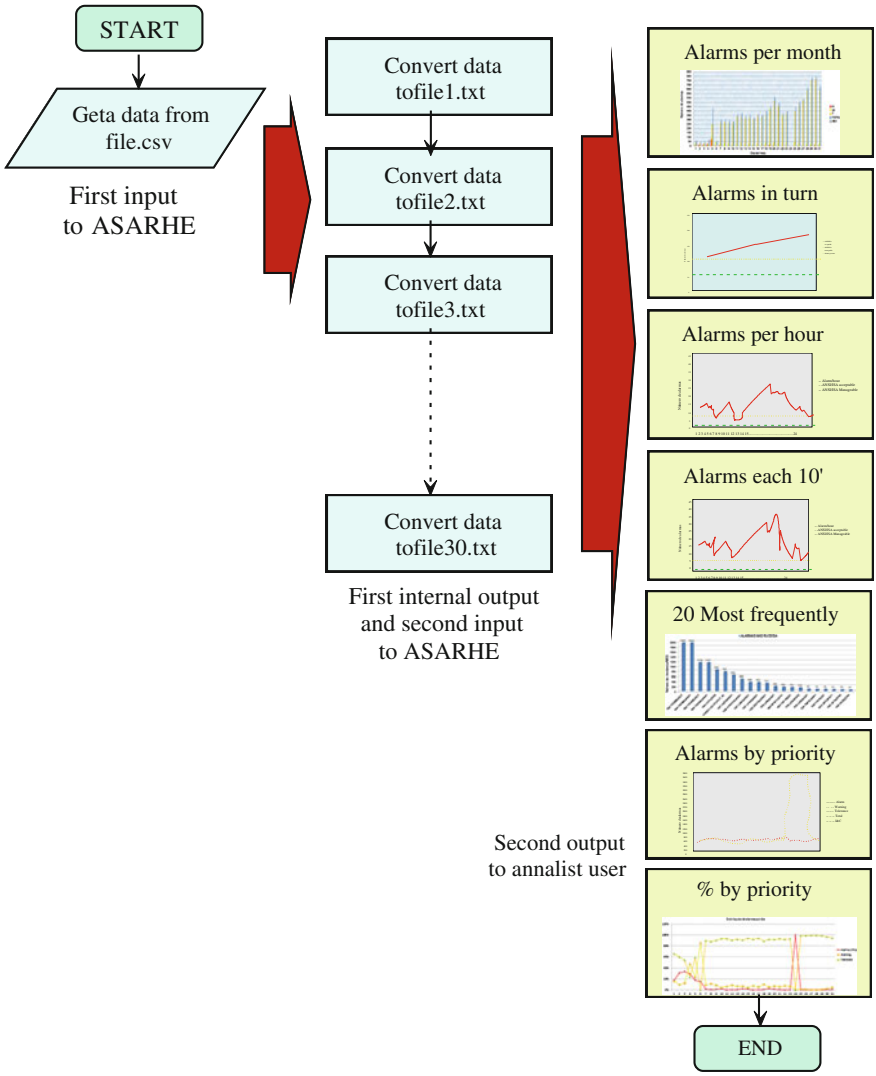


Fig. 3 Flow chart of ASARHE system

4.3 Data Format Conversion

Information conversion initiates from the information registered in the alarm historical record, registered on tape, flexible discs and most modern DCS in order to get the most appropriated selected data. This information is kept in a defined format by the provider and later will be used by the analysts in chief to check the sequence of operations, alarms occurrence, operator's answer, etc. before the destabilization of normal conditions of operation. A typical conversion needed by ASARHE appears in Fig. 4, in which differentiation of the order of the information and the separation of the alarms happened by daily periods for one month can be observed.

A code segment for one data conversion module written in Visual Basic is shown in Fig. 5.

4.4 Identification of Alarm Types and References

As soon as the information is converted to ASARHE format, the graphs are generated with statistics of quantity of alarms per month, for operator's shift, per hour, and per every 10 min. The ANSI/ISA norm establishes that an alarm must appear of the following way. See Figs. 6 and 7.

1. In normal situations, an alarm occurs every 10 min.
2. In a disturbance, during the first 10 min, there will be a maximum of 10 alarms.

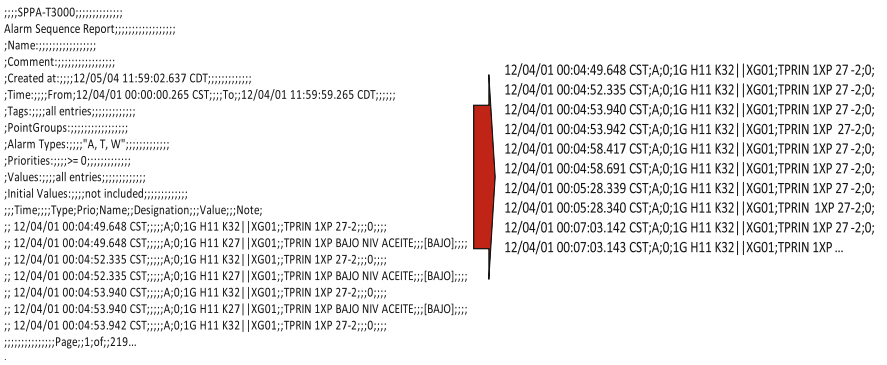


Fig. 4 Data conversion first input to ASARHE

```
Application.ScreenUpdating = False
Application.Calculation = x1CalculationManual
'Application.EnableEvents = False
ActiveSheet.DisplayPageBreaks = False

With ThisWorkbook
   l = 1
   For Each objHojas In Worksheets
    If IsNumeric(objHojas.Name) Then
     H = objHojas.Name
     Sheets(H).Activate
     'Rango de valores celdas a revisar
     Dim xFile As Double

     With.ActiveSheet
        xFile = Cells (Rows.Count, "A").End(x1Up).Row
     End With

     Set rng = .Sheets(h).Range("E1:E" & xFila)
     'Set rng = .Sheets(h).Range("E1:E65500")

     'Llama a la función que saca los elementos únicos
     Lista = UniqueItems(rng, False)
        Sheets(H).Activate
     With.ActiveSheet
     '.Range(Cells(1, "T"), _
     '.Cells(Ubond(Lista), "T:T"))
     '= WorksheetFunction.Traspose(Lista)
        For x = LBound(Lista) To Ubond(Lista)
           Cells(x + 1, 20) = Lista(x)
        Next x
     End With
   End If
   l = l + 1
 Next objHojas
```

Fig. 5 Code segment of 20 most frequently conversion module

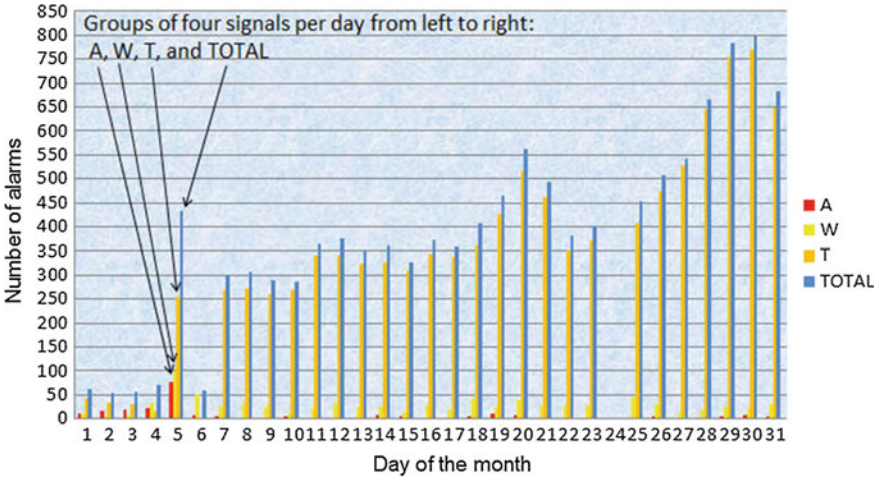


Fig. 6 Type of alarms per period: monthly

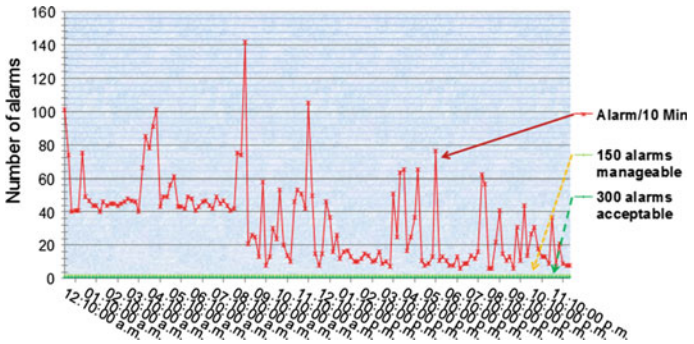


Fig. 7 Type of alarms per period: every 10 min

4.5 *Results Interpretation, Findings, and Performance Level Determination*

As part of the generated results by ASARHE, the 20 most frequent alarms, named *bad actors* are shown. See Fig. 8. From these alarms a tag which serves to identify every instrument and it generally coincides with badly calibrated or aged instrumentation that may need adjustment of its set point.

Nuisance alarms could be duplicate alarms on the system. This is a typical situation between the overloaded and reactive performance level and the identification of these alarms as well point adjustment tasks could be the difference to set the system on a reactive or stable level, and it is often usable in practice during plant upsets.

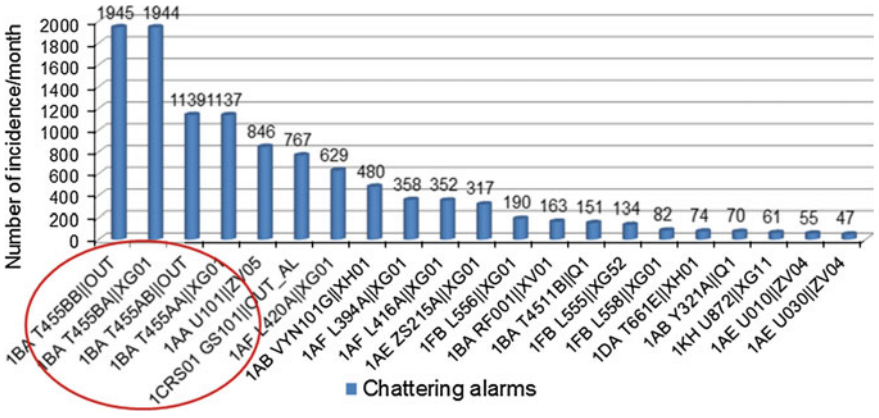


Fig. 8 Bad actors and nuisance alarms

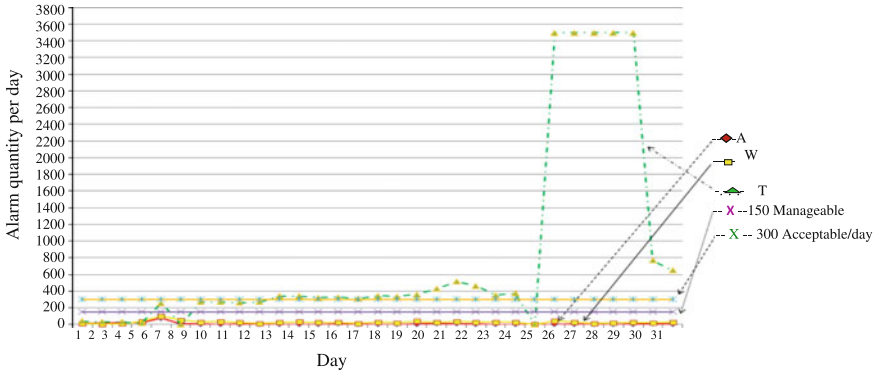


Fig. 9 Finding of critical, warning, and tolerance alarms *quantity per day*

In Fig. 9 also the alarms quantity per day are identified, which is a good indicator of the overall health of the alarm system. In this graph, the most alarms are under maximum acceptable (300) and manageable (150) except during last seven days of the month due to a disturbance occurred in a generating process, but general speaking, the performance of the alarm system was in a good acceptance.

Distribution of alarms occurrence by priority type is shown in Fig. 10.

These last two graphs determine the performance level in accordance with the limits established in the reference norm as indicated in Fig. 11.

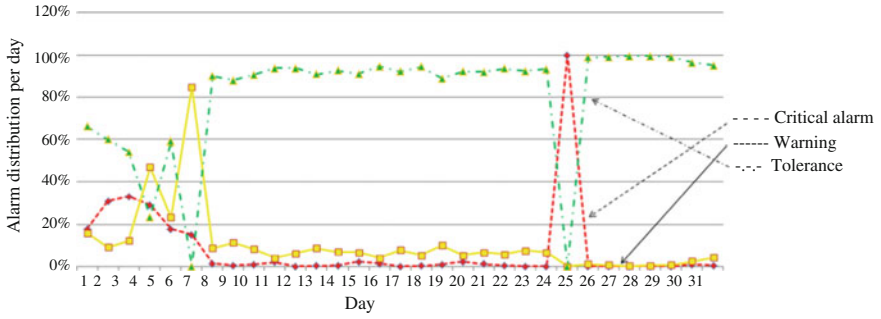


Fig. 10 Distribution alarms by priority per day

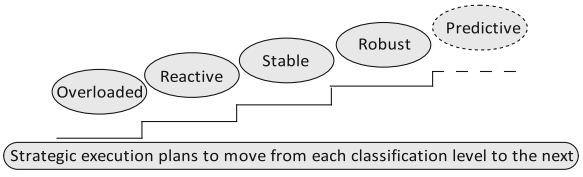


Fig. 11 Performance level of alarm system

Performance levels are [1, 2]: (1) Overloaded—alarms are very difficult to distinguish from less important ones; (2) Reactive—operators react more to the rate of alarm generation than to the purpose of the alarms themselves; (3) Stable—all alarms are meaningful and have a specific response; (4) Robust—operators strongly trust the alarm system, and have time to attend all alarms; and (5) Predictive—alarm system is completely stable and provides the operator with timely, accurate information. EEMUA-191 performance level model, edition 2-2007 recommended this last level as optimum management. For most power plants, to reach this last level depends on the state of the art of the technology at that moment. Same EEMUA-191, third and revision 2-2013, considers robust level as top performance level. This level requires early and adequate fault detection, moving forward the process tendency, as well as incorporating artificial intelligence techniques, among others.

4.6 Preparation for Alarms Rationalization

Common problems refer to the excessive quantity of alarms presented to the operator, to the identification of the chattering alarms, to the distinction of alarms and events, and to the determination of the state of the alarms system.

As previously observed, immediately after DCS modernization, the problems of administration and suitable handling of the alarm systems created disturbances in the control rooms and hence in the power plants where control and monitoring operative processes of whatever the application are. For such reason, since modernization, the concern of restoring alarms systems again has arisen, which is a "regression to" when deployments alarms were done from light box annunciators and appropriate legends allowed the operator to control the normal state of the process (gone are those days). Steps as part of the alarm management of the ANSI/ISA administration cycle under the 18.2 norm is the next labor to be done, which phases shown in Fig. 12 are the following.

Philosophy ①—basic design of alarm system; Identification ②—collection point for potential alarms; Rationalization ③—applying prioritization requirements; Design/re-design ④—basic alarm design, HMI design, and design of advanced alarming techniques; Implementation ⑤—installation alarm system as well as operators training. Operation ⑥—confirm alarm philosophy and purpose of each alarm; Maintenance ⑦—test and adjust if alarm operational is not working properly; Monitoring ⑧—continuously monitoring the overall performance; Changes management ⑨—identifies problem alarms for maintenance; and Audit ⑩—to continuous improvement, closing the alarm management life cycle.

Alarm's operator interface is shown in Fig. 13, which presents different type of alarms during normal operation at a modernized 350 MW unit of a thermal power plant.

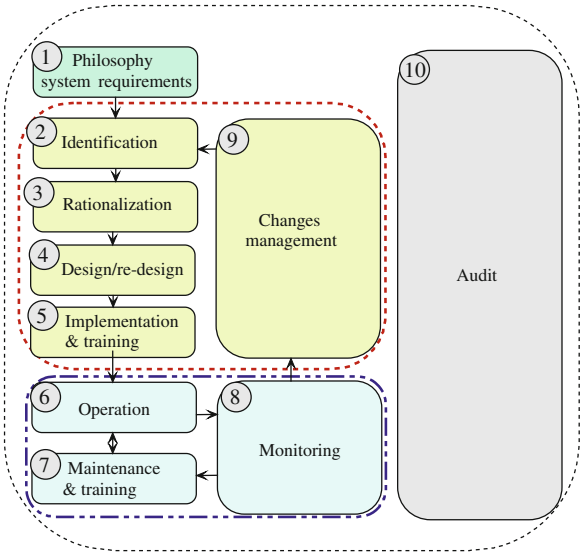


Fig. 12 Alarm management life cycle stages

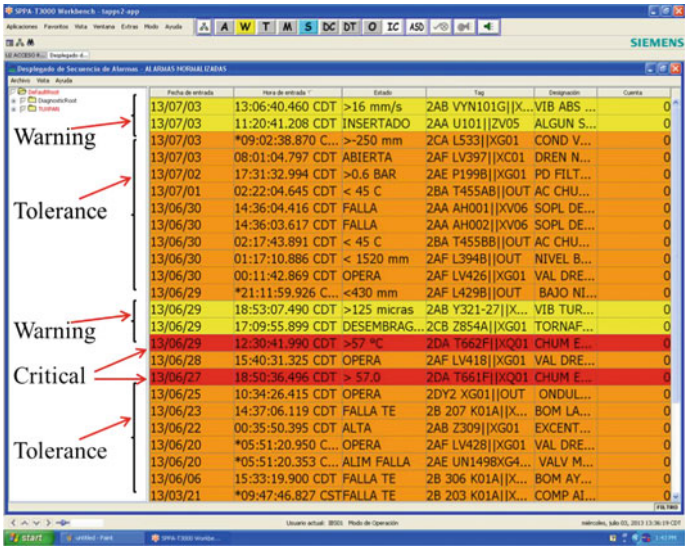


Fig. 13 Alarm's operator interface

5 Alarm Rationalization

Once the unit has been evaluated, it will be necessary to apply the steps of the alarms administration cycle (Fig. 12) and to reduce substantially the quantity of alarms that present to the operator in its interface. This work in general is realized by expert engineers and experienced operators on the operation of the power plant.

1. To prepare the alarm philosophy, definitions and terminology to use, rationalization criteria, alarm priorities definition, deployment criteria of HMI prioritization, monitoring, maintenance plan, test, as well as operators' training. Optimum alarm distribution criteria must be the following. See Fig. 14.
2. To identify the information in the database of the DCS. It is necessary to understand its complete content.
3. To analyze *bad actors*, monitoring the current alarm system and to identify the system performance, to identify the alarm occurrence and to separate alarms from events.
4. To document the alarm book containing the tag, set point, priority and a clear description of (1) Cause of the alarm: why did the alarm occur? (2) Action: what must the operator do to restore the process to its normal condition?, and (3) Consequence: what happens if the alarm is not attended?
5. To focus in the priority and in the possible change of every alarm, which must be checked by the person in charge of the alarm system together with expert engineers in operation, electrical, safety, faults and other related areas.
6. To implement a suitable alarm administration in real-time, this means, to establish a methodology that guarantees the update of every alarm.
7. To keep control of program changes of every alarm.

The first three steps described are always necessary and the last four represent the most arduous part and important labor of alarm rationalization and the performance and optimal operation of the power plant will depend on it.

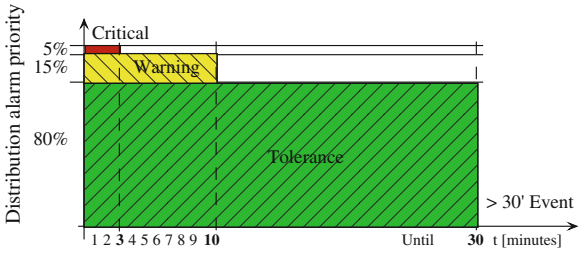


Fig. 14 Operator's response time

6 Conclusion and Future Work

The diagnosis before alarm rationalization is fundamental as improving the performance level of the alarm system in process plants can avoid accidents, losses of production and unnecessary trips unit, that in turn, affect copiously the economic resources in power plants, as well as the reliability, relevant aspect in safety terms [9, 10].

The diagnosis uses a systematical, validated, standardized and highly advisable, comparable methodology on a global scale. In Mexico no reference exists and it represents a challenge of big dimensions for the IIE, since data base systems that operate in the real power plant are re-designed, and any mistake can be of unimaginable consequences.

Up to this moment, there is not another technologist in the country that can diagnose the units of power plants. Nevertheless, the activities of rationalization that the IIE is applying to the alarm systems to have been attended at power plants of the CFE, can be compared with companies on a global scale that also are applying a suitable management rationalization and administration of the alarm systems, such as the cases of products of proprietary analysis of Matrikon, Emerson (United States and Canada), ABB (Switzerland), Finland and Sweden; Siemens (Germany), and Yokogawa (Japan).

From 2010 to 2013, 148 alarm systems were rationalized in 50 different power plants. The alarm books were prepared for every unit. In all units where rationalized alarms were implanted they moved from OVERLOADED to STABLE performance level.

ASARHE system will continue being applied to power plants that modernize its DCS and that adopt, as an integral solution, the alarm management inside the electrical sector as well as like part of its daily activities and of a new culture and continuous progress.

In the future advanced skills of alarm management for optimization of the alarm system will be included. This will contribute to get ROBUST or PREDICTIVE performance level to generating units in accordance to process type and operation mode.

Acknowledgments The authors acknowledge the valuable experience, and enthusiastic personnel's labor during tasks concerning to the used of this application in control rooms at power plants.

This work was supported in part by Comisión Federal de Electricidad, under contract code C-00309GEN.

References

1. ANSI/ISA-18.2-2009, Management of alarm systems for the process industries. ISBN: 978-1-936007-19-6, 23 Jun 2009
2. EEMUA-191, The engineering equipment and materials users association, a guide to design, management and procurement, Ed. 2-2007, ISBN: 0-85931-155-4 and, Ed. 3-2013, ISBN: 978-0-85931-192-2
3. Zabre E et al (2013) Reliability recovery in attending power plants by means of alarm rationalization. In: 2013 IEEE power and energy society general meeting: shaping the future energy industry, Vancouver, Canada, 21–25 Jul 2013, Catalog Number CFP13POW USB ISBN 978-1-4799-1301-5
4. Foong OM et al (2009) ALAP: alarm prioritization system for oil refinery. Lectures notes in engineering and computer science: proceedings of the world congress on engineering and computer science 2009, WECECS 2009, 20–22 October, 2009, San Francisco, USA, pp 1012–1017
5. Matrikon Alarm Manager's Advanced Analysis, 2007
6. Hollifield B, Habibi E (2008) The alarm management handbook, a comprehensive guide. ISBN: 0-9778969-0-0
7. Salinas M et al (2012) Herramienta de diagnóstico y evaluación de los sistemas de alarmas en unidades generadoras de energía eléctrica, Boletín IIE, Año 36 Jul–Sep 2012, vol. 36, núm. 3, ISSN0185-0059
8. Jiménez V, Zabre E (2012) Sistema para el análisis de señales de alarmas basado en registros históricos de eventos en unidades generadoras, Registro INDAUTOR 22/May/2012, 03-2012-050311182000-01
9. HSE (1999) The health and safety executive, 2nd edn. ISBN: 978-0-7176-1709-8
10. Zabre E, Jiménez V (2014) Diagnosis of alarm systems and impact in the maximization of operator's effectiveness at power plants. Lectures notes in engineering and computer science: proceedings of the world congress on engineering and computer science 2014, WCECS 2014, 22–24 October, 2014, San Francisco, USA, pp 289–294

Solutions for the Massive Dirac Equation with Electric Potential, Employing a Biquaternionic Vekua Equation

Marco Pedro Ramirez Tachiquin and Vania Martinez Garza Garcia

Abstract Employing quaternionic analysis and pseudoanalytic functions, we study the massive Dirac equation with electric potential, utilizing one relation with a hyper-complex Vekua equation, from which it is possible to obtain a new class of solutions, considering the elements of the so-called Taylor series in formal powers.

Keywords Biquaternionic functions · Dirac equation · Electric potential · Pseudoanalytic functions · Quaternions · Vekua equation

1 Introduction

The Dirac equation is fundamental in Relativistic Quantum Mechanics, and the study of its solutions is of special relevance in many branches of Theoretical Physics and Engineering, e.g. in Nuclear Medicine. Therefore, the construction of new solutions becomes interesting since it could well allow a dipper understanding of the quantum particles behavior, governed by the equation, under a wide variety of circumstances.

The following pages are dedicated to study one technique that allows the construction of new class of solutions for the massive Dirac equation, with an arbitrary electric potential depending upon one spatial variable interacting with the quantum particles [6], showing that those solutions can be both numerically or analytically constructed.

By utilizing elements of the quaternionic analysis [2], we rewrite the time-harmonic Dirac equation in a quaternionic form, and by considering a special

M.P.R. Tachiquin (✉)
Asaji Audio Internacional S.A. de C.V., Guadalupe I.R. 687, 16030 Mexico, C.P., Mexico
e-mail: marco.ramirez@micei.com.mx

V.M.G. Garcia
The Engineering Faculty of Universidad La Salle, B. Franklin 47, 06140 Mexico, C.P., Mexico
e-mail: vanmtzg@hotmail.com

H.K. Kim et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-94-017-7236-5_9

case of solutions, we obtain a biquaternionic Vekua equation [7], from which it is possible to introduce a generating sequence [1], and in consequence, to pose its general solution in terms of Taylor series in formal powers. Indeed, the biquaternionic Vekua equation constitutes a rediscovering of the results posed in [2] when analyzing a two-dimensional Dirac equation with scalar potential.

Given the solutions of the biquaternionic Vekua equation, we explain in a simplified way the technique to obtain the corresponding solutions for the Dirac equation in classical form, finishing with a brief discussion of the main contributions.

2 Preliminaries

2.1 Elements of Quaternionic Analysis

We will denote the set of complex quaternions [3] (or *biquaternions*) as $\mathbb{H}(\mathbb{C})$, whose elements $Q \in \mathbb{H}(\mathbb{C})$ have the form:

$$Q = q_0 + q_1\mathbf{e}_1 + q_2\mathbf{e}_2 + q_3\mathbf{e}_3, \tag{1}$$

where $q_k \in \mathbb{C}$, $k = 0, 1, 2, 3$. This is $q_k = \operatorname{Re} q_k + i \operatorname{Im} q_k$, being i the standard imaginary unit: $i^2 = -1$. Beside, we have that $\mathbf{e}_1$, $\mathbf{e}_2$ and $\mathbf{e}_3$ are the quaternionic units (e.g. [2]), possessing the following properties of multiplication:

$$\begin{aligned}
\mathbf{e}_1\mathbf{e}_2 &= -\mathbf{e}_2\mathbf{e}_1 = \mathbf{e}_3;\\
\mathbf{e}_2\mathbf{e}_3 &= -\mathbf{e}_3\mathbf{e}_2 = \mathbf{e}_1;\\
\mathbf{e}_3\mathbf{e}_1 &= -\mathbf{e}_1\mathbf{e}_3 = \mathbf{e}_2;\\
\mathbf{e}_1^2 = \mathbf{e}_2^2 &= \mathbf{e}_3^2 = -1.
\end{aligned}$$

Notice that $i\mathbf{e}_k = \mathbf{e}_k i$. In addition, for $q \in \mathbb{H}(\mathbb{C})$, it will be useful to consider the auxiliary notation:

$$q = \operatorname{Sc}(q) + \operatorname{Vec}(q) \tag{2}$$

where

$$\operatorname{Sc}(q) = q_0,$$

whereas

$$\operatorname{Vec}(q) = q_1\mathbf{e}_1 + q_2\mathbf{e}_2 + q_3\mathbf{e}_3.$$

From the formulae above, the reader can verify that the biquaternionic multiplication is not commutative, thus we will denote the right hand-side multiplication of the biquaternion Q by the biquaternion S as follows:

$$M^{(S)}Q = Q \cdot S. \tag{3}$$

Furthermore, there exist a subset of elements $Q, S \neq 0, \in \mathbb{H}(\mathbb{C})$ such that:

$$Q \cdot S = 0. \tag{4}$$

The elements belonging to this subset of biquaternions are usually known as *Zero-Divisors*.

In addition, we will need to introduce the quaternionic Moisil-Theodoresco partial differential operator:

$$D = \mathbf{e}_1\partial_1 + \mathbf{e}_2\partial_2 + \mathbf{e}_3\partial_3; \tag{5}$$

where $\partial_k = \frac{\partial}{\partial x_k}$, $k = 1, 2, 3$; and being x_1, x_2 and x_3 the classical Cartesian axis.

2.2 Elements of Pseudoanalytic Functions

Following the concepts posed in [1], briefly modified on behalf of our main objectives but without loosing generality, let us consider the pair of complex-valued F and G, such that:

$$\mathrm{Im}\left(\overline{F}G\right) \neq 0, \tag{6}$$

where $\overline{F}$ denotes the complex conjugation of F. This is $\overline{F} = \mathrm{Re}\, F - i\,\mathrm{Im}\, F$, and just as explained before, i is the standard imaginary unit: $i^2 = -1$. A pair of functions (F, G) satisfying (6) is called a *Bers generating pair*, and any complex-valued function W can be expressed by means of the linear combinations of these functions:

$$W = \phi F + \psi G.$$

Here ϕ and ψ are purely real valued-functions. Employing this idea, L. Bers introduced in [1] the definition of the (F, G)-derivative of a complex function W:

$$\partial_{(F,G)}W = (\partial_z\phi)F + (\partial_z\psi)G, \tag{7}$$

where $\partial_z = \partial_x - i\partial_x$, and $z = x + iy$, being x and y the classical Cartesian coordinates system, in the plane. Yet, the derivative (7) will only exist iff the following condition is hold:

$$(\partial_{\overline{z}}\phi)F + (\partial_{\overline{z}}\psi)G = 0. \tag{8}$$

Here $\partial_{\overline{z}} = \partial_x + i\partial_y$, whereas $\overline{z} = x - iy$. Notice the partial differential operators ∂_z and $\partial_{\overline{z}}$ are usually introduced with the factor $\frac{1}{2}$, for they are indeed the Cauchy-Riemann operators. Nevertheless, it will become somehow more convenient to work without this factor in the following pages. Shall we introduce the notations:

$$\begin{aligned} A_{(F,G)} &= \frac{\overline{F}\partial_z G - \overline{G}\partial_z F}{F\overline{G} - G\overline{F}}, \quad a_{(F,G)} = \frac{\overline{G}\partial_{\overline{z}} F - \overline{F}\partial_{\overline{z}} G}{F\overline{G} - G\overline{F}}, \\ B_{(F,G)} &= \frac{F\partial_z G - G\partial_z F}{F\overline{G} - G\overline{F}}, \quad b_{(F,G)} = \frac{F\partial_{\overline{z}} G - G\partial_{\overline{z}} F}{F\overline{G} - G\overline{F}}; \end{aligned} \tag{9}$$

the expression (7) will turn into

$$\partial_{(F,G)} W = A_{(F,G)} W + B_{(F,G)} \overline{W}, \tag{10}$$

as well the condition (8) can be written as:

$$\partial_{\overline{z}} W - a_{(F,G)} W - b_{(F,G)} \overline{W} = 0. \tag{11}$$

The last equation is known as the *Vekua equation*, since it was deeply analyzed by I. Vekua in [7], playing a central role for the present work, and any complex-valued function W satisfying (11) will be called (F, G)-*pseudoanalytic*, or simply *pseudoanalytic*, if not risk of confusion is taken. Beside, the set of functions (9) will be referred as the *characteristic coefficients* of the generating pair (F, G).

Remark 1 Let p be a non-vanishing real-valued function within a bounded domain Ω. The pair of functions:

$$F = p, \quad G = \frac{i}{p}, \tag{12}$$

will satisfy the condition (6), therefore they constitute a generating pair. Moreover, their characteristic coefficients (9) satisfy the relations:

$$\begin{aligned} A_{(F,G)} &= a_{(F,G)} = 0, \\ B_{(F,G)} &= \frac{\partial_z p}{p}, \quad b_{(F,G)} = \frac{\partial_{\overline{z}} p}{p}. \end{aligned} \tag{13}$$

Thus the corresponding Vekua equation (11) will have the form:

$$\partial_{\overline{z}} W - \frac{\partial_{\overline{z}} p}{p} \overline{W} = 0. \tag{14}$$

Hereafter we will assume all generating pairs posses the form (12).

Definition 1 Let (F_0, G_0) and (F_1, G_1) be two generating pairs of the form (12). If their characteristic coefficients satisfy the relation:

$$B_{(F_1,G_1)} = -b_{(F_0,G_0)}, \tag{15}$$

we say that (F_1, G_1) is a successor of (F_0, G_0).

Definition 2 Let us consider the set of complex-valued functions:

$$\{\ldots, (F_{-2}, G_{-2}), (F_{-1}, G_{-1}), (F_0, G_0), (F_1, G_1), (F_2, G_2), \ldots\};$$

and let us assume that every (F_k, G_k) is a successor of (F_{k-1}, G_{k-1}). This set will be called a *generating sequence*. If $(F, G) = (F_0, G_0)$, we say that (F, G) is embedded into the generating sequence. Furthermore, if there exist a constant c such that:

$$(F_{k+c}, G_{k+c}) = (F_k, G_k),$$

the generating sequence will be called *periodic*, with period c.

Remark 2 Let us consider a generating pair of the form (12), and let us assume that the function p depends only upon the spatial variable x: $p = p(x)$. Then the generating pair (F, G) is embedded into a periodic generating sequence, with period $c = 1$. In other words, this implies that:

$$\ldots(F_{-2}, G_{-2}) = (F_{-1}, G_{-1}) = (F_0, G_0) = (F_1, G_1) = (F_2, G_2)\ldots$$

Definition 3 The *adjoin* pair (F^*, G^*) of a generating pair (F, G) of the form (12) is defined as:

$$F^* = -iF, \quad G^* = -iG. \tag{16}$$

The concept of an (F, G)-*integral* corresponding to an (F, G)- pseudoanalytic function W, was also introduced by Bers in [1]. We refer the reader to the cited bibliography if complete explanations are needed about the existence and properties of the integral operators composing the (F, G)-integral, since all functions W within the integral expressions in this work will be, by definition [1], (F, G)-integrable.

Definition 4 The (F, G)-*integral* of a complex-valued function W, when it exist [1], is defined as follows:

$$\int_\Gamma W d_{(F,G)} z = F \cdot \mathrm{Re} \int_\Gamma W \cdot G^* dz + G \cdot \mathrm{Im} \int_\Gamma W \cdot F^* dz, \tag{17}$$

where Γ denotes a rectifiable curve going from a fixed point z_0 till z, and the functions F^*, G^* are the element of the adjoin pair corresponding to (F, G), defined

in (16). Particularly, let $\partial_{(F,G)}W$ be the (F,G)-derivative of a pseudoanalytic function W. Then we will have that:

$$\int \partial_{(F,G)}W d_{(F,G)}z = W - \lambda F - \psi G. \tag{18}$$

But since, as proven in [1], the following properties hold:

$$\partial_{(F,G)}F = \partial_{(F,G)}G = 0,$$

the integral expression introduced in (18) can be considered the *antiderivative* in the sense of Bers of the function $\partial_{(F,G)}W$.

2.2.1 Formal Powers

The formal power $Z^{(0)}(a_0, z_0; z)$, with formal exponent 0, complex coefficient $a_0 = \operatorname{Re} a + i \operatorname{Im} a$, depending upon $z = x + iy$, being $x,\ y \in \mathbb{R}$, and with center at the fixed point $z_0 = x_0 + iy_0$; is defined according to the expression:

$$Z^{(0)}(a_0, z_0; z) = \lambda F + \mu G,$$

where λ and μ are constants that fulfill the condition:

$$\lambda F(z_0) + \mu G(z_0) = a_0.$$

The formal powers with higher coefficients are determined by the recursive expressions:

$$Z^{(n+1)}(a_{n+1}, z_0; z) = n \int_{\Gamma} Z^{(n)}(a_n, z_0; z) d_{(F,G)},$$

where the integral expression is indeed an (F,G)-antiderivative, defined in (17).

The formal powers posses the following properties:

- Each $Z^{(n)}(a_n, z_0; z)$ is (F,G)-pseudoanalytic. This is, they are solutions of the special Vekua equation (14).
- We have that:

$$Z^{(n)}(a_n, z_0; z) = \operatorname{Re} a_n \cdot Z^{(n)}(1, z_0; z) + \operatorname{Im} a_n \cdot Z^{(n)}(i, z_0; z). \tag{19}$$

- The following equality holds:

$$\lim_{z\to z_0} Z^{(n)}(a_n, z_0; z) = a_n(z - z_0)^n.$$

Remark 3 Any complex valued function W, solution of (14), accepts the representation:

$$W = \sum_{n=0}^{\infty} Z^{(n)}(a_n, z_0; z). \tag{20}$$

This implies that the last expression is indeed an analytic representation for the general solution of the special Vekua equation (14):

$$\partial_{\overline{z}} W - \frac{\partial_{\overline{z}} p}{p}\overline{W} = 0.$$

Following [1], the right side of the expression (20) will be called *Taylor series in formal powers*.

3 Solutions of the Dirac Equation with Electric Potential

Let us consider the Dirac equation with electric potential:

$$\left[\gamma_0\partial_t - \sum_{k=1}^{3}\gamma_k\partial_k + im + \gamma_0 u(x_1)\right]\Phi(t, \mathbf{x}) = 0. \tag{21}$$

Here where m is the mass of the particle, $u(x_1)$ denotes the electric potential, $\partial_t = \frac{\partial}{\partial t}$, being t the time variable, $\mathbf{x} = (x_1, x_2, x_3)$, and $\gamma_k, k = 0, 1, 2, 3;$ are the Pauli-Dirac matrices:

$$\gamma_0 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix}, \quad \gamma_1 = \begin{pmatrix} 0 & 0 & 0 & -1 \\ 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix},$$

$$\gamma_2 = \begin{pmatrix} 0 & 0 & 0 & i \\ 0 & 0 & -i & 0 \\ 0 & -i & 0 & 0 \\ i & 0 & 0 & 0 \end{pmatrix}, \quad \gamma_3 = \begin{pmatrix} 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \end{pmatrix}.$$

We shall now consider the harmonic representation for the solution $\Phi(t,\mathbf{x}) = e^{i\omega t}\varphi(\mathbf{x})$, where ω is the energy of the particle. Then the Eq. (21) becomes:

$$\left[i\gamma_0\omega - \sum_{k=1}^{3}\gamma_k\partial_k + im + \gamma_0 u(x_1)\right]\varphi(\mathbf{x}) = 0. \tag{22}$$

According to [3], let us consider the pair of linear matrix operators $\mathbf{A}$ and $\mathbf{A}^{-1}$:

$$\mathbf{A} = \begin{pmatrix} 0 & -1 & 1 & 0 \\ i & 0 & 0 & -i \\ -1 & 0 & 0 & -1 \\ 0 & i & i & 0 \end{pmatrix},\ \mathbf{A}^{-1} = \begin{pmatrix} 0 & -i & -1 & 0 \\ -1 & 0 & 0 & -i \\ 1 & 0 & 0 & -i \\ 0 & i & -1 & 0 \end{pmatrix}. \tag{23}$$

Following [4], we shall apply the Pauli-Dirac matrices γ_k, $k = 0, 1, 2, 3$, $\mathbf{A}$ and $\mathbf{A}^{-1}$, to the differential operator of (21) as follows:

$$-\mathbf{A}\gamma_1\gamma_2\gamma_3\left[\gamma_0\partial_t - \sum_{k=1}^{3}\gamma_k\partial_k + im + \gamma_0 u(x_1)\right]\mathbf{A}^{-1},$$

in order to obtain a biquaternionic Dirac equation:

$$\left(D - M^{g(x_1)\mathbf{e}_1 + m\mathbf{e}_2}\right)f(\mathbf{x}) = 0, \tag{24}$$

where $f(x)$ is a full biquaternionic function of the form (1), the complex function $g(x_1)$ denotes:

$$g(x_1) = iu(x_1) + i\omega,$$

and

$$f(x) = \mathbf{A}\varphi(x).$$

3.1 Special Solutions of the Biquaternionic Dirac Equation

Employing a variation of the technique presented first in [4], and in [6], let us assume that the biquaternionic function $f(\mathbf{x})$, solution of (24), possesses the form:

$$f = \alpha \cdot Q, \tag{25}$$

where α is a purely scalar function and Q is a biquaternionic function. Thus, expanding the Eq. (24) we will obtain:

$$\alpha DQ + D\alpha \cdot Q - \alpha Qg\mathbf{e}_1 - \alpha Qm\mathbf{e}_2 = 0.$$

In order to obtain exact solutions for this equation, let us assume that $\alpha \neq 0$, and that:

$$DQ - Qg\mathbf{e}_1 = 0. \tag{26}$$

In consequence, we will have:

$$D\alpha \cdot Q - \alpha Qm\mathbf{e}_2 = 0. \tag{27}$$

Let us suppose now that:

$$Q = q_0 + q_2\mathbf{e}_2, \tag{28}$$

where q_0 and q_2 are complex-valued functions. In addition, Q shall not be a zero divisor, *ergo* there must exist a biquaternion Q^{-1} such that $Q^{-1} \cdot Q = 1$. Then (27) can be rewritten as:

$$D\alpha - \alpha m\mathbf{e}_2 = 0,$$

from which we immediately obtain:

$$\alpha = Be^{mx_2}, \tag{29}$$

where B is an arbitrary complex constant. On the other hand, when analyzing term by term the Eq. (26), we will notice that:

$$\partial_2 q_0 = \partial_2 q_2 = 0,$$

thus the remaining terms are:

$$\partial_1 q_0 + \mathbf{e}_2\partial_3 q_0 + \mathbf{e}_2\partial_1 q_2 - \partial_3 q_2 - gq_0 + \mathbf{e}_2 g q_2 = 0.$$

Introducing the auxiliary notations:

$$\partial_{z_{\mathbb{H}}} = \partial_1 + \mathbf{e}_2\partial_3, \quad p_{\mathbb{H}} = \mathrm{e}^{\int g dx_1},$$

the Eq. (26) can be rewritten into a biquaternionic Vekua equation of the form:

$$\partial_{\bar{z}_{\mathbb{H}}} Q - \frac{\partial_{\bar{z}_{\mathbb{H}}} p_{\mathbb{H}}}{p_{\mathbb{H}}} \mathbf{C}_{\mathbb{H}}[Q] = 0, \tag{30}$$

where $\mathbf{C}_{\mathbb{H}}[Q]$ indicates the quaternionic conjugation of Q. This is:

$$C_{\mathbb{H}}[Q] = C_{\mathbb{H}}[q_0 + q_2 e_2] = \mathrm{Sc}\, Q - \mathrm{Vec}\, Q = q_0 - q_2 e_2.$$

Equation (30) is a biquiaternionic Vekua equation for which, adapting the elements of the pseudoanalytic function theory, we can construct the set of solutions in terms of formal powers. As a matter of fact, (30) is a corrected equation for that presented in [6], and constitutes a rediscovering of a biquaternionic Vekua equation related with a two-dimensional quaternionic Dirac equation presented in [2], where a scalar potential was considered. More precisely, considering $z_0 = 0$, we will obtain:

$$Z^{(0)}_{\mathbb{H}}(1, 0, z_{\mathbb{H}}) = e^{\int g dx_1}, \quad \text{and} \quad Z^{(0)}_{\mathbb{H}}(\mathbf{e}_2, 0, z_{\mathbb{H}}) = \mathbf{e}_2 e^{-\int g dx_1},$$

where $z_{\mathbb{H}} = x_1 + \mathbf{e}_2 x_3$. Thus, adapting the (F, G)-antiderivative introduced in (17), we can approach numerically or analytically the formal powers with higher coefficients:

$$Z^{(1)}_{\mathbb{H}}(1, 0, z_{\mathbb{H}}), Z^{(2)}_{\mathbb{H}}(1, 0, z_{\mathbb{H}}), \ldots, Z^{(1)}_{\mathbb{H}}(\mathbf{e}_2, 0, z_{\mathbb{H}}), Z^{(2)}_{\mathbb{H}}(\mathbf{e}_2, 0, z_{\mathbb{H}}), \ldots$$

3.2 Solutions for the Dirac Equation

The solutions of the biquaternionic Dirac equation (24) are constructed by combining (29) with the solutions of (26), according to (25):

$$f = B \cdot e^{m x_2} Z^n_{\mathbb{H}}(a_n, 0, z_{\mathbb{H}}),$$

whereas the solutions for the Dirac equation (22), remarking that $f = f_0 + f_2 \mathbf{e}_2$, are approached according to the expression [4]:

$$\varphi = \mathbf{A}^{-1}[f] = \begin{pmatrix} -\tilde{f}_2 \\ -\tilde{f}_0 \\ \tilde{f}_0 \\ -\tilde{f}_2 \end{pmatrix}, \tag{31}$$

where $\tilde{f}_k$, $k = 0, 1$; denotes the projection

$$f_k(x_1, x_2, x_3) \to \tilde{f}_k(x_1, x_2, -x_3).$$

4 Conclusions

The Dirac equation has posed many interesting paths since its very discovering in the early 20th Century, and it became even more relevant when several applied disciplines of Science related its study to specific topics, e.g., the Nuclear Medicine [5]. Therefor, researching new classes of solutions when a variety of potentials are applied, represents a useful task in order to study the particles behavior.

Specifically, a wide class of solutions will be available when considering the close relation between the biquaternionic Dirac equation, fully equivalent to the classical, and a special class of biquaternionic Vekua equation. It is interesting to remark that the original theory of pseudoanalytic functions is directly connected with several two-dimensional partial differential elliptic equations (see [2]), whereas the study of the Dirac equation, a very important example of partial differential hyperbolic equations, can be achieved by writing down the Vekua equation in terms of a biquaternionic variable, without introducing important changes in the postulates of pseudoanalytic function theory [1].

Thus, the relation between the Dirac equation and the biquaternionic Vekua equation could well reach a variety of interesting questions when analyzing the physical meaning of precise elements. For example, the results presented in these pages could immediately start the discussion about the physical interpretation of the formal powers with coefficient 1 and those with coefficient $\mathbf{e}_2$.

Based upon the last pages, we consider the study of the Dirac equation through the hypercomplex Vekua equation a very useful tool because its actual and potential contributions to Theoretical Physics.

Acknowledgments Marco Pedro Ramirez Tachiquin would like to acknowledge the support of Mantenimiento a Instalaciones Conservacion de Equipo e Inmuebles S.A. de C.V., Mexico.

References

1. Bers L (1953) Theory of pseudoanalytic functions. New York University, IMM
2. Kravchenko VV (2009) Applied pseudoanalytic function theory. Frontiers in mathematics. ISBN 978-3-0346-0003-3
3. Kravchenko V (2003) Applied quaternionic analysis. Heldermann, Berlin
4. Kravchenko V, Ramirez T. MP (2000) New exact solutions for the massive Dirac equation with electric or scalar potential. Math Methods Appl Sci 23:769–776
5. McParland BJ (2010) Nuclear medicine radiation dosimetry: advanced theoretical principles. Springer, London
6. Ramirez MPT, Martinez-Garza-Garcia V (2014) New solutions for the massive Dirac equation with electric potential, employing biquaternionic and pseudoanalytic functions. Proceedings of the world congress on engineering and computer science 2014, WCECS 2014. Lecture notes in engineering and computer science, 22–24 Oct 2014, San Francisco, pp 256–260
7. Vekua IN (1962) Generalized analytic functions, international series of monographs on pure and applied mathematics. Pergamon Press, United Kingdom

Charatrization of Building Penetration Loss for GSM and UMTS Signals at 850 MHz and 1900 MHz Bands

Hisham Elgannas and Ivica Kostanic

Abstract Building penetration loss is one of the main points of interest in the outdoor-to-indoor propagation. This chapter describes measurements of the building penetration loss for mobile signals within 850 and 1900 MHz bands. Four buildings are studied aiming to provide first-order statistics of radio coverage inside buildings of signals transmitting from outdoor base stations. The effects of operating frequency, signal bandwidth, receive antenna height, and the presence of line of sight are investigated and numerically presented.

Keywords Building penetration loss · Floor height gain · GSM-1900 MHz · Office building · Outdoor-to-indoor propagation · Suburban environment · UMTS-1900 MHz

1 Introduction

Current wireless systems are evolving to meet the increasing demand for high speed data services and capacity. In the most recent surveys, it has been found that a high proportion of speech and data traffic originates from inside buildings [1]. Therefore, one of the main goals of the upcoming wireless networks like LTE and LTE-Advanced is to extend the indoor coverage of radio signals transmitting from outdoor base stations. In outdoor-to-indoor scenarios, the ability to provide high speed data services to users inside building is dependent on the available Signal to Interference and Noise Ratio (SINR) throughout the building. The outdoor-to-indoor channel is complex and its understanding is critical in designing of a wireless communication system. Therefore, for accurate design of a modern

H. Elgannas (✉) · I. Kostanic
Department of Electrical and Computer Engineering, Florida Institute of Technology, Melbourne, FL 32901, USA
e-mail: helgannas2010@my.fit.edu

I. Kostanic
e-mail: kostanic@fit.edu

H.K. Kim et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-94-017-7236-5_10

wireless system, real detailed spatial data measurements that describe the received signal power levels inside buildings are of fundamental importance. In addition, a reliable characterization of the wireless channel parameters, which are statistical in nature, requires robust statistical analysis [2]. Several studies regarding measurements based outdoor-to-indoor building penetration loss characterization and modeling have been published [2–10]. The buildings examined in these publications range from small residential buildings to large office buildings. Building penetration loss at frequencies up to 8 GHz has been studied. However, the results and the observations presented in those studies are limited by the assumptions, the working conditions, and the environment that have been considered. The study presented in this chapter describes extensive detailed measurement campaign conducted inside and around office buildings. The goal of the campaign is to characterize mainly the building penetration loss and its effects on in-building coverage. A spatial description of radio coverage of mobile signals that are transmitted from six different GSM and UMTS base stations serving the examined buildings is presented in details. The building penetration loss data are presented graphically and coupled with site-specific information.

2 Experimental Setup

2.1 Environment Description

The measurements are carried out on the campus area of Florida Institute of Technology (FIT) in Melbourne, FL. The measurements are dedicated to test the outdoor-to-indoor building penetration loss for GSM and UMTS signals within 850 and 1900 MHz frequency bands. The campus area of FIT is considered as a part of suburban area having buildings of different construction materials and layouts. Macro-cell base station antennas are well above the average heights of the building and allow wide area coverage. Buildings are typically low detached office and education buildings with two to seven floors. There are occasional open areas such as parks and playgrounds and the vegetation is modest. Locations of buildings where measurements were conducted are shown in Fig. 1. The figure also shows the relative position of the examined buildings with respect to GSM-850, GSM-1900, and UMTS-1900 MHz base stations.

2.2 Buildings Description

The examined buildings are Crawford tower (CRAW), Olin Engineering complex (O.ENG), Olin Life Sciences (O.LS), and Olin Physical Science Center (O.PSC). These buildings include specialized research and teaching laboratories, classrooms,

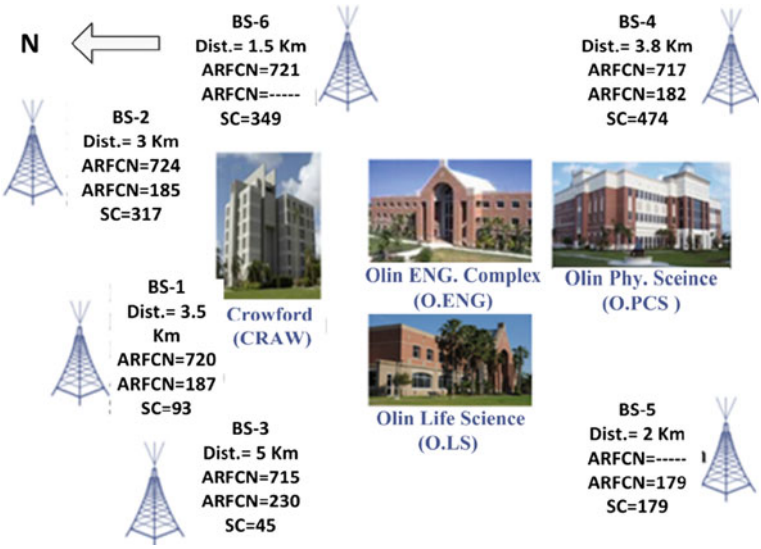


Fig. 1 Locations of examined buildings and base stations

and conference rooms. Offices on all floors of all buildings were equipped with typical office furniture. The teaching or research laboratories contained experimental equipment according to the specialized demand of each area of study. The floor height in all buildings has mean value of about 4 m. The external walls of all buildings are made of brick and concrete masonry except for Crawford building whose external walls are made of concrete masonry only. A brief description of the measured buildings is summarized in Table 1.

2.3 Measurement Procedures

The building penetration loss in this study is defined as the difference in mean received signal level between the measurement obtained inside and a reference signal level measured outside near the unobstructed walls of the measured building [3]. In order to provide reliable outdoor references for the respective buildings, over two hundred data points at different locations around each building are considered.

Table 1 Brief description of the 4 measured buildings

Building	Floor area (m^2)	Windows percentage	Const. date	No. of floors
CRAW	360	10 % of building sides	1961	7
O.PSC	2300	16 % of building sides	2005	4
O.ENG	2100	28 % of building sides	1999	3
O.LS	2000	22 % of Building sides	1999	2

For indoor measurements, the floor area is divided into small squares. The area of each square is 4 m^2.

With this setup, more than 8 thousand in-building locations are measured. The measurement are conducted using Agilent digital receivers. Two GSM receivers are used to measure more than twelve Broadcast Control Channels (BCCH) of GSM-850 and GSM-1900 MHz base stations. For UMTS-1900 MHz system, one receiver is used to measure the scrambling codes (SC) of six different UMTS-1900 MHz base stations. The receivers perform measurement of received signal levels for all frequency channels simultaneously. The analyzer displays the amplitude of the Absolute Radio Frequency Channel Number (ARFCNs) and the scrambling codes (SC) for all measurements with the corresponding power levels in dBm. Figure 2 shows Agilent digital receivers that are used to conduct the measurements. The data collection system setup (for both indoor and outdoor locations) is also shown in the same figure. The receiving antennas are mounted on the top of a wooden mast at height of 1.5 m above the examined floor. The collection system shown in Fig. 2 is wheeled around and inside the measured buildings on trolley.

In order to provide site-specific information about the radio coverage inside buildings, color coded plots are used. Figure 3 shows a sample scatter plot of received power level measurements on ground floor of Crawford building (left). Dots symbolize location of successive measurement points and the color of each dot indicates basic received signal level in dBm. An interpolated view of the same measurements is presented on the right side of Fig. 3. The interpolated view is created by using spatial

Fig. 2 The equipment used and the data collecting system in two different locations

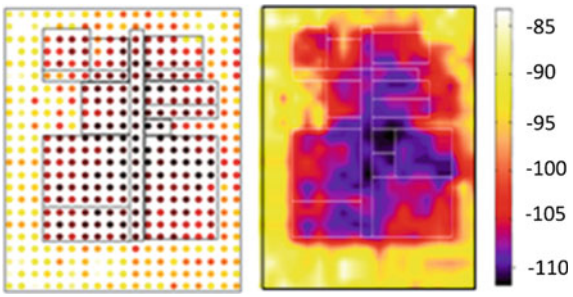


Fig. 3 Scatter plot and its interpolated view of a radio coverage inside building

interpolation. Inverse Distance Weighting (IDW) interpolation method is used to predict the expected average signal level in places where measurements are not possible.

3 Measurement Results and Data Analysis

There are various factors that may influence propagation into buildings. Operating frequency, signal bandwidth, building construction materials, transmission conditions, and floor height where the receiver is located are considered important parameters in characterizing the outdoor-to-indoor building penetration loss [4]. In what follows, the effects of the main factors that influence measured building penetration loss are discussed in detail.

3.1 Average Building Penetration Loss

Building penetration loss is usually calculated according to an adopted definition. Different methods of calculating building penetration loss are presented in the literature [5, 6]. In this study, the method of calculating the building penetration loss used in [6] is adopted. In [3, 6, 7], the building loss is calculated as the difference between the average received power at a reference point outside building (measured at base station side of the building) and the individual averages of received powers at all specified data points inside building. Various propagation models designed to predict the signal power levels at street levels treat the building penetration losses as additional losses. This makes the calculation of building loss with respect to a reference point located outside that building more practical. Table 2 summarizes the mean (Mean) and the standard deviation (STD) values of the building penetration loss obtained for all measured buildings.

The reported Mean and STD values of the building penetration loss represent the average Mean and the average STD over the four measured buildings. The most noticeable trend from the table is the trend of decreasing in standard deviation versus frequency for GSM signals.

Table 2 Average and standard deviation of building penetration losses for the different buildings for 850 and 1900 MHz bands at the ground floor

Building	GSM-850 MHz		GSM-1900 MHz		UMTS-1900 MHz	
	Mean	STD	Mean	STD	Mean	STD
CRAW	13	4.7	13.7	3.0	13.7	4.5
O.ENG	14.6	4.5	17.2	2.47	18.2	4.8
O.PHY	14	4.2	17	1.5	19	4.2
O.LS	9.5	4.16	11.1	2.3	12.4	4

3.2 The Impact of Operating Frequency

Often, different electrical properties of building construction materials and objects interacting with electromagnetic waves make building penetration losses frequency dependent [4]. Dependency of the building penetration loss on frequency are reported in [8, 9]. However, some recent studies show slight or no frequency dependency as in [10]. Figure 4 shows a comparison plot of the mean and the standard deviation of ten signals transmitted from five different base stations operating at both 850 and 1900 MHz bands.

As shown in Fig. 4, and for all the measured BCCH channels in 850 and 1900 MHz bands, the operating frequency has slight impact on the building penetration loss in three buildings. However, no frequency dependency was noticed in Crawford building. According to Table 2, the building penetration loss increased by an average of 2.8 dB when the operating frequency increases from 850 to 1900 MHz. These results are in good agreement with the conclusion drawn in [9]. These results may be attributed to the difference in building construction materials between Crawford building and the other three buildings.

Figure 5 shows a sample plot of the difference in building loss between the 850 and 1900 MHz channels for the ground floor of O.PHY building which exhibits the highest penetration loss values. This exemplary plot represents an example of building loss distribution in O.ENG, O.LS, and O.PHY buildings.

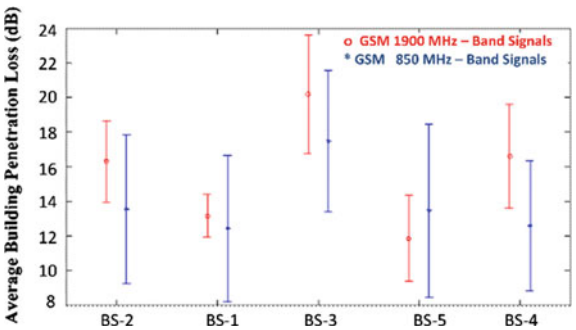


Fig. 4 Computed building penetration mean and standard deviation for GSM-850 MHz versus GSM-1900 MHz channels using measured data from all buildings

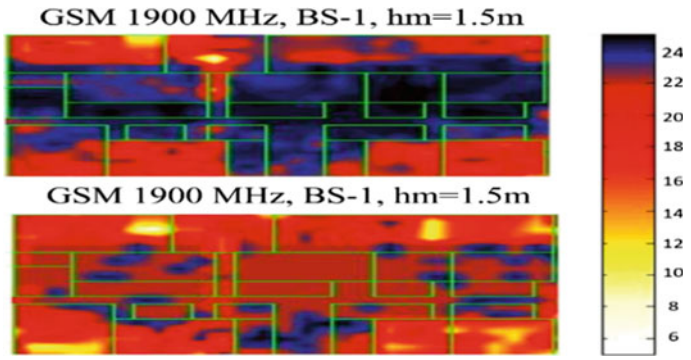


Fig. 5 Penetration loss in O.PHY building of GSM channels transmitting from same base station. One channel is in 1900 MHz band (*up*) and one is in 850 MHz band (*bottom*)

It is obvious from Fig. 5 that 1900 MHz channels exhibits higher building loss than 850 MHz channels especially at the center region of the building. In general, the frequency dependency can be attributed to the difference in dominating propagation mechanisms at different frequencies as the electrical properties of the building construction materials vary with frequency. For example, signals at higher frequencies reach the receiver mainly due to the wave guiding and the diffraction mechanisms. Therefore, buildings that show frequency dependency are more transparent to the lower frequency bands than to the higher ones.

The average building penetration loss values are useful for modeling purposes but they are not always sufficient especially when it comes to interpreting individual results. For example, Fig. 6 shows a comparison plot of the building loss of four different channels in Crawford building where no frequency dependency was noticed. On the top of Fig. 6, comparison plot shows the building loss map of GSM 850 and GSM 1900 MHz that were transmitting from the same base station (BS1). The mean building penetration loss for both channels is the same but the STD and distribution are different. Similar comparisons are made between another 850 and 1900 MHz channels transmitting from (BS2) on the bottom of Fig. 6. Overall, and for Crawford building, 1900 MHz-transmitters provide smoother coverage when compared to 850 MHz-transmitters with lower values of standard deviation as reported in Table 2.

3.3 The Impact of Channel Bandwidth

Figure 7 shows a comparison plot of the Mean and the STD of the building penetration loss obtained in four measured buildings. Eight signals transmitted from four different base stations of GSM-1900 and UMTS-1900 MHz systems are considered. It is clear from Fig. 7 that the UMTS signals experience nearly the same Mean and STD values of loss within each individual building. This consistency is

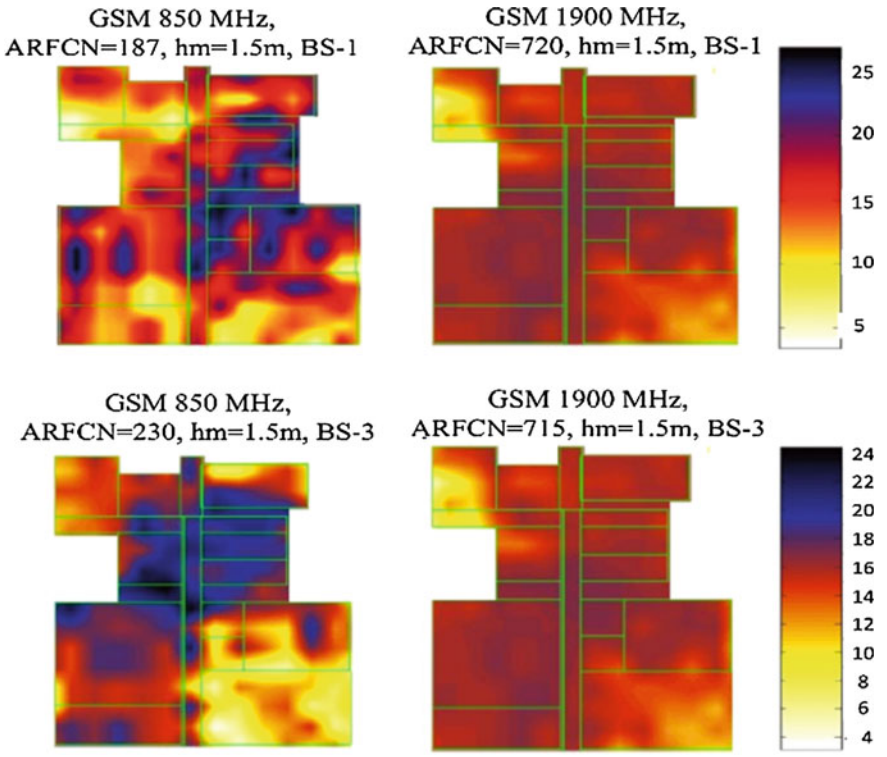


Fig. 6 Measured penetration loss of four channels transmitting from different base stations. GSM-850 MHz signals (*left*) and their counterpart in GSM-1900 MHz band (*right*)

due to the fact that all measured UMTS signals operate at the same carrier frequency but they are separated by different scrambling codes. On contrary, GSM signals transmitted from different base stations encounter different Mean and STD values of building penetration loss in each individual building.

It is noticeable also in Fig. 7 that the building penetration loss of all measured buildings is slightly higher for wideband signals than narrowband signals for three different base stations. However, three buildings (CRAW, O.PSC, and O.LS) exhibit higher penetration loss for narrowband signals that are transmitted from BS1 than the wideband signals received from the same base station.

The results also show that both wideband and narrowband signals that are transmitting from same base station may encounter nearly the same average building penetration loss and show no significant signal band dependency. For example, the received GSM and UMTS signals at the ground floor of Crawford building that were transmitting from BS2 and BS3 undergo a mean building loss of about 14 dB (See Fig. 7). Figure 8 shows a comparison plot of the building penetration loss of four different signals (narrowband and wideband) that were received from BS2 and BS3.

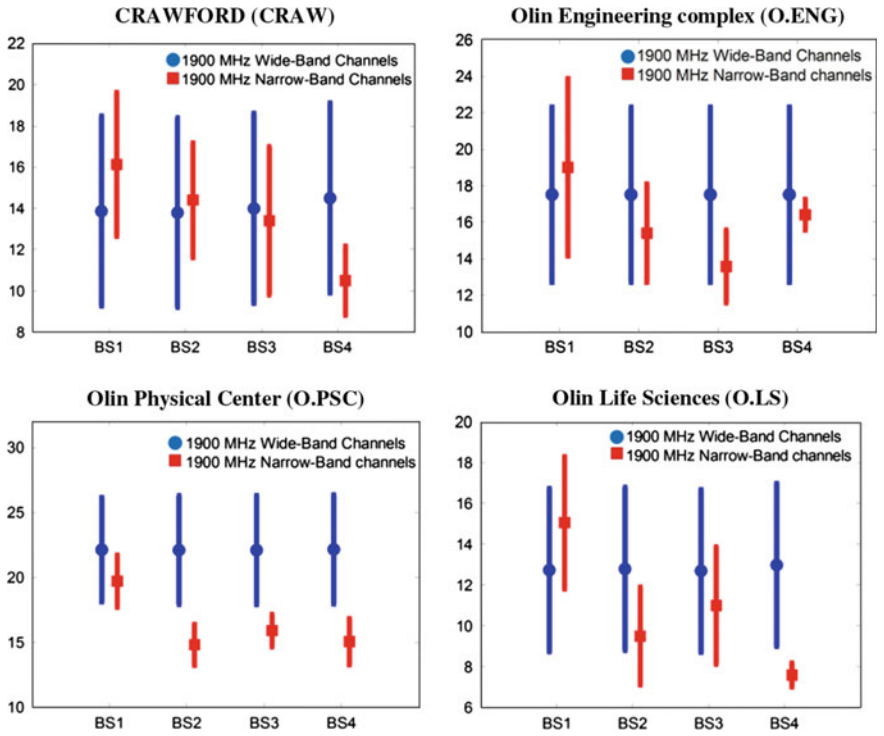


Fig. 7 Computed building penetration mean and standard deviation for UMTS-1900 MHz versus GSM-1900 MHz channels in the four measured buildings

3.4 The Impact of the Presence of Line-of-Sight

In outdoor-to-indoor propagation, the LOS condition is not related to presence or absence of direct path between transmit antenna and receive antenna [2]. Instead, the LOS path is defined as the direct path between the outdoor antenna and the immediate exterior environment to the indoor antenna.

For all measurements, it was observed that the LOS condition affects the standard deviation of the received power levels inside and in the vicinity of the measured buildings. The presence of LOS path has also obvious influence on penetration distance of measured signals (penetration depth). These results confirm obtained observations reported in previous work for GSM signals [2, 3, 7]. Outdoor measurements in locations placed in both illuminated and non-illuminated facades of the four measured buildings show relatively high values of standard deviations. For GSM-850 MHz signals, the standard deviation of the outdoor measurements takes values between 4 and 7 dB, while, it ranges between 2.3 and 5.3 dB for GSM-1900 MHz signals. Additionally, outdoor measurements of UMTS signals in both the illuminated and non-illuminated facades of the measured buildings show

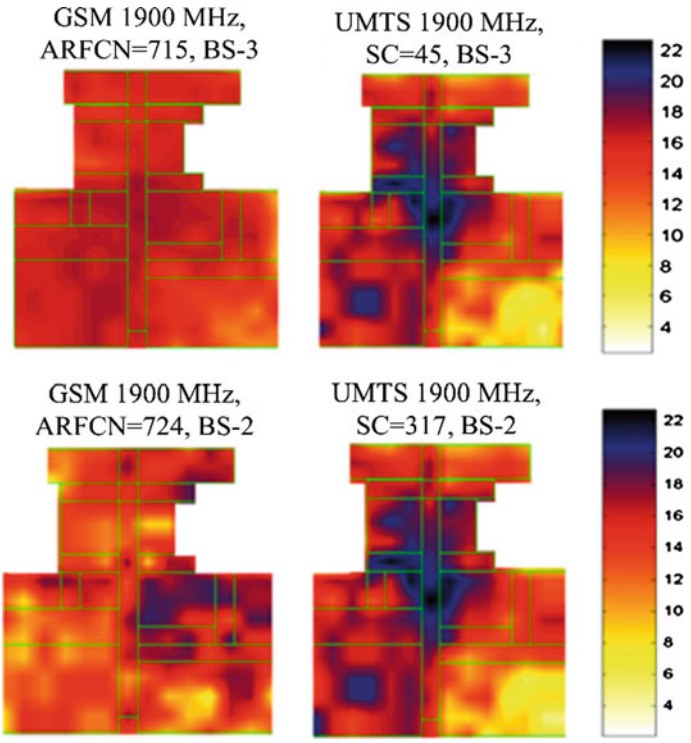


Fig. 8 Measured penetration loss of four channels. Two channels are in GSM-1900 MHz band (*left*) and two in UMTS-1900 MHz band (*right*)

relatively consistent higher values of standard deviations. It takes values between 4 and 4.8 dB for UMTS-1900 MHz signals.

On the other hand, indoor measurements confirmed an expected decrease of building penetration loss within LOS-areas inside building when compared to NLOS-areas. Figure 9 shows a comparison plot of received signal power of two GSM-850 MHz transmitters in two different floors that are subjected to different transmission conditions.

In Fig. 9 (top), the 3rd floor of O.PHY building is subject to clear LOS path from the west side of the building while the plot in the bottom shows the coverage of a different channel that is obstructed by neighboring buildings. At the same antenna height and with presence of windows and sufficient clearance to the operating base station, an average difference of 7 dB in penetration losses between LOS-areas and NLOS-areas inside buildings are observed.

In the macro-cellular environment, determination of the accurate transmission condition is not always possible. One way to investigate the transmission condition is to first plot an interpolation view of the radio coverage of each measured signal in the respective building floor. Then relate the plot to the relative position of each BS to the respective building floor.

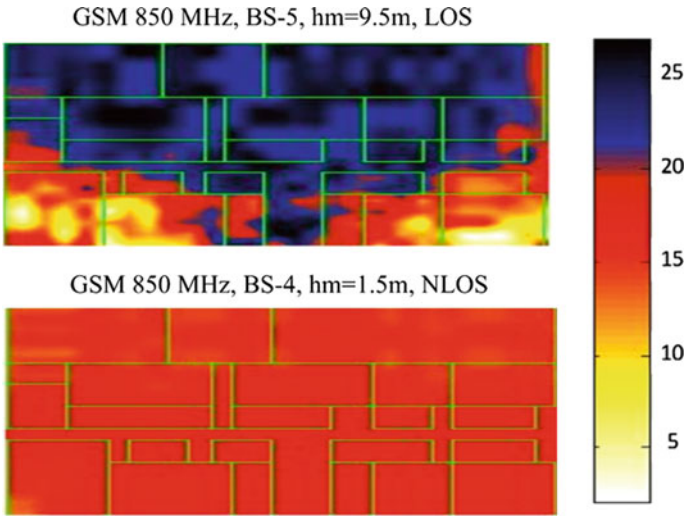


Fig. 9 Radio coverage of two different GSM 850 MHz transmitters in two different transmission condition with O.PHY building (LOS in *top*) and (NLOS in *bottom*)

Figure 10 shows a comparison plot of building penetration loss of the ground floor of O.PSC building. Four GSM 1900 MHz channels that are transmitting from different base stations are considered in the plot. It is assumed here that the ground floor of O.PSC building has four different transmission conditions with the measured GSM base stations.

On the top left corner of Fig. 10, the transmission path between BS1 and the respective ground floor is partially blocked by O.ENG building. As a result of the NLOS condition, the building penetration losses take higher values in comparison with the other transmission conditions. In the other plots of Fig. 10, the effect of the direct transmission path (LOS) is obvious on in-building areas, especially at the base station side of the respective building. In such cases, direct transmission paths encounter higher diffraction than other paths which in turn causes significant fluctuations in the received signal power. This explains the relatively higher standard deviation values observed at LOS-areas inside buildings.

3.5 The Impact of Floor Height Gain

The floor height gain is well known model used to describe the decrease in building penetration values at higher building levels with respect to its values at ground floor [8]. Regression analysis is used to evaluate the floor height gain effect on building penetration loss in this study. Figure 11 shows a sample plot of line that best fits in-building measurements of the received power levels of a list of line-of-sight-UMTS signals. Measurements in seven floors of Crawford building

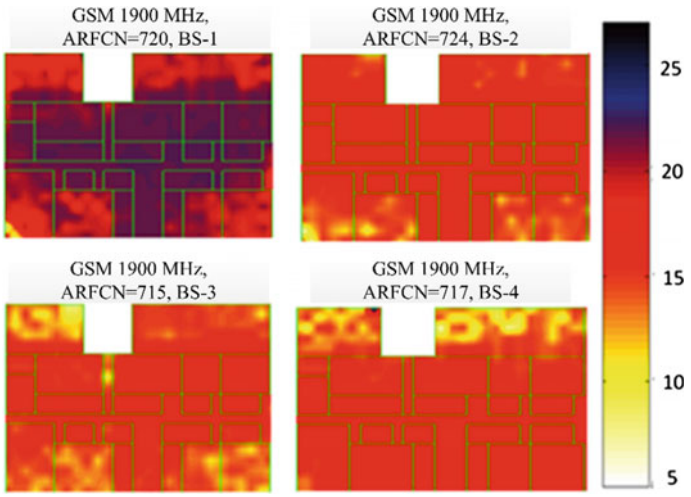


Fig. 10 Building penetration loss plots of four different GSM 1900 MHz transmitters in different transmission conditions for the ground floor of O.PHY building

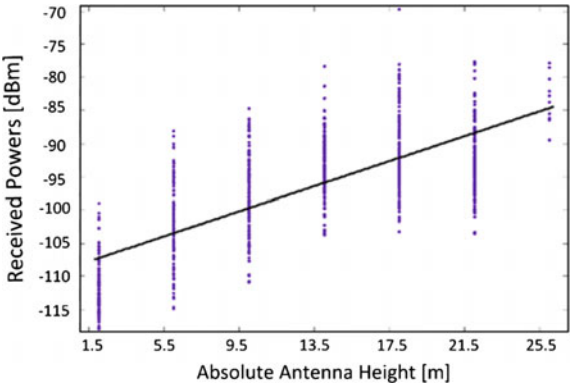


Fig. 11 Average received power levels of UMTS channels versus receiver heights in high-rise building together with best fit line

are used. To eliminate other factors that may affect the received power levels of different signals, only in-building locations at the base station side are considered.

It may be seen clearly that as the receive antenna height increases, a steadily increase in the received power levels becomes higher as well. Therefore, the building penetration loss values decreases with an increase of the received antenna's height. For all measurements in Crawford building, the floor height gain takes values between 0.662 and 0.96 dB/m for GSM-850 MHz signals, while, this gain takes values between 0.49 and 0.60 dB/m for GSM-1900 MHz signals. For UMTS-1900 MHz signals, the floor height gain is consistent and has a mean value of 0.95 dB/m. On average, the floor gain is 0.58 dB/m. These results are in good agreement with the findings reported in [2, 8]. However, the effect of floor height is

not reliable and as pronounced in the other measured buildings (O.ENG, O.LS, and O.PSC). In these buildings, the difference in received signal power levels between ground floor and the highest floor is clear but not consistent for GSM signals. In contrast, the floor height gain for UMTS signals in those three buildings has a value of 0.8 dB/m.

The other observed influence of the floor height is the increase in penetration depth of radio signal as it propagates through building. When mobile station moves to higher floors, the outdoor-to-indoor transmission condition may also change. Then, the chance to receive direct transmission paths (LOS path) from different transmitters increases.

The impact of floor height on the received power level, and therefore on the building penetration depth, is depicted in Fig. 12. In the figure, three sample plots represent the radio coverage of GSM signal in three different floor heights of Crawford building (CRAW).

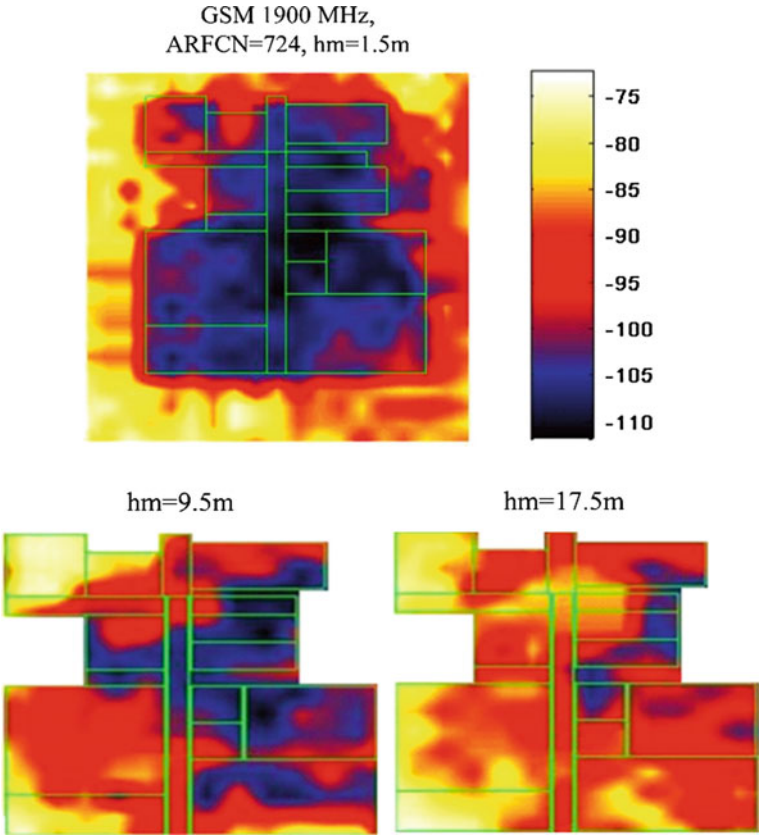


Fig. 12 Increasing in the penetration depth of the signal propagation for one base station within three different floors inside Crawford building

The measurements of the received power levels of signal transmitted from BS1 within the 1st floor are shown on the top of Fig. 12. The radio coverage map of the 3rd floor is shown in the bottom left side whereas the coverage map of the 5th floor is shown in the bottom right side. In higher floors, the attenuation of the penetration distance becomes more linear which in turn makes the process of modeling the building penetration loss more practical in such floor areas. A possible explanation for this observation is that at higher floors, the effect of other attenuation factors is limited to the minimum with the presence of LOS transmission condition.

4 Summary

In this chapter, the building penetration loss in macro-cellular environment is investigated for mobile signals of GSM-850, GSM-1900 and UMTS-1900 MHz systems. The building penetration loss was evaluated for 4 office buildings in suburban environment.

On average, the building loss was 16 dB for both narrowband and wideband signals with higher standard deviation of 4.5 dB for wideband signals. The results showed that the building loss increases by 2.8 dB as the operating frequency increases from 850 to 1900 MHz for buildings with same construction materials. However, one of the measured buildings, which is constructed with different materials showed no frequency dependency. For some base stations, dependency of the loss on the signal bandwidth was noticeable. The received UMTS signals exhibited higher losses when compared to GSM signals coming from the same base stations. Additionally, UMTS signals encounter consistent building losses in each measured building with mean standard deviation of 4.5 dB. For GSM signals, an average floor height gain of 0.58 dB/m confirms previous results reported in the literature. This gain was higher for UMTS signals and it had a value of 0.95 dB/m.

References

1. De la Roche G, Glazunov AA, Allen B (2013) LTE-advanced and next generation wireless networks: channel modelling and propagation. Wiley, Chicago
2. Berg J (1999) 4.6 building penetration: digital mobile radio toward future generation systems. Brussels, Belgium
3. Elgannas H, Kostanic I (2014) Outdoor-to-indoor propagation characteristics of 850 MHz and 1900 MHz bands in macro-cellular environments. In: Proceedings of the world congress on engineering and computer science, WCECS 2014. Lecture notes in engineering and computer science, 22–24 Oct 2014, San Francisco, USA, pp 685–690
4. Stavrou S, Saunders SR (2003) Factors influencing outdoor to indoor radio wave propagation. In: 12th international conference on antennas and propagation, 2003 (ICAP 2003), vol 2. (Conf. Publ. No. 491), pp. 581–585
5. De Toledo AF, Turkmani AMD, Parsons J (1998) Estimating coverage of radio transmission into and within buildings at 900, 1800, and 2300 MHz. IEEE Pers Commun 5:40–47

6. LaSorte NJ, Burette Y, Refai HH (2010) Experimental characterization of electromagnetic propagation of a hospital from 55–1950MHz. In Proceedings of IEEE Asia-Pacific Symposium on Electromagnetic Compatibility, pp 826-829
7. Elgannas H, Kostanic I (2014) Outdoor-to-indoor propagation characteristics of 1900 MHz signals in macro-cellular environments for GSM and UMTS systems. Univers J Electr Electron Eng 3:24–30
8. Turkmani AMD, Parsons JD, Lewis DG (1988) Measurement of building penetration loss on radio signals at 441, 900 and 1400MHz. J Inst Electr Radio Eng 58:S169–S174
9. Rose DM, Kurner T (2012) Outdoor-to-indoor propagation 2014; accurate measuring and modelling of indoor environments at 900 and 1800 MHz. In: 6th European conference on antennas and propagation (EUCAP) 2012, pp 1440–1444
10. Okamoto H, Kitao K, Ichitsubo S (2009) Outdoor-to-indoor propagation loss prediction in 800-MHz to 8-GHz band for an urban area. IEEE Trans Veh Technol 58:1059–1067

Experimental Validation of Lafortune-Lacrous Indoor Propagation Model at 1900 MHz Band

Ali Bendallah and Ivica Kostanic

Abstract With an ever increasing demand for cellular voice and data services, many of the cellular providers are finding deployment of the in-building infrastructure inevitable. The in-building deployments allow for further re-use of the allocated spectrum and hence, they provide one of the most effective ways for increasing the overall system capacity. However, such deployments face some formidable challenges. The performance of the in-building systems depends heavily on the characteristics of the indoor radio channel. Excessive path loss within the building can cause lack of coverage. The indoor mobile radio channel is not easily modeled. The channel varies significantly from one building to the other. Furthermore, the channel depends heavily on factors which include building structure, layout of rooms, and the type of construction materials used.

Keywords Cellular networks · Indoor propagation · Lafortune model · Measurements scenario · Model accuracy · Path loss estimation · Personal communication network · Propagation modeling · Signal strength · Spectrum clearing

1 Introduction

In the past couple of decades personal communication systems have experienced a tremendous growth. As a result of grow, the user expectation have risen. Nowadays one expects to enjoy seamless connectivity. The signal needs to be brought to the user to provide such connectivity in the frequency reuse environment. Many service

A. Bendallah (✉) · I. Kostanic
Electrical and Computer Engineering Department, Florida Institute of Technology, Melbourne, FL 32901, USA
e-mail: abendallah2009@my.fit.edu

I. Kostanic
e-mail: kostanic@fit.edu

H.K. Kim et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-94-017-7236-5_11

providers are installing large number of indoor cellular systems. To effectively plan such installation, Cellular network operators have to have a thorough understanding of the radio signal propagation inside building across all frequency bands that are primarily used in Personal Communication Network (PCN).

Despite some efforts in development of propagation models for the in-building propagation, there is still a significant gap. A good example of the successful implementation of a PCS is indoor wireless communication. Indoor wireless communication covers a wide variety of situations ranging from communication with individual walking in office facility or residential buildings, hospitals, factories, etc. to fixed stations sending signals that control some machines or devices in a service or factory environment.

Due to reflection and refraction, the indoor environment is prone to interference. Radio signals strength varies due to structures layout of the building and obstacles inside it, the signal either deteriorate or reach the receiver by more than one path, resulting in a phenomenon is known as multiple fading. Multipath cause deep fading and pulse spreading of the signal.

1.1 Basic Radio Propagation

Path loss is defined as the reduction in signal power as it propagates through space. Path loss is a major component in the analysis and design of the link budget of a telecommunication system. When deployed, a multi-hop relaying system traverses two radio links.

The most basic model of radio wave propagation widely known is the free space propagation model. In this model, radio waves radiate from a point source of radio energy, traveling in all directions in straight line. The radio waves fill the entire spherical volume of space with signal power decreases by 20 dB per decade increase in distance. Real world radio propagation rarely follows this simple model.

Three main phenomena which are reflection, diffraction and scattering are associated with transmission of signals inside buildings. All three of the phenomena cause radio signal strength deterioration and signal fading, as well as additional signal propagation losses. These reflections come from different directions or can be in a form of diffracted or scattered signal components. Reflection, diffraction, and scattering are the three basic propagation mechanisms that affect the in-building propagation as well. Due to their importance, those mechanisms will be briefly explained as follows.

1.1.1 Reflection

Reflection occurs when a propagation signal impinges on an object which of large dimensions compared to the wavelength of the propagated signal. As shown in

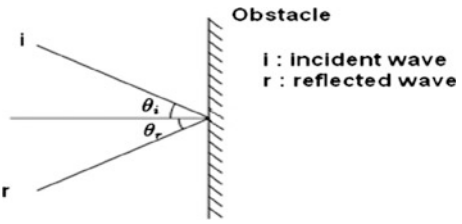


Fig. 1 Reflection of a propagated signal

Fig. 1, the incident wave is reflected with an angle θ_r, which is dependent on the type of obstacles, causes the reflection.

1.1.2 Diffraction

The diffraction occurs when the radio path between the transmitter and receiver is obstructed by an obstacle.

1.1.3 Scattering

Scattering occurs when the dimensions of the objects in the wave path are small compared to the wavelength, and when the number of the obstacles per unit volume is large [1].

In practice, the scattering in mobile communications can be caused because of existence of for example, walls and obstacles.

1.2 Indoor RF Propagation

The indoor propagation environment is characterized by several multi-path propagation. The line of sight may not exist and the characteristics of this environment can change drastically over a very short distance or time due to the amount of multipath and movement of the people, equipment and doors [2].

In the indoor environment, there are two types of RF propagation obstacles: Hard and soft partitions. The Hard partitions are defined as a part of the physical/structural components of a building. On the other hand, the soft partitions are the obstacles formed by the office furniture, fixed or movable or portable structures that do not extend to a building ceiling. The radio signals effectively penetrate both kinds of obstacles or partitions in ways that are extremely hard to predict. In free space case, the propagation model is used to predict the strength of the received signal when the transmitter and receiver aren't obstructed and have a

clear path which is called Line of Sight (LoS) which is not the case with the indoor environment where there is no direct line of sight between the transmitter and receiver.

To develop a propagation model for indoor environment, a comprehensive understanding of radio propagation inside buildings is essentially required. In effect, it is the nature of the radio channel that makes wireless network design and deployment challenging when compared to their wired counterparts. Radio propagation is heavily depending on the site and hence can vary significantly depending on building structure and frequency of operations.

To get an appropriate model that can be used to predict the propagation inside a building, an extensive measurements are required to come up with a model because there is no generally accepted model that can be implemented since each building has its own structure and layout. Also, the propagation model is an essential tool required for modeling the in-building Cell deployment. This is one of the most significant trends in cellular mobile communication network.

Propagation prediction is a major factor in communication network planning. If the modeling is too conservative, considerable high costs may be incurred, whereas too liberal version of modeling can result in an inaccurate prediction and unsatisfactory results [3].

For communication network planning, the modeling of the propagation behavior is for the purpose of predicting the received signal strength at the end of the link. In addition to signal strength, there is other channel impairments that can degrade link performance such as delay spread which occur due to multipath fading.

1.3 Lafortune-Lacrous Indoor Propagation Model

This model was developed using measurements from a two story building by Jean-Francois Lafortune. The experiments have been conducted with a continuous wave emission at 917 MHz by using a synthesized signal source was used as transmitter. The transmit power was of 10 dBm for most measurements and 19 dBm for the experiments between two floors and more. Spectrum analyzers were used as receivers.

The transmit antenna was kept stationary for the measurement at a height of 1.7 or 2.5 m. For the signal strength measurements, the receiving antenna was moving, at height of about 1.7 m. over an area of approximately 1 m^2 and the average attenuation was noted, as visually estimated from the screen of the spectrum analyzer; this manual procedure was judged to carry an uncertainty of 1 dB [4].

For calculation of the correction factors, the model provides a set of cases and a set of corresponding factors empirically derived equations. The summary of cases and associated equations is provided in Table 1. The table of cases and associated equations summarizes only the cases that pertain to strictly in-building and single floor propagation.

Table 1 Cases for calculation of correction factors in Lafortune–Lecours model at 900 MHz

Case	Description	L_{OB}, G_{RM} in dB	
C.1	n walls between TX and RX	$L_{OB} = 3.7 - 1.5n - 10.7\log(d) + \begin{cases} 0, & \text{if } d' < 4\text{m} \\ -7.8 + 15.3\log(d'), & \text{if } d' \geq 4\text{m} \end{cases}$ where d' is the distance to the closest wall. *Note* Corner uses n = 1, thin wall uses n = ½ and a thick wall may use n = 2. $G_{RM} = 0$	
C.2	Door between antennas	y: distance behind the door, θ: angle between TX-RX line and door wall • Door open ($\theta > 30°$), no other wall If $y \leq 2$, then $L_{OB} = 0,\ G_{RM} = 2$ If $2 < y < 10$, then $L_{OB} = 0,\ G_{RM} = 0$ If $y > 10$, then $L_{OB} = -1,\ G_{RM} = 0$ • Door closed If $y \leq 2$, then $L_{OB} = -2,\ G_{RM} = 0$ If $y > 2$ use case C.1 with n = 1 • Doors and walls (x_1 door) Use case C.1 with $n = x_1 + x_2$	
C.3	Windows between antenna	1 window, $\theta > 45°,\ L_{OB} = 0,\ G_{RM} = 0$ 1 window, $\theta < 45°$, use C.1 with n = 1 1 window, x walls, use C.1 with $n = x$	2 windows, use C.1 with n = 1 2 windows, x walls, use C.1 with $n = x$ 3 windows, x walls, use C.1 with $n = x + 1$
C.4	Furniture between antennas	Non-metallic furniture L_{OB} = Case C.1 − 1, $G_{RM} = 0$ High, metallic furniture with wall L_{OB} = Case C.1 − 2, $G_{RM} = 0$ High, metallic furniture without wall $L_{OB} = -4,\ G_{RM} = 0$	
D.1	Emission in main corridor	Main corridor, no transversal doors: $L_{OB} = 0,\ G_{RM} = 0.2 + 1.8\log(d)$	
D.2	End of corridor	Last 8 m: $L_{OB} = 0,\ G_{RM} = 1.6 + 3.9\log(d)$	
E	Lateral corridor, opening in main corridor	E.1: No door at the junction: $L_{OB} = -5.6 - 12\log(h+1),\ G_{RM} = 0$ Where h is 'geometric diffraction parameter'	E.2: door at the junction Door open: same as E.1 Door closed: $L_{OB} = -7.6 - 11.5\log(h+1),\ G_{RM} = 0$
F	Room adjacent to corridor (d > 30 m)	$L_{OB} = -7.6 - 11.5\log(h+1),\ G_{RM} = 0$	
G	Emission in the same room	$L_{OB} = 0,\ G_{RM} = 0.2 + 1.8\log(d)$	

Note All distances are in meters

From Table 1, one readily observes that the model identifies many different propagation cases. The definition of the cases is somewhat qualitative and as a result, one may expect to see differences in different application of the model. Also, one may note, that all the equations for correction factors are derived using path loss measurements obtained at 900 MHz. This raises a question of their validity when used in different frequency bands.

2 Experimental Setup

2.1 Equipment Description

The data collection system consists of a transmitter, transmit antenna, receiver, receive antenna, Global Positioning System (GPS) antenna and a laptop with installed measurement software called Wireless Measurement System (WMS) from Grayson wireless. The software is used to measure the strength of the received signal and map its value on the spectrum tracker screen along with the corresponding frequency. The transmitter is mounted in the middle of the building, or inside one of the rooms in the building, depending on the scenario implemented. The receiver is placed in a cart as shown in Fig. 2 and moved along the hallway in the third floor of a multi-story building. Spectrum clearing of the area was performed before measurements were conducted, to verify that the frequency used for the path loss measurements is free from any source of interference.

Table 2 shows the technical specifications of the Tx antenna Andrew ASPP-2933E.

Fig. 2 Illustration of the transmitter (*left*) and the receiver (*right*)

Table 2 Technical specifications of Tx antenna

Frequency range	1850–1990 MHz
Gain	5.1 dBi
VSWR	<1.5:1
Nominal impedance	50 Ω
Polarization	Vertical
Maximum power	250 W
Dimension/diameter	635 mm × 38 mm/25 in. × 1.5 in.
Connector	7/16 DIN female
Connector place	Bottom
Pattern	Omni directional
Hardware material	Galvanized steel

2.2 *Environment Description*

The measurements obtained in this study are collected in OLIN Engineering building at Florida Institute of Technology (FIT) Melbourne, Florida, USA. The third floor plan of the building is presented in Fig. 3.

There are 29 offices, 11 labs, one elevator, two emergency exits, stairs and hallways. The offices and labs have metal stud walls while the walls of the emergency exits and elevator are made out of concrete white brick. The height of ceiling is 9′ 2″ covered with acoustic ceiling pressed fiber tiles. The size of surveyed area is 20,770 square feet.

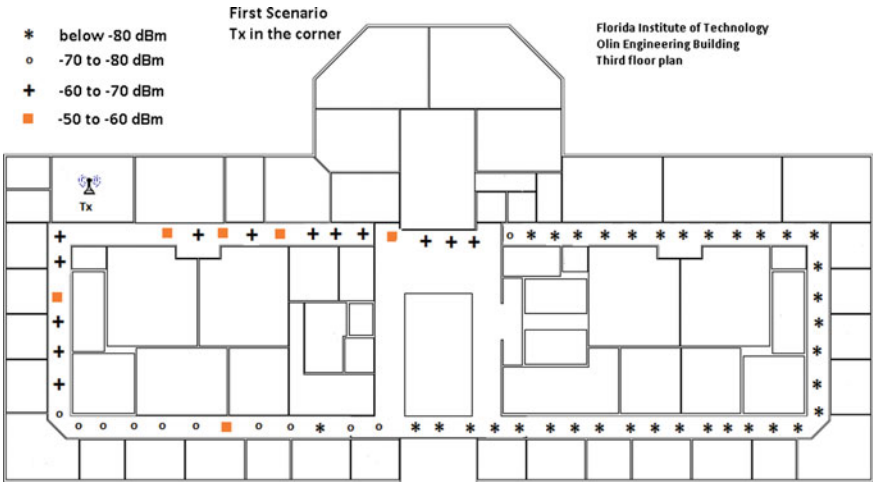


Fig. 3 First measurement scenario

2.3 Measurement Procedure

The 1900 MHz band path loss measurements used for the model verification are collected in a three-story building at the campus of Florida Tech. A Continuous Wave (CW) transmitter is used. The transmitter operating frequency is 1925 MHz with transmit power of 6 dBm. For the measurements reported in this chapter, both the transmitter and the receiver are located on the same floor of the building.

2.4 Measurements Scenarios

In-building RF propagation prediction is the first step in modeling the wireless channel behavior inside the building. There are many factors that impact the wireless propagation such as the structure of the building, the materials inside it, and the associated factors which lead to more attenuation and cause loss of signal strength. Spectrum clearing conducted for 12 h. The average reading value was −130.8 dBm with standard deviation of 2.66 dB.

During the measurements, the transmitter is kept stationary, while the receiver is moving to measure the signal strength in different points. The measurements are taken when the receiver is standing still. The measurements points are 8 feet apart in the hallways through the building.

The measured signal strength is averaged at each point to eliminate the fast fading from the motion of the environment that is surrounding both the transmitter and the receiver. The averaging time is 3 min. Therefore, the measurements represent the average path loss.

The investigation process considers two experimental scenarios. The scenarios are described as follows.

2.5 Measurements Scenario 1

In this measurements scenario, the transmitter is mounted inside a room located in the corner of the building as depicted in Fig. 3. In this scenario, all measurements points have no line of sight. Before being measured, the signal has to pass through at least one obstruction.

2.6 Measurements Scenario 2

In this scenario, the transmitter is placed in the middle of the building. Many measurements have a clear line of sight to the transmitter.

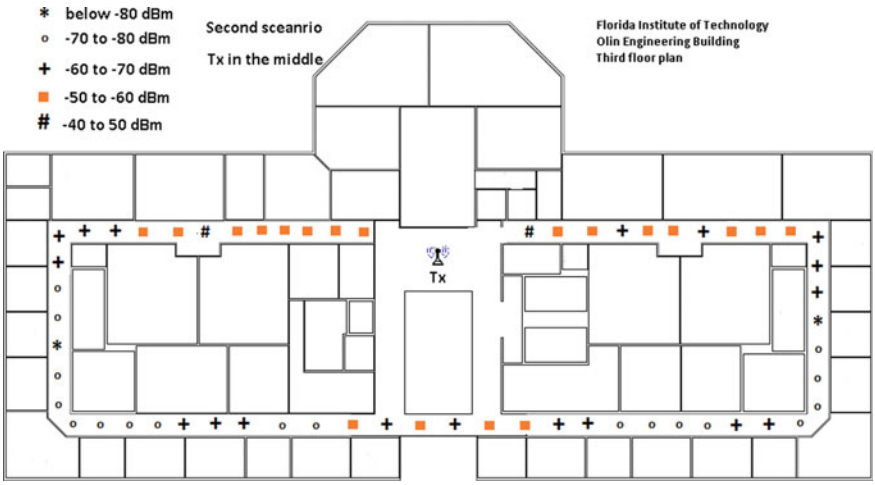


Fig. 4 3rd floor plan and measurements positions

Also, for the measurements points that do not have LOS, the signal is propagating through spectral reflections from the floor, ceiling and the building walls. The outline of the measurements in the second scenario is depicted in Fig. 4.

2.7 Measurements Procedure

The measurements are collected in 1900 MHz band inside OLIN Engineering building and Table 3 shows the parameters associated with the measurements scenarios.

Table 3 Parameters associated with the measurement scenarios

Parameter	Value
Operating frequency	1925 MHz
Transmit antenna height	1.7 m
Transmitting power	43 dBm
Transmit antenna gain	6 dBi
Cable and connector losses	0.7 dB
Receive antenna height	71.2 cm
Receive antenna gain including cable and connector losses	5 dBi
Noise figure	2 dB

3 Investigation of the Accuracy of Lafortune Model at 1900 MHz Band

Lafortune conducted measurements of signal strength at 900 MHz at two story building and came up with a model that computes the path loss inside buildings. The next step to be done is to investigate the validity of this model in Florida Tech. at OLIN Engineering building. To improve the model's performance, results from multiple buildings should be incorporated. The idea behind that is to conduct signal strength measurements in different floors; in the first stage; of OLIN building, and analyzing the data to find out if there is a possible improvement of the model proposed by Lafortune by adding correction parameters, for example.

4 Results and Analysis

The signal strength inside the building is measured. The signal strength is decreasing as the distance increases. The transmission path is obstructed by walls. As depicted in Figs. 5 and 6, the signal strength decreases as a function of the distance between the transmitter and receiver and the number of obstacles located in the path of propagation.

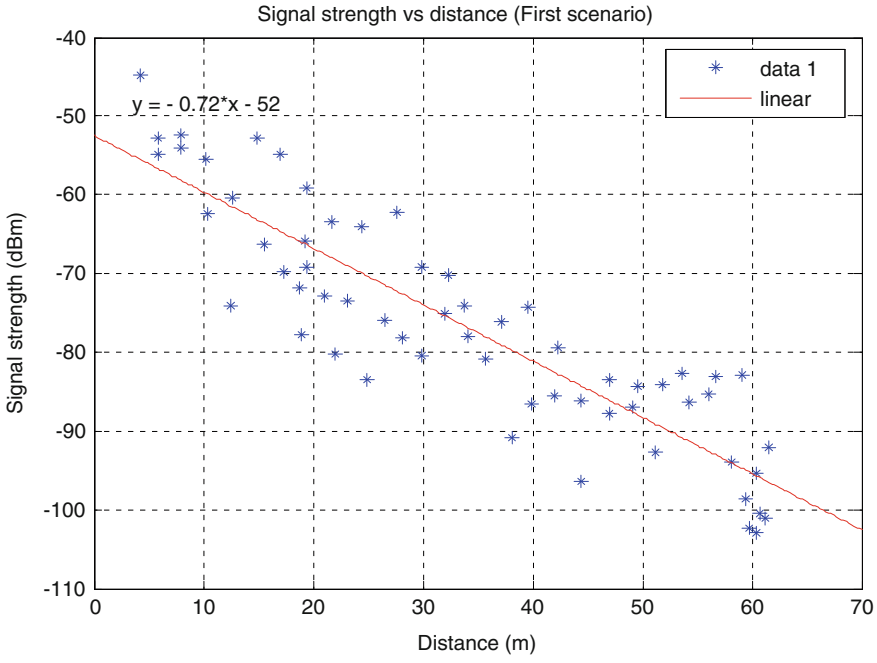


Fig. 5 Signal strength versus distance (scenario 1)

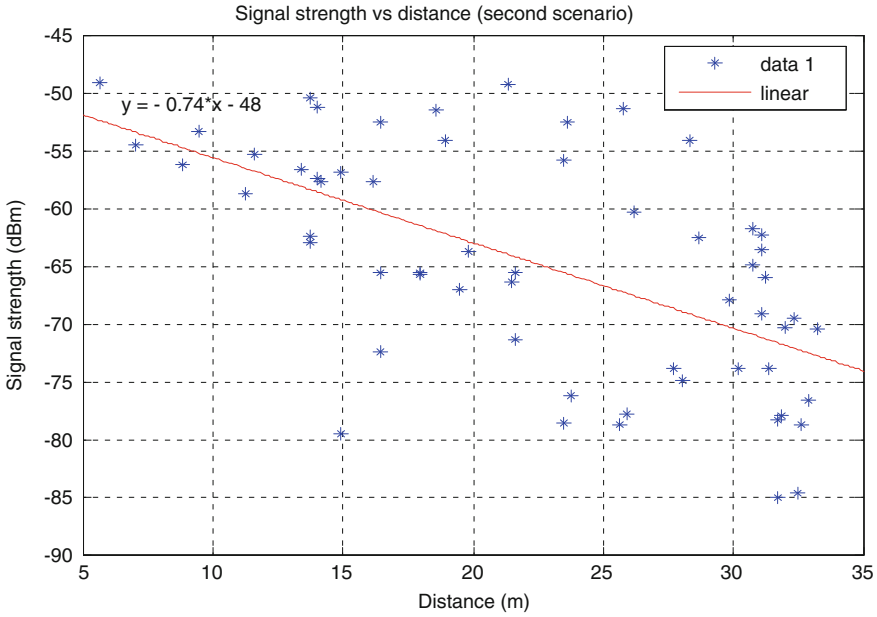


Fig. 6 Signal strength versus distance (scenario 2)

In some locations, the signal experiences high loss because of the number and kind of obstacles that are on the path of radiation. The spots with star sign represent where the severe loss of the signal that might occur due to presence of different kinds of walls other that in the rest of the building (Brick and concrete in this specific case).

The path loss caused by the existence of obstructions is expressed as in section C1 in Table 1. Due to refraction from the walls, the signal attenuation is smaller because of the gain caused by the refraction. The gain is expressed as in section G in Table 1. Different measurements points have almost the same distance experience comparable signal strength.

The path loss is computed according to Lafortune-Lecrous model and compared to the measured values of the path loss resulted from the measurements as seen in Figs. 7 and 8.

To investigate the validity of Lafortune-Lecrous model, the results of the measured values of path loss are compared and can be shown in Figs. 9 and 10.

The difference between the measured and predicted values of the path loss is depicted in two scenarios of measurements (different position of transmit antenna). The error resulted from the difference between the predicted and measured path loss is shown in a histogram form Figs. 11 and 12.

The difference between the values of the path loss predicted by Lafortune-Lecrous model and the measured values from the experiment is calculated. The standard deviation is 5.7 dB. There is some measurement show big

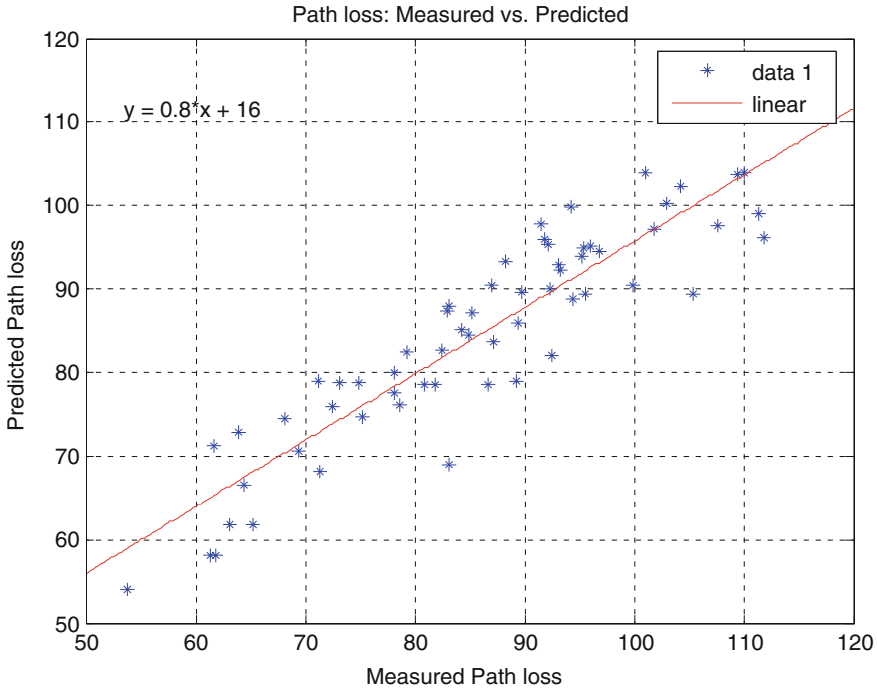


Fig. 7 Predicted versus measured path loss (scenario 1)

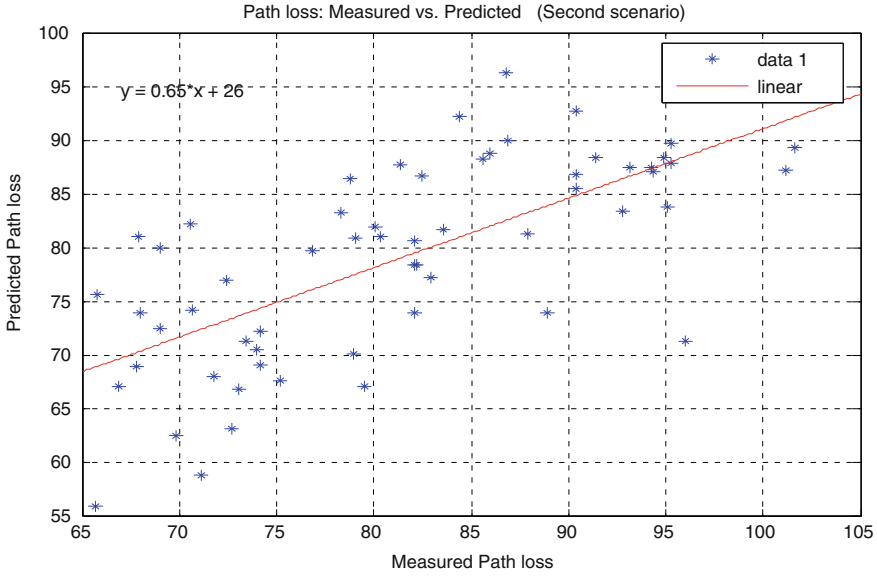


Fig. 8 Predicted versus measured path loss (scenario 2)

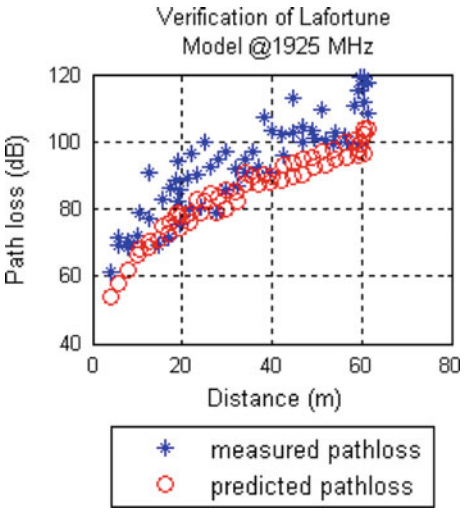


Fig. 9 Verification of Lafortune model (scenario 1)

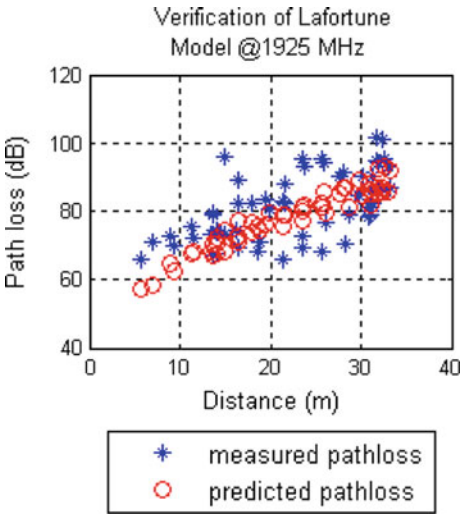


Fig. 10 Verification of Lafortune model (scenario 2)

difference between the predicted and measured path loss where the signal path goes through different types of obstructions as shown in Tables 4 and 5.

In this case, the measured values of the path loss are greater than path loss values predicted by Lafortune model. Table 4 depicts some path loss vales where the radiation path is obstructed by walls made of concrete bricks. By taking a look at Tables 4 and 5, it can be clearly seen the effect of the Brick walls. The difference between the two vales of the path loss is quite large.

In the first scenario, the average error resulted from the difference between the predicted and a measured value is 11.5 dB with 5.7 dB Standard deviation [5].

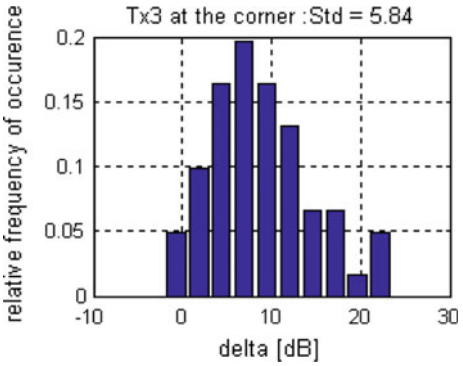


Fig. 11 Error histogram (scenario 1)

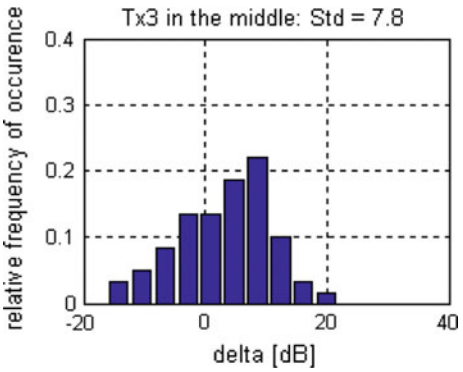


Fig. 12 Error histogram (scenario 2)

In the second scenario, the average difference between the measured and predicted is 3.03 dB while the standard deviation is 7.8. By comparing both scenarios, it has been found that standard deviation of the difference between the actual and predicted values in the first scenario is less than that of second scenario. Lafortune-Lecrous model shows a good approximation in the second scenario

Table 5 Predicted and measured values of path loss (scenario 2)

Rx (dBm)	No of walls	Path loss measured (dB)	Path loss predicted (dB)	Delta
−79.5	4	96	74.2	21.8
−99.8	9	116.3	97.1	19.2

Table 4 Predicted and measured values of path loss (scenario 1)

Rx (dBm)	No of walls	Path loss measured (dB)	Path loss predicted (dB)	Delta
−94.9	7	111.4	95.4	16
−99.8	9	116.3	97.1	19.2

(where the transmit antenna is mounted in the middle of the building) when compared to the results got from first scenario measurements.

In the second scenario, the error average of the difference between the actual and predicted values of the path loss is greater than that of first scenario. Since the transmitter location was in the middle of the building as depicted in Fig. 4, the amount of reflection is higher than that of first scenario, that is why the model shows almost 2 dB difference when the location of the transmitter had been changed.

Some values of Delta (error values: difference between measured and predicted value of the path loss) as depicted in Tables 4 and 5 show a large difference and needed to be interpreted again as Laforune-Lecrous model is incapable to show an acceptable prediction.

References

1. Rappaport T (1996) Wireless communications principles and practice. Prentice-Hall Inc., New Jersey
2. Papoulis A, Pillai U (2002) Probability, random variables and stochastic processes, 4th edn. McGraw-Hill, New York
3. Lafortune JF, Lecours M (1990) Measurement and modeling of propagation losses in a building at 900 MHz. IEEE Trans Veh Technol 39:101–108
4. Murch RD, Cheung KW, Fong MS, Sau JHM, Chuang JCI (1994) A new approach to indoor propagation prediction. In: 44th vehicular technology conference, 1994, vol 3. IEEE, 8–10 June 1994, pp 1737–1740
5. Bendallah A, Kostanicmes I (2014) Experimental validation of Lafortune-Lacrous indoor propagation model at 1900 MHz band. Lecture notes in engineering and computer science. In: Proceedings of the world congress on engineering and computer science 2014, WCECS 2014. San Francisco, USA, 22–24 Oct 2014, pp 708–712

Distributed Protection for the Enterprise

William R. Simpson

Abstract Entities in the enterprise are deployed with a standard configuration. Over time, patches, updates, new software versions, and mistakes or malicious activity all lead to deviations across the enterprise from this standard baseline. Malicious or unknown software on a system can cause harm or unexpected behavior. To mitigate these problems where possible, and help fix them in other cases, an enterprise plan for quality of protection is needed. This involves eliminating certain actions on machines that could harm the machine itself or the enterprise. The level of protection is dependent upon the type of enclave (an enclave is defined as a collection of entities with a common set of security and assurance mechanisms in place). Certain mitigations will be exercised based upon the cyber environment and enclave, and they may be exercised in different ways when communication is needed across enclaves of differing security and assurance. Mitigations include virus scanners and disabling of devices or interfaces. These mitigations also involve identifying and fixing issues that were not stopped. This requires a central visualization of the enterprise to quickly identify potential issues and a method of remotely taking action to either fix the affected system or freeze it until further action can be taken. This chapter discusses the current approach to centralized monitoring of communication as opposed to a more distributed approach. The latter relies on a well-formed security paradigm for the enterprise. The paper concludes with a proposal for a distributed inspection system that is currently being developed and tested.

Keywords Encryption · Security · Key management · Appliance · Packet inspection · Protection · Traffic inspection

W.R. Simpson (✉)
Institute for Defense Analyses, 4850 Mark Center Drive, Alexandria, VA 22311, USA
e-mail: rsimpson@ida.org

H.K. Kim et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-94-017-7236-5_12

1 Introduction

This paper is based in part on a paper published by WCECS [1]. Enterprise protection is based upon the device, the environment, and the enclave type. An enclave is a collection of entities and assurance mechanisms that uniformly employ the same security. Protection includes on-device software, in-line monitoring of communications and the particular security model that provides confidentiality, integrity, and availability. Many times this security model is compromised in trying to provide basic levels of protection. These compromises may include policy-based instructions to configure intrusion detection devices that provide the capability for selecting which attacks are being monitored. These policy selections can provide capabilities to select what responses will be taken for each detected intrusion.

2 Current Protection Approaches

Elements involved in implementing Quality of Protection are numerous and complicated. A wide range of appliances are used to provide functionality ranging from quality of service to the user or quality of protection of network resources and servers. These appliances are often placed in-line and some require access to content to provide their service. Figure 1 illustrates how these appliances are installed between the user and the application.

The number of appliances can be quite high [2–23]. Below is a partial list of functional types:

- Header-based scanner/logger:
  - Views only unencrypted portion of traffic
  - Synchronous or asynchronous operation
  - Scans for suspicious behavior, logs traffic
- Content-based scanner/logger:
  - Views decrypted content
  - Synchronous or asynchronous operation
  - Scans for suspicious behavior, logs traffic and/or content
- Header-based firewall:
  - Views only unencrypted portion of traffic
  - Synchronous operation
  - Scans for and blocks suspicious behavior

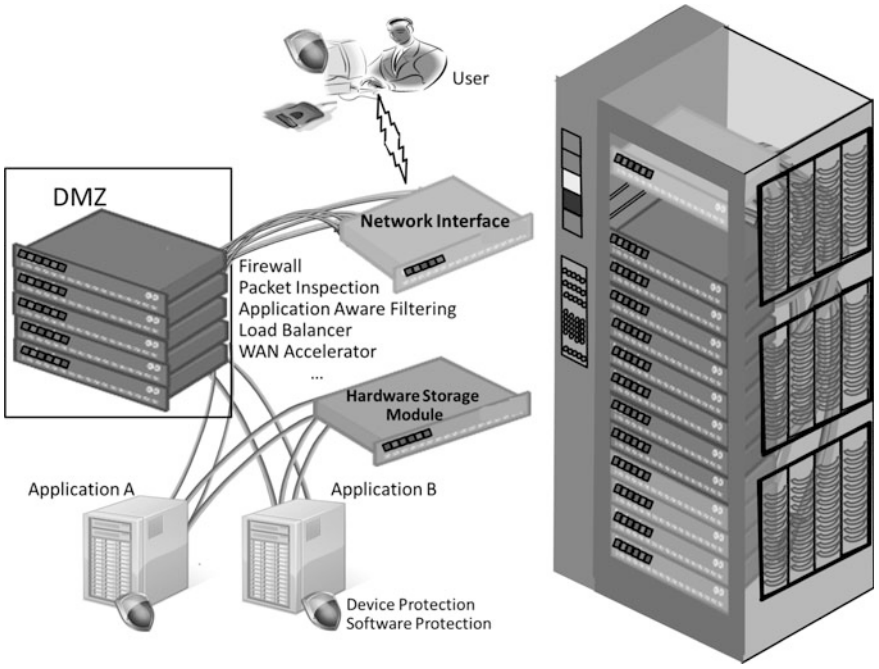


Fig. 1 End-point access

- Content-based firewall—block only:
  - Views decrypted content
  - Synchronous operation
  - Scans for suspicious behavior and blocks (terminates) connection
- Content-based firewall—modifies malicious content:
  - Views decrypted content
  - Synchronous operation
  - Scans for suspicious content, and blocks connection or removes suspicious content while preserving the connection
- Web accelerator:
  - Views decrypted content
  - Synchronous operation
  - Modifies content for performance
- WAN accelerator:
  - Views decrypted content
  - Multi-party system

  - Synchronous operation
  - Modifies content representation between parties.

- Load Balancers:
  - Distributes load among destination end points.
  - May decrypt content:

    May combine encrypted flows through an "encryption accelerator"
    May distribute content by request to different servers based on load
    These load balancers are considered active entities.

  - May not decrypt content:

    Using "sticky" balances individual requests to the same server.
    These load balancers are considered passive entities.

Each of the appliances above offers some functionality and increases the threat exposure. None of these are bullet-proof from a security standpoint and they do increase the threat surface and the vulnerability space. Use of any appliance must be balanced by the increased functionality and the increased vulnerability. The situation is further complicated by vendor offerings of load balancers with firewall capability, "smart" accelerators that scan content, and software only offerings that provide most of these functionalities in a modular fashion.

2.1 Current—Unencrypted Traffic

To understand the current paradigm, a review of what is done through a portal for unencrypted traffic is provided. HTTP traffic is unencrypted from browser to portal, and unencrypted from portal to web applications providing content.

1. Examples:
   a. www.amazon.com
   b. www.va.gov
   c. www.af.mil
2. Man-in-the-middle (MITM) model for appliances (e.g., firewalls, deep packet inspection, accelerators) that perform analysis of headers and content (IP, TCP, HTTP, HTML, XML, JavaScript, etc.)
3. Portal is the endpoint for browser requests.

The process is shown in Fig. 2.

Note that while the traffic is unencrypted, the requester may or may not be authenticated using a smart card with public key infrastructure credentials. The traffic is pulled in-line through a number of appliances to protect ports, and inspect content. We have included a web accelerator in this figure, because they have a number of characteristics in common with these other appliances. Other appliances,

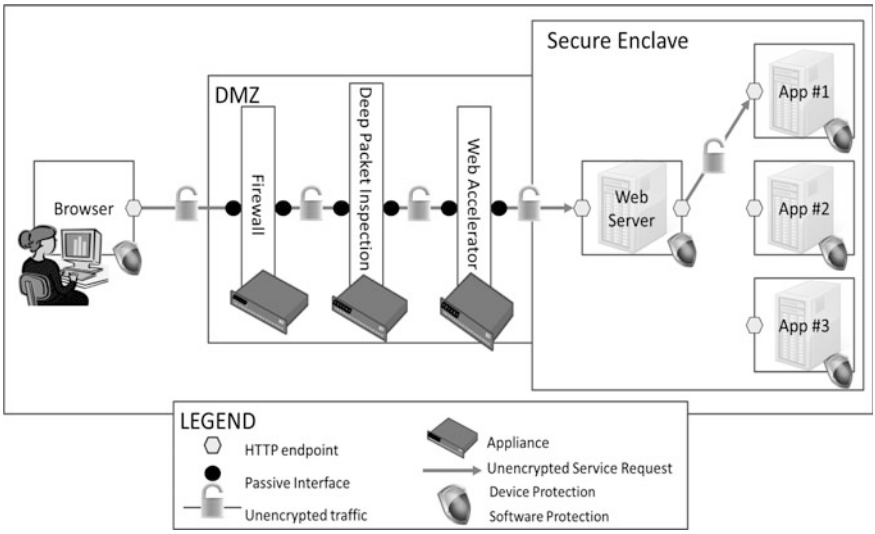


Fig. 2 Current paradigm for unencrypted traffic

including load balancers and WAN accelerators may also be included in this stack. Load balancers and some firewalls may be treated a passive entities in this treatment. WAN Accelerators are not covered in this paper.

2.2 Current—Encrypted Traffic

When traffic is encrypted, the same basic approach is adapted to handling traffic inspection. HTTPS traffic is successively decrypted and re-encrypted when needed. End-to-end HTTPS traffic is encrypted using Transport Layer Security (TLS) from browser to portal and using separate TLS sessions from portal to web applications

1. Example: a. https://www.mybank.com
2. MITM model for appliances
3. Some can function without decryption (e.g., firewall)
4. Some require decryption, using portal private key
5. Portal is the endpoint for browser requests.

The process is shown in Fig. 3.

In order to be able to decrypt the packages, the private key and initialization vectors (IVs) from the portal are provided to the in-line appliances. As such, they see the handshake and exchange and have access to the keying material sufficient to compute session and Message Authentication Code (MAC). A MAC is a cryptographic checksum on data that uses a session key to detect both accidental and intentional modifications of the data (for integrity). While sharing of private keys is

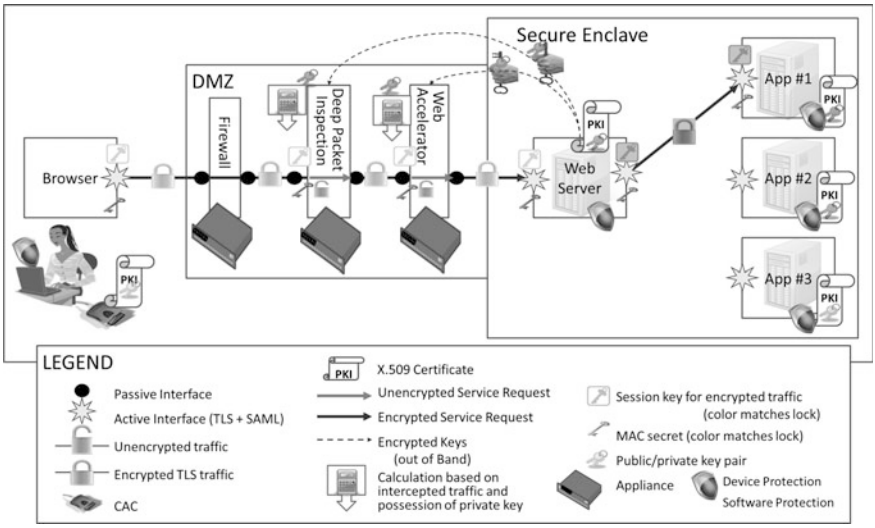


Fig. 3 Current paradigm for encrypted traffic

an easy way to provide the appliance's visibility to the content, it is a singularly bad idea from a security standpoint. Loss of a private key will compromise identity and all sessions and will allow an adversary to impersonate the entity. Further, in high assurance systems, the private keys are locked in a Hardware Storage Module (HSM) and are not shareable. Loss of a session key will entail loss of session confidentiality only. The situation is a bit more complicated in that some of the devices need to see content and some do not. For example, the following data is available for all encrypted packets in the header without decryption:

1. IP Header:
   (a) Time to live
   (b) Source IP
   (c) Destination IP
   (d) IP version and flags
2. TCP Header:
   (a) Source port
   (b) Destination port
   (c) Sequence number
   (d) Acknowledgment number
   (e) Windows size
   (f) TCP flags

3. TLS Handshakes and Header:

   (a) Version, cipher suite, compression algorithm, extensions
   (b) Server certificate and partial chain to root CA
   (c) Client certificate and partial chain to root CA
   (d) Session IDs, client and server random values
   (e) Alerts (full content) sent prior to encryption
   (f) Message types and lengths
   (g) Renegotiations
   (h) Trusted CA list at server
   (i) Client supported cipher suites list and preference
   (j) Client supported compression algorithms

4. Firewall Capabilities not requiring Decryption:

   (a) Blacklist/Whitelist based on IP/port numbers
   (b) Block unacceptable TLS versions, cipher suites
   (c) Block insecure TLS renegotiations
   (d) Block compromised or unknown CAs

3 An Alternative to Private Key Passing

For most interactions using Enterprise Level Security (ELS) approaches, traffic does not need to be inspected. The firewall functionality will still be available using the headers that are not encrypted. However it is recognized that certain circumstances, including cyber-attack indications and/or insider suspicions, and others may require content inspection. For these conditions we recommend an alternative to the sharing of private keys as follows:

HTTP traffic is encrypted using TLS from browser to web application; gateway router at boundary provides access to internal IPs.

1. Web app shares the TLS session keys that are needed to function.

   (a) Firewall: no keys needed, uses headers.
   (b) Deep Packet Inspection: encryption key only, no modification needed.
   (c) Accelerator: encryption key and MAC, to inspect and modify content.

2. No shared private keys.
3. Web application is endpoint for browser requests.

Figure 4 shows the alternative recommendation. The figure illustrates the importance of key management. The user session locks and keys have a life equal to the user session. The web server to appliance locks and keys have a fairly long life to accommodate passing of multiple session keys. The public and private keys have a life specified by the certificate.

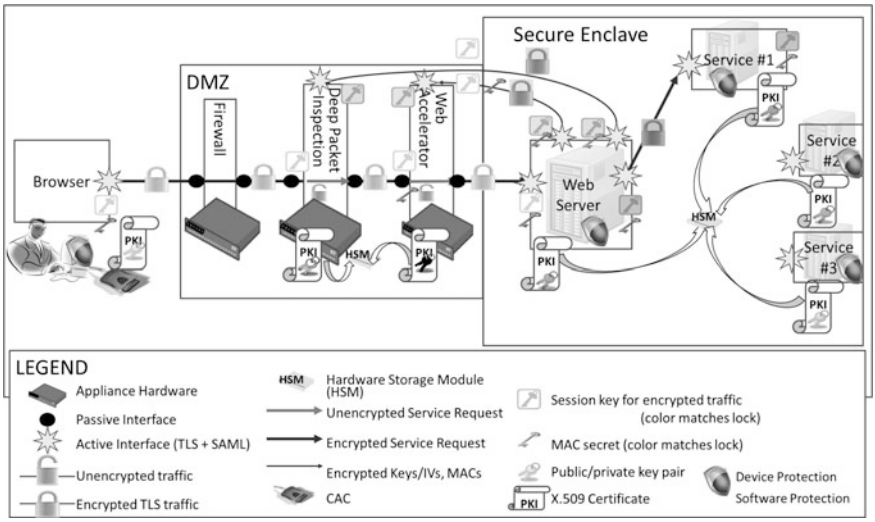


Fig. 4 Alternative encrypted web server communications

4 A Distributed Protection System

The protection system has the capability to monitor, filter, and shut down traffic to given ports. It scans for malicious code. It examines incoming and outgoing traffic for anomalies or known exploits. The protection system checks ports and protocols and can perform other functionality. The protection system acts in the security context of the endpoint for both requester and provider and examines not only the encrypted traffic but also the unencrypted TLS traffic for malicious behavior or code. This requires access to the unencrypted traffic as well as the encrypted traffic. Not all of the checks are provided by the protection system. Fig. 5 walks through checks in a high assurance enclave provided by the protection system, the server handlers, the service handlers and the service itself, minimizing the need for in-line appliances

4.1 Appliance Functionality In-Line

For enclaves that are characterized by full bi-lateral based authentication, private keys should be required to be under the control of hardware storage modules, and the end-to-end paradigm cannot be broken by a passive entity. In-line appliances may be configured as active entities but this is not recommended. The in-line system can only observe headers, apply white and black lists and pass through

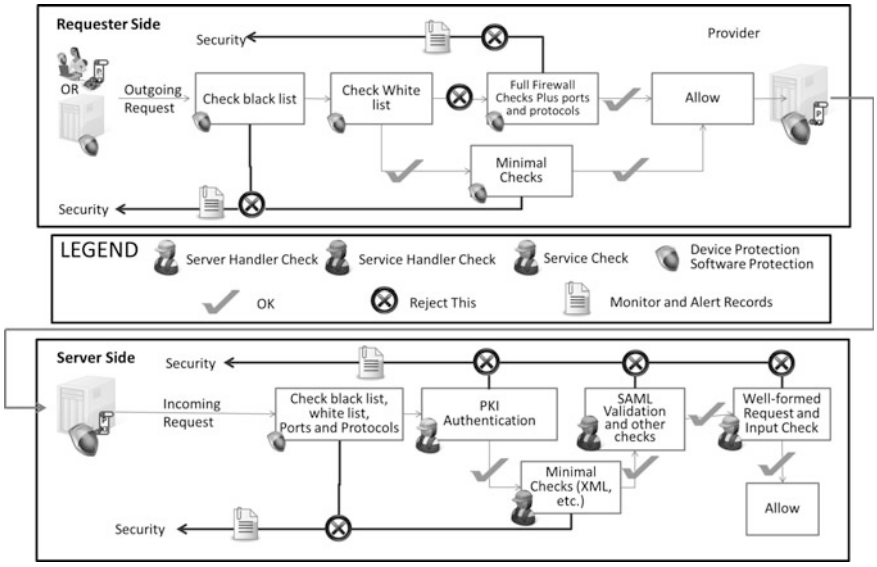


Fig. 5 Distributed protection provided without in-line servers

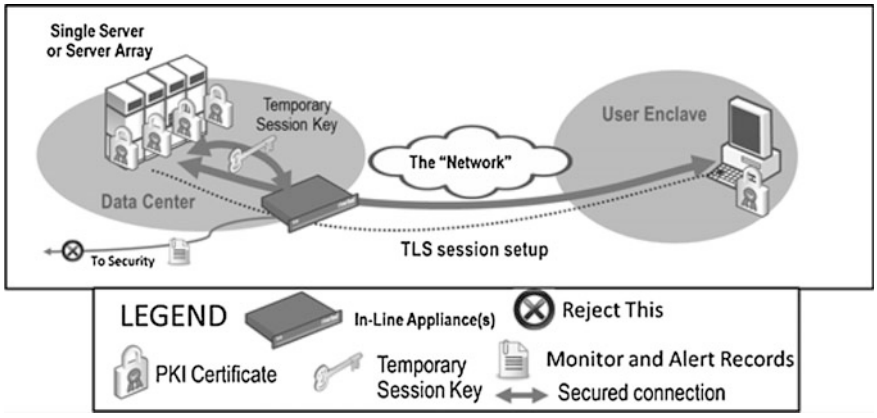


Fig. 6 In-Line appliance functionality

encrypted content without inspection unless it is provided assistance from the service as shown in Fig. 6. This assistance is in the form of passing the session keys where appropriate.

The in-line device may decrypt and inspect incoming traffic. It should either deny the communication (with appropriate logging of information and/or security alerts as appropriate) or pass the communication (unmodified) on to the server. The device may also scan outgoing traffic.

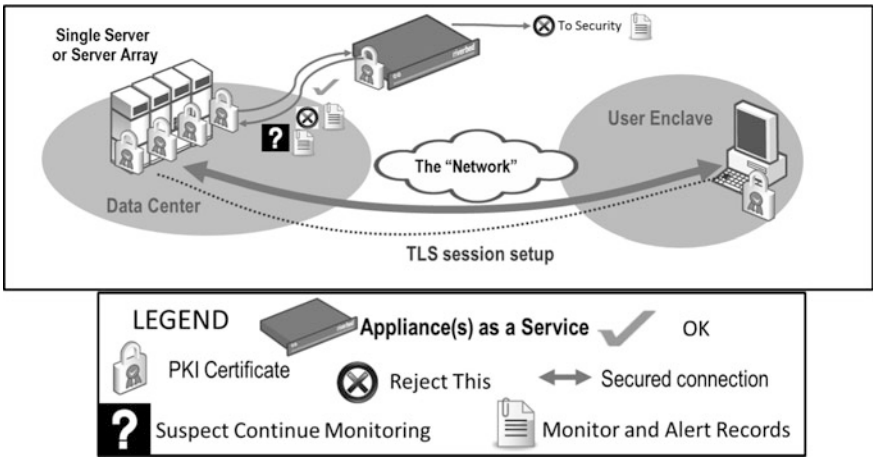


Fig. 7 Appliance functionality as-a-service

4.2 Appliance Functionality as a Service

The appliance system may be used as a service by the application and as such follows full bi-lateral authentication and the TLS security paradigm as shown in Fig. 7.

In order to prevent attackers from gaining access to networks, each device must monitor DHCP requests and report to the central monitor all such requests. This provides listeners throughout the network that allow the central monitor to quickly identify the requesting entity, determine whether it is a known and trusted device or a rogue entity, and take action accordingly.

Any system that is found on the network, through DHCP or other traffic, must identify itself to the protection system before any services are provided to it. This identification is through protection system communications, through which each device authenticates to the central authority and also authenticates the central authority. All such traffic uses end-to-end security, and all devices and their protection systems are registered with enterprise. Unknown entities are not given services and are marked as rogue, which enables local devices to ignore their traffic.

5 Software Only Functionality

Most appliance functionality is now available as software only. Figure 8 shows the conversion of the particular piece of the inspection chain that applies to a particular server as a pseudo appliance.

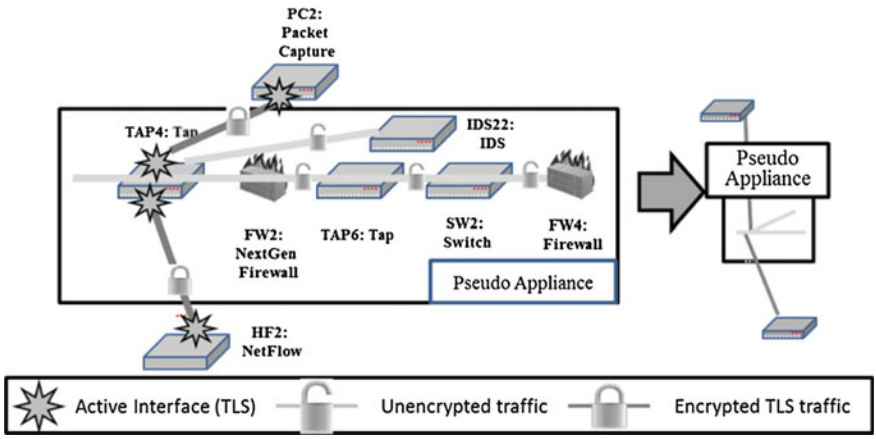


Fig. 8 Creation of the pseudo appliance

The packets are decrypted on entry into the pseudo appliance and stay that way with the exception of an offload to an external source (such as a network monitoring appliance where packets are counted and graded, etc. This offloading will be done by a full ELS communication (for security, of course). This sounds pretty similar to the current approach. How does this change anything? The difference is in where the pseudo appliance lives. For the moment let's assume the software appliance lives in the application server.

5.1 Protecting the Server

Of course we cannot send all of the incoming packets to the application server. It would be inefficient and dangerous. Some packets can wreak havoc before they are even processed, so we would want to be sure that the server was the intended recipient and that the server is communicating with a credentialed entity with valid credentials.

Figure 9 above has identified an intelligent tagging device that will identify the traffic by observing the first few packets. In the case of official enterprise traffic the first few packets are not encrypted and these involve the exchange of PKI certificates that can be identified and the owners can be compared to a white list. The target will be the destination for traffic. Before identification the packets can be passed through the normal set of inspection appliances—sometimes referred to as the de-militarized zone (DMZ). If no identification is made, the packets will continue through the DMZ. When identified, they can be passed directly to the server or the load balancer in front of the server.

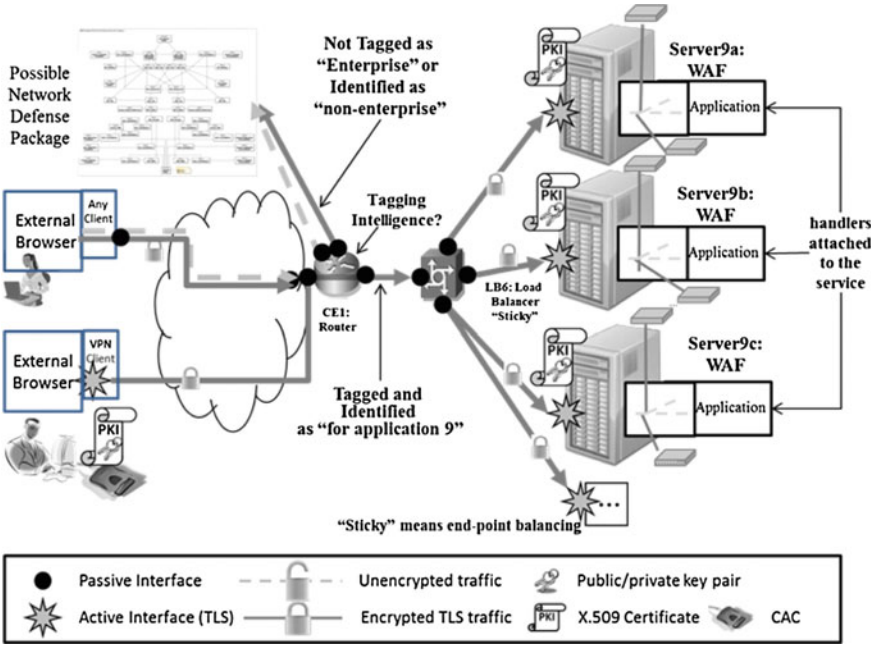


Fig. 9 Tagged and embedded functionality

5.2 Handlers in the Server

The only remaining problem is to reduce the software functionality to handlers in the handler chain as shown in Fig. 10.

Note that the handlers are embedded in the server handler chain at the point that the communication is prepared for their use, and that the functionality has been divided along those lines as opposed to the previous functionality such as virus scan, ports and protocols, intrusion detection or blacklist/whitelist, etc. These are distributed to packet header inspection, packet content inspection, and message content inspection. Each of these may perform inspection related to intrusion detection or blacklist blocking, etc. Pilots are being worked on, stay tuned for results. This is the preferred embodiment for enterprise applications. It moves the inspections to the point of the application itself, by inserting handlers within the server and service to do the inspections at the point it makes most sense. The inspections that can be done without decrypting the packets may be done at the front of the web server because they are passive entities. Moving inspections of decrypted traffic inside the server, not only preserves the end-to-end paradigm, but encapsulates the security and allows tailoring for the application itself. The encapsulated security with the application is virtualization ready.

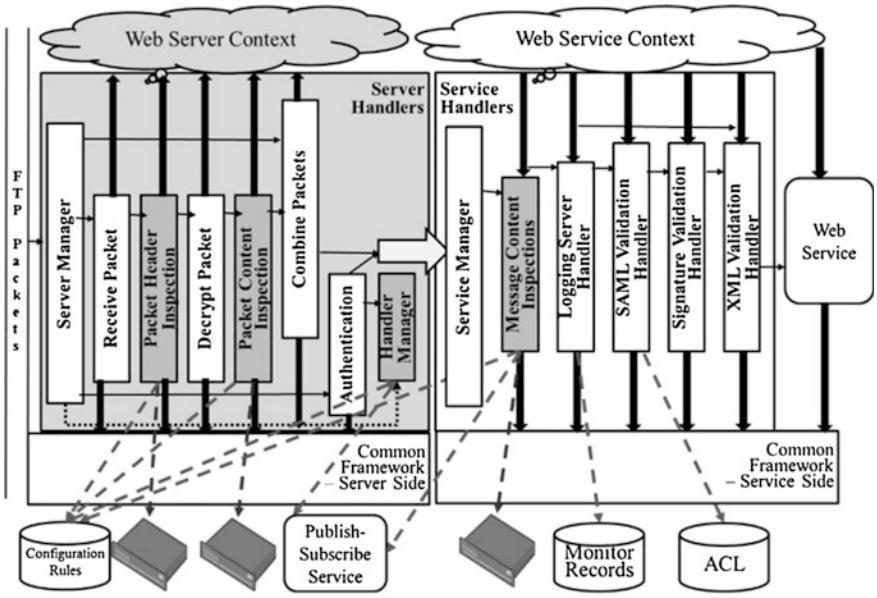


Fig. 10 Server side handlers

6 Summary

We have reviewed the basic approaches to communication protection in computing environments. We have also described high assurance architectures and protection elements they provide. In many instances the high assurance elements provide equivalent protection. In cases where additional protective measures are needed, we have provided three mechanisms for their incorporation. None of these mechanisms requires distribution of a private key which is often done with today's appliances. The distribution of private keys is a fundamental violation of a high assurance model. What remains is the need for high reliability and secure code for passing of private keys, or the establishment of service interfaces on the appliances. More recently we have developed an approach to software only—server embedded appliances which by-passes the need for key passing altogether. This work is part of a body of work for high assurance enterprise computing using web services. Some elements of this work are described in [24–37].

References

1. Chandersekaran C, Foltz K Simpson WR (2014) Distributed versus centralized protection schema for the enterprise. In: Proceedings world congress on engineering and computer science 2014, WCECS2014. Lecture notes in engineering and computer science, San Francisco, USA, pp 68–73
2. Oppliger R (1997) Internet security: FIREWALLS and beyond. Commun ACM 40(5):94
3. Ingham K, Forrest S (2002) A history and survey of network firewalls (pdf)
4. Alkharobi T, Firewalls, presentation found at http://www.ccse.kfupm.edu.sa/~talal/Sec/Firewall.pdf
5. Ingham K, Forrest S (2002) A history and survey of network firewalls (pdf). p 4. Retrieved 25 Nov 2011
6. Cheswick WR, Bellovin SM, Rubin AD (2003) Google books link. Firewalls and internet security: repelling the wily hacker
7. Duhigg C (2003) Virus may elude computer defenses. Washington Post
8. Conway R (204). Code hacking: a developer's guide to network security. Charles River Media, Hingham, p 281. ISBN 1-58450-314-9
9. Chang R (2002) Defending against flooding-based distributed denial-of-service attacks: a tutorial. IEEE Commun Mag 40(10):42–43
10. Almeida V, Bestavros A, Crovella M, de Oliveira A (1996) Characterizing reference locality in the WWW. In: Proceedings of the fourth international conference on parallel and distributed information systems, Miami Beach, Florida, USA, 18–20 December 1996, pp 92–107
11. Altinel M, Bornhövd C, Krishnamurthy S, Mohan C, Pirahesh H, Reinwald B (2003) Cache tables: paving the way for an adaptive database cache. In: Proceedings of the 29th international conference on Very large data bases, Berlin, Germany, 09–12 Sept 2003, pp 718–729
12. Amiri K, Tewari R, Park S, Padmanabhan S (2002) On space management in a dynamic edge cache. In: Proceedings of the fifth international workshop on the web and databases (WebDB 2002) (Madison, Wisc.). ACM, New York, pp 37–42
13. Anton J, Jacobs L, Liu X, Parker J Zeng Z, Zhong T (2002) Web caching for database applications with Oracle web cache. In: Proceedings of the 2002 ACM SIGMOD international conference on management of data, 03–06 June 2002, Madison, Wisconsin. doi: 10.1145/564691.564762
14. Apache HTTP Server Project (2003) Apache HTTP server. http://httpd.apache.org/
15. BEA Systems (2003) Weblogic application server. http://www.bea.com
16. CacheFlow (1999) Accelerating e-commerce with cacheflow internet caching appliances (a CacheFlow white paper)
17. Cain B, Spatscheck O, May M, Barbir A (2001) Request-routing requirements for content internetworking. http://www.ietf.org/internet-drafts/draft-cain-request-routing-req-03.txt
18. Candan KS, Li WS, Luo Q, Hsiung WP, Agrawal D (2001) Enabling dynamic content caching for database-driven web sites. In: Proceedings of the 2001 ACM SIGMOD international conference on management of data, Santa Barbara, California, USA, 21–24 May 2001, pp 532–543. doi: 10.1145/375663.375736
19. Challenger J, Dantzig P, Iyengar A (1999) A scalable system for consistently caching dynamic web data. In: Proceedings of the 18th annual joint conference of the IEEE computer and communications societies (INFOCOM) (New York, NY). IEEE Computer Society Press, Los Alamitos, California, pp 294–303
20. Cunha C, Bestavros A, Crovella M (1995) Characteristics of WWW Client-based traces. Boston University, Boston
21. ESI Consortium (2001) Edge side includes http://www.esi.org
22. Gadde S, Rabinovich M, Chase J (1997) Reduce, reuse, recycle: an approach to building large internet caches. In: Proceedings of the 6th workshop on hot topics in operating systems (HotOS-VI), 05–06 May 1997, p 93

23. Gamma E, Helm R, Johnson R, Vlissides J (1995) Design patterns: elements of reusable object-oriented software. Addison-Wesley Longman Publishing Co. Inc, Boston
24. 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, USA May 2008, pp 12–18
25. 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., USA, June 2008, pp 313–318
26. Chandersekaran C, Simpson WR (2008) The case for bi-lateral end-to-end strong authentication. World Wide Web consortium (W3C) workshop on security models for device APIs, London, England, December 2008, 4 pp
27. Simpson WR, Chandersekaran C (2009) Information sharing and federation. In: The 2nd international multi-conference on engineering and technological innovation: IMETI 2009, vol I, Orlando, FL., USA, July 2009, pp 300–305
28. 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: IMETI 2010, vol 2, Orlando, FL., July 2010, pp 303–308
29. Simpson WR, Chandersekaran C (2010) Use case based access control. In: The 3rd international multi-conference on engineering and technological innovation: IMETI 2010, vol 2, Orlando, FL., USA, July 2010, pp 297–302
30. Chandersekaran C Simpson WR (2012) 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, Berlin-Heidelberg, 2–4 Jan 2012, Bangalore, India, 20 pp
31. 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., USA, April 2011, pp 84–89
32. Simpson WR, Chandersekaran C (2011) An agent-based web-services monitoring system. Int J Comput Technol Appl (IJCTA) 2(9):675–685
33. Simpson WR, Chandersekaran C Wagner R (2011) High assurance challenges for cloud computing. In: Proceedings world congress on engineering and computer science 2011 WCECS 2011. Lecture notes in engineering and computer science, 19–21 Oct 2011, San Francisco, USA, pp 61–66
34. Chandersekaran C, Simpson WR (2012) Claims-based enterprise-wide access control. In: Proceedings world congress on engineering 2012. Lecture notes in engineering and computer science, WCE 2012, pp 524–529
35. Simpson WR, Chandersekaran C (2012) Assured content delivery in the enterprise. In: Proceedings world congress on engineering 2012, WCE 2012. Lecture notes in engineering and computer science, 4–6 July 2012, London, UK, pp 555–560
36. Simpson WR, Chandersekaran C (2012) Enterprise high assurance scale-up. In: Proceedings world congress on engineering and computer science 2012, WCECS 2012. Lecture notes in engineering and computer science, 24–26 Oct 2012, San Francisco, USA, pp 54–59
37. 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

Comprehensive Non-repudiate Speech Communication Involving Geo-tagged Featuremark

A.R. Remya, A. Sreekumar and M.H. Supriya

Abstract The study of audio watermarking has become significant due to its strategic as well as commercial importance. This dissertation addresses the key focus area, concerns in the current audio watermarking practices and what we have done for guaranteeing a secure audio communication. A novel watermarking technique is suggested which utilize audio characteristics and geo-location information to formalize the watermark is later embedded within the audio signal using FWHT. The geo-location information can aid as a secondary security scrutiny in user authentication system or can assist in discovery of the geo-coordinates of the place where the communication happened. The research can further be extended to implement information security in e-governance and in the interest of Law Enforcement, particularly Computer Forensics.

Keywords Audio watermarking · Fast Walsh-Hadamard transform · Featuremarking · Geo-tagging · Mel-frequency cepstral coefficients · Non-repudiation · Spectral-flux · Zero-cross rate

A.R. Remya (✉) · A. Sreekumar
Department of Computer Applications, Cochin University of Science and Technology, Kochi 682 022, Kerala, India
e-mail: remyacusat@gmail.com

A. Sreekumar
e-mail: askcusat@gmail.com

M.H. Supriya
Department of Electronics, Cochin University of Science and Technology, Kochi 682 022, Kerala, India
e-mail: supriyadoe@gmail.com

H.K. Kim et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-94-017-7236-5_13

1 Introduction

Advances in digital technology and widespread use of digital communication in various areas including government, legal, banking, military applications in turn has increased the reproduction and re-transmission of multimedia data through both legal and illegal channels. However in today's information driven society, illegal usage of digital media is a serious threat to the content owner's authority or proprietary right which emphasize the utmost want to authenticating the information that is sent across various communication channels. In certain dispute circumstances involving digital communication lack of evidence may result in denial of authorship, denial of sending or receiving the signal, denial of time or location of occurrence.

Audio watermarking originated as a sub-discipline of digital signal processing and elevated itself to main stream research and development. This discipline mainly focuses on appropriate signal processing techniques to embed vital information to audio sequences. Recent copyright infringements in digital communication make us believe that the stronger analytical tools and methods need to be researched on. In order to combat this malicious usage of digital audio communication we need to understand the existing audio watermarking schemes especially that are aligned towards Intellectual Property Rights (IPR);understand some of the best practices in existing watermarking schemes and foster a differentiator methodology which in turn transcend these breaches.

Developing a non-repudiate [1] voice authentication scheme is a challenging task in the context of audio watermarking. The embedded watermark should be robust to any signal manipulations and can be unambiguously retrieved at the other end. Presently different audio watermarking methods are available, most of them inclined towards copyright protection and copy protection. This was the motivation for the key notion to develop a speaker verification scheme that guarantee non-repudiation services and the proposed scheme is its outcome.

The intended work introduce a novel but comprehensive voice signal authentication schemes that assure non repudiation by utilizing the key acoustic signal features towards the preparation of the watermark. ANN, k-NN and SVM classifiers are employed in determining the appropriate acoustic features in the new FeatureMark. The acoustic characteristics such as Mel-frequency cepstral coefficients (MFCC), spectral flux and zero-cross rate are taken into consideration for the watermark preparation. Geo-tag information as well as cryptographic hash function is opted in the preparation of the watermark to improve authenticity. Embedding the FeatureMark generated involve Fast Walsh-Hadamard Transform (FWHT) technique.

Labelling geo-coordinates or geographical location in a digital media is termed as Geotagging which is invisible to the human eyes. Existing geotag application employ techniques to hold the location information as textual metadata in case of mp3 it may be ID3 or Vorbis comment in case of Vorbis and Opus file formats. The metadata containers hold title, artist, album, track number, and other information about the media and can be retrieved easily from the audio without much toil.

The proposed system will embed the geo information implicitly in the audio signal and not as metadata there by making it difficult to retrieve. The system devises a unique watermark which is composed of key audio features in the signal along with the geotag information.

The primary objective of proposed work is to assess the prevailing audio watermarking methods and formalize a scheme designed to exploit the acoustic characteristics, geo location information of the parties involved in the communication and craft a novel scheme that ensure non-repudiation service. It is not possible to formally prove the rightness or falsehood of the proposed system but the experimental results provide strong evidence to substantiate its betterment in terms of imperceptibility, robustness and capacity when compared to the existing schemes.

2 Related Works

A better understanding in the field of audio watermarking is obtained by conducting a comprehensive literature review which leads to the identification of some of the unresolved issues within the existing implementations.

The paper [2] demonstrates a security enhanced biometric incorporated speaker identification system that can be employed in forensic applications. In order to prevent unauthorized copying of digital audio, a psychoacoustic model is designed which embeds the watermark into the frequency domain of the signal by employing the DSSS methodology [3]. Another scheme suggested in [4] also inserts the watermark bits by DSSS methodology and employing the LPC parameters is also a copy-right protection scheme which includes a psychoacoustic model of audio coding [5]. A content-based audio authentication scheme introduced in [6] embeds the watermark data in the same domain as the content feature extraction without causing any impairment of the same. Automatic identification of audio recordings of ethnic music by employing hidden Markov models (HMMs) is presented in [7]. Research on psychoacoustics exploits the use of embedding an arbitrary message, the watermark, in a digital recording without altering sound perception [8]. A cyclic pattern embedding watermarking algorithm to detect speech forgery is presented in the paper [9]. An audio dependent watermarking procedure suggested in the work [10] creates its watermark by segmenting the audio clips and adding a perceptually shaped pseudo-random sequence, which in the complete system acts as the author signature and guarantees the copyrighting of each time-domain signal. The work implemented in [11] helps to identify the owner and the distributor of digital data by employing the frequency domain of each multimedia files.

An author identification scheme with speech watermarking presented in [12] utilizes a combination of spread-spectrum methodology with frequency masking. Employing CPN networks in the context of audio watermarking is presented in [13] to confirm the copyright authentication. The paper [14] introduces an audio watermarking technique for content integrity authentication and copy-right

protection. A procedure for logo watermarking of speech signal presented in [15] which selects the DFT coefficients on time-frequency basis and inserts the logo in it.

Perceptible watermarking for encrypting the audible data into inaudible data as the first step and imperceptible watermarking for taking care of the copyright issues is presented in the work [16]. A signal-dependent sequence using modified Swanson's method is suggested [17] as the watermark and ANN is used towards embedding the watermark. An authentication scheme for music audio is introduced in [18] using DWT in which the audio is authenticated using the features extracted by the wavelet transform and characteristics coding. The work done in [19] deals with content-fragile as well as invertible watermarking approaches. In order to overcome the problem of backward-compatible audio authentication, a distributed source coding based scheme presented in [20] is robust against legitimate encoding variations and detects any illegitimate modifications.

3 Fundamental Theory

3.1 Audio Signals

In this present world, we are coming across different kinds of signals in various forms. An audio signal is a representation of sound, usually in decibels and rarely as voltages [21]. An exciting human sensory capability is the hearing system [22]. Audio frequency range denotes the limits of human hearing and is in the range of 20–20,000 Hz [23] and intensity range of 120 dB. A digital audio signal is the result of suitable sampling and quantization performed on an audio signal with a sampling rate of 44,100 Hz. Audio signal processing, sometimes referred to as audio processing, is the intentional alteration of auditory signals or sound, often through an audio effect or effects unit. As audio signals may be electronically represented in either digital or analog format, signal processing may occur in either domain. Analog processors operate directly on the electrical signal, while digital processors operate mathematically on the digital representation of that signal [23].

The properties of an audio event can be categorized as temporal or spectral properties. The temporal properties refer to the duration of the sound and any amplitude modulations; the spectral properties of the sound refer to its frequency components and their relative strengths. As discussed in [24], features are designed with the help of salient signal characteristics in terms of signal production or perception.

3.1.1 Mel-frequency Cepstral Coefficients [MFCC]

Mel-frequency cepstral coefficients introduced by Davis and Mermelstein in 1980s are treated as the best parametric representation of the acoustic signals employed in the recognition of speakers and have been the state-of-the-art ever since. Mel-scale

relates perceived frequency or pitch of a pure tone to its actual measured frequency. MFCCs are based on a linear cosine transform of a log power spectrum on a non-linear Mel-scale of frequency.

The formula for converting from frequency to Mel scale is:

$$\mathrm{M(f)} = 1125\ln(1 + \frac{\mathrm{f}}{700}) \tag{1}$$

To go from Mels back to frequency:

$$\mathrm{M}^{-1}(m) = 700(\exp\left(\frac{\mathrm{m}}{1125}\right) - 1) \tag{2}$$

The Mel-frequency scale, a perceptual scale like the critical band scale, is linear below 1 kHz and logarithmic above this frequency. Logarithm of the band-wise power values are taken and de-correlated by applying a DCT to obtain the cepstral coefficients. The log transformation serves to de-convolve multiplicative components of the spectrum such as the source and filter transfer function. For instance, in 16 kHz sampled speech, 13 low-order MFCCs are adequate to represent the spectral envelope across phonemes [23, 25].

3.1.2 Spectral Flux

The spectral flux also termed as spectral variation is a measure of how quickly the power spectrum varies corresponding to each frames in a short-time window. It can be described as the local spectral rate of change of an acoustic signal. Timbre of an audio signal can also be derived from it [23]. A high value of spectral flux stands for a sudden change in the spectral magnitudes and therefore a possible spectral boundary at rth frame.

Spectral flux can be calculated as follows:

$$F_r = \sum_{k=1}^{\frac{N}{2}} \left(|X_r[k]| - |X_{r-1}[k]|\right)^2 \tag{3}$$

where $X_r[k]$ represents the normalized magnitudes of spectral distribution corresponding to signal frame X_r.

3.1.3 Zero-Cross Rate

Zero-cross rate another key feature used in classifying voice signals or musical sounds. ZCR is calculated for each frame and is defined as the rate of sign changes along a signal [23].

$$ZCR = \frac{1}{T-1}\sum_{t=1}^{T-1}\prod\{S_t S_{t-1} < 0\}. \tag{4}$$

where S is a signal of length T and the indicator function Π{A} is 1 if its argument A is true and 0 otherwise.

3.2 Walsh-Hadamard Transform

Unlike most transforms Hadamard transform also known as Walsh-Hadamard transform, is not based on sinusoidal function instead, elements in its transform matrix are either 1 or −1 and needs no multiplications for its computation, leading to simple hardware implementations. Walsh transforms are the discrete analog of the Fourier transforms. When N = 2n, its transform matrix is defined by the following recursion [1].

$$H_1 = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \quad H_n = \frac{1}{\sqrt{2}}\begin{bmatrix} H_{n-1} & H_{n-1} \\ H_{n-1} & -H_{n-1} \end{bmatrix} \tag{5}$$

From the above equations it is evident that the Hadamard transform is unitary. The forward and inverse Walsh transform pair for a signal x(t) of length N are

$$y_n = \frac{1}{N}\sum_{i=0}^{N-1} x_i WAL(n,i), \quad n = 1,2,\ldots,N-1. \tag{6}$$

$$x_i = \frac{1}{N}\sum_{n=0}^{N-1} y_n WAL(n,i), \quad n = 1,2,\ldots,N-1. \tag{7}$$

Walsh transform decomposes an arbitrary input vector into a superposition of Walsh functions that comes under generalized class of Fourier transforms. FWHT () is an efficient algorithm to compute the Walsh transform leading to reduced complexity of O(NlogN). Walsh analysis is given in the previous work [1].

3.3 Hash Functions

Hash is defined in [26] as a kind of signature or digest for a stream of data that represents the contents. In informal terms it is the "checksum". A cryptographic hash function takes variable sized input and delivers fixed sized output and these functions are computationally efficient and pseudorandom. Popular hash algorithms include MD4 (128 bits, obsolete), MD5 (128 bits), RIPEMD-160 (160 bits), SHA-1

(160 bits), SHA-256, SHA-384 and SHA-512 (longer versions of SHA-1 with slightly different designs).

Avalanche effect denotes effect of changes in hash with a very small change in input data stream. All input streams yields hashes of same length with the hashing algorithms and is a one-way operation that converts a stream of data into a digest with a fixed size. An important application of cryptographic hash function is message authentication and in this work MD5 hashing algorithm is utilized. Either the message M or hash or both can be encrypted for improved confidentiality.

3.4 Geo-tagging

Geo-tagging is the term used to represent assigning automatic labelling of geographical information about the multimedia content. As stated in [27], the purpose of assigning labels to content is to achieve a representation that encodes a higher level of semantic abstraction and from this paper it is evident that automatic labelling or tagging falls in any one of the following categories: open-set tagging or closed-set tagging. In open-set tagging the identity of the labels to be assigned need not be known in advance. But in the second approach labels are assigning from a fixed set or known set of labels.

GPS coordinates may be represented in text in a number of ways such as decimal degrees with negative numbers for South and West; degrees and decimal minutes with N, S, E or W suffix for North, South, East or West; degrees and decimal minutes with N, S, E or W prefix for North, South, East or West; degrees, minutes and seconds with N, S, E or W suffix for North, South, East or West; degrees, minutes and seconds with N, S, E or W prefix for North, South, East or West etc. [23]. In the proposed scheme latitude and longitude are stored in units of degrees with decimals:

GPS Latitude: 10.044826
GPS Longitude: 76.327547
GPS Position: 10.044826 76.327547

4 Non-repudiate Featuremarking Scheme

The work presented in this paper is an enhancement of the scheme suggested in [1] that employs MFCC, spectral flux and zero-cross rate towards its watermark preparation. Another modification that can be identified in this scheme is the use of GPS system for obtaining the location details of the spoken input as well as the use of MD5 hashing technique for generating a fixed bit data stream for the watermark preparation.

4.1 Watermark Preparation

Signal dependent physical features from the speech signal such as mel-frequency cepstral coefficients (MFCC), spectral-flux and zero-cross rate (ZCR) along with the exact geo location of the speaker with the latitude and longitude values obtained using a standalone GPS system are employed towards the preparation of the FeatureMark. Feature selection was based on the classification module that works with ANN, kNN and SVM classifiers.

Encoding data as its 1-D and 2-D form can be achieved with Barcodes and Datamatrix codes [1] respectively. A data carrier represents data in a machine readable form; used to enable automatic reading of the Element Strings. This scheme suggest Datamatrix code as its watermark and use the signal dependent perceptual feature such as MFCC, spectral flux and zero-cross rates for its preparation. A Datamatrix code is the two-dimensional matrix representation which encodes text or numeric data and can be represented as a square or rectangular pattern with varying arguments of black and white cells in accordance to the information to be encoded. The length of encoded data depends on the number of cells in the matrix. This large data stream is given as input to an MD5 hash algorithm and obtains a fixed 32-bytes hexa-decimal data stream which includes the two 16-bytes output half buffers. Decimal equivalent of this data stream and its modulus 7 value is derived which later is fed to an encryption scheme that inserts the modulus value to the hexadecimal hash value to improve the confidentiality.

$$y = \bar{x} \biguplus_k \tag{8}$$

y is obtained by inserting $\bar{x} \bmod 7$ at the kth position of $\bar{x}$, where $\bar{x} = x_1 x_2 \ldots x_n \cdot x_{n+1} x_{n+2} \ldots x_{n+k}$ is the hash value obtained by MD5 execution and $0 \leq k \leq m$ and $m = n + k$;

Obtained hash value is given as input to an online Datamatrix code generator that generates a signal dependent, location dependent Datamatrix code. Generated Datamatrix code is unique and guarantees the individuality of the owner and is used as the watermark for the proposed scheme (Fig. 1).

Fig. 1 Encrypted Datamatrix code

4.2 Embedding Scheme

Synchronization code embedding is necessary with all audio watermarking schemes and detection of watermark bits gets easy by embedding the synchronization code bits in the signal. Any variation in synchronization involving any spatial or transform domain modifications cause the detector loses its synchronization and results in false detection. In the proposed approach, a 16-bit Walsh code is embedded in front of the FeatureMark to locate its position. In order to embed these bits into the audio signal, the matrix is read row wise and then calculated its auto-correlation function. Obtained 16-bit sequence is employed as the synchronization code and embedded in time-domain sequence of the speech signal [1].

Embedding module in this scheme functions through different levels: Synchronization code embedding and FeatureMark embedding. The First step associated with the embedding module is the segmentation where the original speech signal is divided into a set of segments and then each segment to two sub-segments. Secondly, with the spatial watermarking technique, the Walsh code bits are embedded into the sub-segments. After embedding the synchronization code, the subsequent segments are transformed using the Fast Walsh Transform.

4.2.1 Synchronization Code Embedding

Robust and transparent nature of audio watermarking scheme is achieved by embedding synchronization code bits in it. In this scheme we embedded the Walsh code into the time-domain samples according to the algorithm presented in the work [1].

4.2.2 FeatureMark Embedding

In the FeatureMark embedding scheme, Arnold transform is applied to the generated FeatureMark image in order to dissipate its pixel space relationship. Scrambled image thus obtained is converted into a 1-dimensional sequence of 1s and 0s (binary digits) in order to embed the bits accurately into the 1-dimensional audio signal.

Let $V = v(i), 0 \leq i <$ Length represent a host digital audio signal with Length samples. $FM = FM(i,j), 0 \leq i < M, 0 \leq j < N$ is a binary image to be embedded within the host audio signal and $FM(i,j) \in 0, 1$ is the pixel value at $(i,j), S = s(i), 0 \leq i \leq L_{syn}$ is a synchronization code with L_{syn} bits, where $s(i) \in 0, 1$.

Applying Arnold transform results in a scrambled structure which can be represented as $FM1 = FM1(i,j), 0 \leq i \leq M, 0 \leq j \leq N$.

Scrambled structure will then be converted into a sequence of 1s and 0s as follows:

$$FM2 = fm2(k) = FM1(i, j),$$
$$0 \le i \le M,\ 0 \le j \le N,\ k = i \times N + j,\ fm2(k) \in 0, 1$$

SYNC code embedding is followed by actual mark embedding scheme where speech signal is transformed using the fwht () transform and the mark bits of 1s and 0s are embedded. Embedding the mark bits follow the same sequence of steps presented in [1]. If the embedding condition fails in any case, we need to actually replace the original Walsh coefficients with the corresponding mark bits. After embedding all the watermark bits into the signal, the signal should be reconstructed by inverse FWHT transform and also by combining the set of overlapping and non-overlapping window-frame series.

4.3 Detection Scheme

Inverse Walsh transform involving IFWHT function on the transform domain data recovers the original time domain signal. The 'symmetric flag' helps to nullify the numerical inaccuracies, in other words to zero out the small imaginary components and to recover the original signal. The original time domain signal and the recovered time domain signal behave almost identically and does not reveal any significant difference while playing the audio.

The Blind FeatureMark detection scheme functions as a two-step process where synchronization code detection followed by the FeatureMark detection. Synchronization code detection involves identifying the presence of the Walsh code in the signal segments/sub-segments. Since the Walsh code bits are embedded in the time-domain, we can directly check the presence of these bits in the sub-segments. Presence of FeatureMark bits are determined by evaluating the spectrum of each frame by Fast Walsh Transform. Mark bits are now fetched out of the signal and are arranged in a matrix of order $M' \times N'$ to finalize the FeatureMark.

In an ideal case, extracted mark bits generate the original FeatureMark image which is the Datamatrix code and the values used in generation of Datamatrix code can be retrieved by the help of an online Datamatrix code scanner. Identifying and retrieving the modulus value provides the original hash output. As MD5 algorithm is reversible we can retrieve the original data stream from the two 16-bytes half buffers which includes the signal features as well as the geo location details of the speaker.

5 Experimental Results and Discussions

Various experiments were conducted to evaluate the performance of this scheme by piloting imperceptibility tests, robustness tests and capacity tests. Malayalam speech signals were primarily considered in this assessment from around 10 members with signal characteristics of 16 bits/sample with sampling rate of 44,100 kHz. Matlab version R2009b and music editor sound recorder were used for inducing common signal manipulations and desynchronization attacks. Signal duration varies for each sample signal with some selected speech signals enduring 300 s. Apparently, results of the experiments conducted, substantiate that the embedding scheme is sound.

GPS system reveals the current location as (10.044826, 76.327547) that represents the latitude and longitude value corresponding to Cochin university of science and technology. MFCC, spectral flux and zero-cross rate values along with this GPS position is given to produce the corresponding MD5 hash value.

MD5 algorithm produces a message digest of length 32-byte which constitutes two 16-byte half buffers. In this work, the obtained 16-byte half buffers for a spoken input are

Buffer 1: 3f1234a66d4e2fb3 (16 byte)
Buffer 2: a8d4266215bee4da (16 byte)

Next phase is to obtain the modulus 7 of the obtained hash value. For that we have converted the hash value to its decimal equivalent and employed the Matlab's mod (x, y) function towards this. In this work, decimal equivalent of the hash value is 83,835,892,948,906,000,000,000,000,000,000,000,000.00 and its modulus 7 gives 0. Employing the encryption algorithm Eq. (8) results in a data stream 3f12340a66d4e2fb3a8d4266215bee4da of 33 bytes. For this signal the key selected for inserting modulus value is 7. The extracted FeatureMark without being attacked is obtained with BER = 0.0 and PSNR = 0.0. With single channel audio signals we can directly use the selected FWHT coefficients for embedding the FeatureMark bits and in the case of stereo signals watermark bits are embedded by considering the coefficients from two channels without affecting its signal properties.

In this work, the digital watermark is embedded into the original voice signal by transforming it using FWHT. During the transmission process, the FeatureMarked voice signal may suffer various signal manipulations including noise addition, silence addition, echo addition, re-sampling, re-quantization, low-pass and band-pass filtering or other desynchronization attacks such as amplitude variation, pitch shifting, random cropping and time-scale modification. These attacks are simulated in experiments conducted to evaluate the proposed work.

5.1 Transparency Tests

Transparency of the proposed scheme is evaluated by employing subjective listening tests. Selection of listening panel is based on a set of non-expert listeners involving fellow research scholars and colleagues by considering the fact that non-expert listeners may be representative of the general population. The evaluation is based on a 5-point grade scale [28] that demonstrates the quality versus impairment of a watermarked audio signal. In this, excellent signal quality denotes imperceptible, good indicates perceptible but not annoying, fair dictates slightly annoying, poor is for annoying and bad stands for very annoying. The imperceptibility criteria obtained for this scheme is excellent for the set of non-expert listeners.

5.2 Robustness Tests

Robustness tests evaluate the FeatureMark detection accuracy against common signal manipulations and desynchronization attacks. The procedure used for these common signal manipulations such as silence addition where a silence of 100 ms duration is inserted at the beginning of the FeatureMarked signal, echo addition where an echo with a delay of 200 ms is added, low-pass and band-pass filtering which are done with cut off frequencies 10 kHz and 200 Hz are performed using the freely available music editor tool. Other signal manipulations include noise addition where a white Gaussian noise with an SNR of 50 is added, re-sampling where the signal is down samples to frequencies 22.05 kHz and then up sampled to its original 44.1 kHz and re-quantization where signal with 16-bit quantized to 8-bit and then back to its original 16-bit are performed in Matlab. Robustness tests on common signal processing functions give a low value for BER.

The procedure used for desynchronization attacks such as amplification to double volume and half volume, pitch change such as frequency fluctuation to the signal, speed changes such as lengthening (slow down) or shortening (double speed) and random cropping are done using the Music editor software. The evaluation is performed on both single channel and multi-channel audio signals. Performance of the proposed FeatureMarking scheme is analyzed by evaluating the bit error rate (BER) of the extracted FeatureMark to the actual embedded FeatureMark.

Recovery rate is calculated using the equation $(1 - BER) \times 100\,\%$ that denotes the knack of the proposed scheme in detection and reconstruction of the embedded watermark without any failure. The BER values obtained and recovery rate shows that the proposed scheme can be applied with the real-time applications involving audio/speech watermarking. Average of the recovery rate obtained for some signals are plotted in Fig. 2.

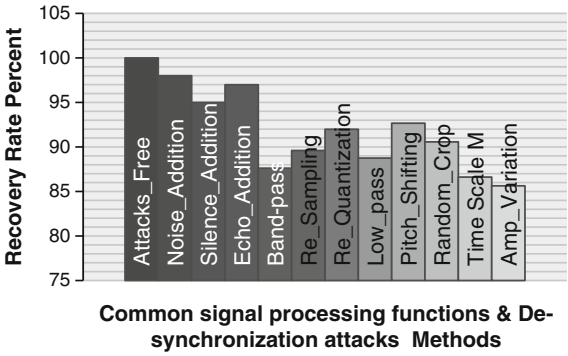


Fig. 2 Average recovery rate

5.3 Capacity Tests

Capacity or data payload demonstrates the amount of information that can be embedded and recovered in the audio stream. It can be evaluated using the Eq. 9.

$$C = \frac{M}{L} \text{bps}, \tag{9}$$

where M refers to the number of watermark bits and L the length of the audio signal.

Theoretical evaluation demonstrates that this scheme holds (4356/300) = 14 watermark bits per second. The capacity obtained for this new algorithm varies depending upon the length of the watermark and with the signal duration we have taken. For a sample audio signal with 3 s duration and 1560 bits watermark length, the results show a capacity of 520 bps.

6 Summary and Concluding Remarks

This paper addresses the key focus area audio watermarking practices with an improvement on what we have done for guaranteeing a secure audio communication [1]. In this method audio characteristics and geo location information were used to formalize the watermark which is later embedded within the audio signal using FWHT. The geo-location information can aid as a secondary security scrutiny in user authentication system or can assist in discovery of the geo-coordinates of the place where the communication happened.

A comparative study of the newly suggested watermarking scheme with existing schemes highlight the fact that unlike existing schemes the watermark generated is dynamic watermark which holds the audio characteristics of the users involved in the communication along with their geo location information. Another key highlight is use of Walsh transform for embedding the watermark in the audio signal.

The use of voice signal features, its classification and feature marking offers an improved scheme for authentic voice communication. The scheme devises Walsh code as its synchronization code to improve the robustness nature. A Cryptographic hash function, message digest MD5 is incorporated that adds one more layer of security to the signal dependent dynamic watermark. Thus the proposed audio watermarking scheme can outperform the existing techniques in terms of high success rates. Results from various experiments endorse the proposed system to be outstanding in terms of imperceptibility, robustness and capacity. The system is proficient of assuring copy protection, copyright protection, and authorship proof.

The research can further be extended to implement information security in e-governance and in the interest of Law Enforcement, particularly Computer Forensics. In the future, we will investigate the combination of multiple acoustic features and can integrate the proposed system as a module of a complete geo-tagging framework which lacks any kind of textual metadata. The results also indicate that the proposed scheme neither compromised audibility of the speech signal nor robustness against various signal processing nor desynchronization attacks.

Acknowledgments This work was funded by the Department of Science and Technology, Government of India under the INSPIRE Fellowship (IF110085).

References

1. Remya AR, Sreekumar A (2014) An FWHT based non-repudiate speech communication. 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 115–120
2. Faundez-Zanuy M, Hagmuller M, Kubin G (2007) Speaker identification security improvement by means of speech watermarking. Pattern Recogn 40(11):3027–3034
3. Seok J, Hong J (2001) Audio watermarking for copyright protection of digital audio data. Electron Lett 37(1):60–61
4. Seok J, Hong J, Kim J (2002) A novel audio watermarking algorithm for copyright protection of digital audio. etri J 24(3):181–189
5. Lin P-L (2001) Digital watermarking models for resolving rightful ownership and authenticating legitimate customer. J Syst Softw 55(3):261–271
6. Gulbis M, Muller E, Steinebach M (2009) Content-based audio authentication watermarking. Int J Innovative Comput Inf Control 5(7):1883–1892
7. Orio N (2010) Automatic identification of audio recordings based on statistical modeling. Sig Process 90(4):1064–1076
8. Boney L, Tewfik A, Hamdy K (1996) Digital watermarks for audio signals. In: Proceedings of the third IEEE international conference on multimedia computing and systems. IEEE, pp 473–480
9. Park C, Thapa D, Wang G (2007) Speech authentication system using digital watermarking and pattern recovery. Pattern Recogn Lett 28(8):931–938
10. Swanson M et al (1998) Robust audio watermarking using perceptual masking. Sig Process 66 (3):337–355

11. Yucel Z, Ozguler AB (2010) Watermarking via zero assigned filter banks. Sig Process 90 (2):467–479
12. Faundez-Zanuy M, Hagmuller M, Kubin G (2006) Speaker verification security improvement by means of speech watermarking. Speech Commun 48(12):1608–1619
13. Chang C, Wang H, Shen W (2010) Copyright-proving scheme for audio with counter-propagation neural networks. Digit Signal Proc 20(4):1087–1101
14. Lei B, Soon Y (2012) A multipurpose audio watermarking algorithm with synchronization and encryption. J Zhejiang Univ Sci C 13(1):11–19
15. Orovic I et al (2008) Speech signals protection via logo watermarking based on the time–frequency analysis. In: Annals of telecommunications-annales des telecommunications, vol 63 (7–8), pp 369–377
16. Dutta M, Gupta P, Pathak V (2012) A perceptible watermarking algorithm for audio signals. In: Multimedia tools and applications, pp 1–23
17. Tsai H, Cheng J (2005) Adaptive signal-dependent audio watermarking based on human auditory system and neural networks. Appl Intell 23(3):191–206
18. Yoshitomi Y et al (2011) An authentication method for digital audio using a discrete wavelet transform. J Inf Secur 2:59
19. Steinebach M, Dittmann J (2003) Watermarking-based digital audio data authentication. EURASIP J Appl Sig Process 2003:1001–1015
20. Varodayan D, Lin Y, Girod B (2008) Audio authentication based on distributed source coding. In: IEEE international conference on acoustics, speech and signal processing. ICASSP 2008. IEEE, pp 225–228
21. MusicTech (2011) Audio signal levels tutorial. http://www.musictech.net/2011/03/10mm-187-audio-signal-levels-explained/
22. Plack C (2007) Auditory perception. http://socialscientist.us/nphs/psychIB/psychpdfs/PIP_Auditory_Perception.pdf/
23. Wikipedia: the free Encyclopedia
24. Prasad B, Prasanna S (2008) Speech, audio, image and biomedical signal processing using neural networks. Springer, Berlin
25. Lyons J (2009) Practical cryptography. http://practicalcryptography.com/
26. Friedl S (2005) An illustrated guide to cryptographic hashes. Retrieved from Unixwiz. http://www.unixwiz.net/techtips/iguide-crypto-hashes.html
27. Larson M et al (2011) Automatic tagging and geotagging in video collections and communities. In: Proceedings of the 1st ACM international conference on multimedia retrieval. ACM
28. ITU-R (2002) Methodology for the subjective assessment of the quality of television pictures. http://www.itu.int/

Comparison of Conceptual Class Diagrams for Verifying Software Model Redesign

Pattamaporn Saisim and Twittie Senivongse

Abstract An existing software system sometimes needs to be redesigned to accommodate various change requirements. Since incomplete software requirements will lead to incorrect design of the new system, the system analyst needs to verify that the gathered requirements for the new system are complete, i.e., those that should be retained in the new system are not missing and those that are changed or newly introduced are included. This paper presents a method to help the system analyst to verify the redesign of the software system. As an initial model created from the new software requirements, the conceptual UML class diagram of the new system is compared with that of the existing system. The comparison algorithm called S-UMLDiff considers similarity of the diagram structure and semantic similarity of names in the two diagrams. The reported similarities and differences between the diagrams can assist the system analyst in reviewing the conceptual model of the new system to verify early on whether the new design is correct and built upon a complete set of change requirements. The paper also presents a comparison tool and good results of the evaluation of S-UMLDiff performance.

Keywords Conceptual model · Semantics · Software change requirements · Software redesign · UML class diagram · WordNet

P. Saisim · T. Senivongse (✉)
Department of Computer Engineering, Faculty of Engineering, Chulalongkorn University, Phyathai Road, Pathumwan, Bangkok 10330, Thailand
e-mail: twittie.s@chula.ac.th

P. Saisim
e-mail: onzony@gmail.com

P. Saisim
KTB Computer Services, 22/1 Sawai Brown 2 Building, Sukhumvit Soi 1, Klongtoey Nua, Wattana, Bangkok 10110, Thailand

H.K. Kim et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-94-017-7236-5_14

1 Introduction

Software systems need to undergo changes constantly. Changes may either be applied directly to existing systems to add or fix certain functions, or they require the systems to be redesigned and reconstructed. The motivation behind redesigning an existing system can be that there are changes in concepts, processes, or functions within the business domain which necessitate changes in the software system structure. To redesign the software system, the system analyst restarts the whole development process by eliciting new user requirements to gather change requirements as well as studying the requirement specification of the existing system. While certain requirements of the existing software system should remain in the new version of the system, some of them may be dropped and some new requirements are to be added. The problem that may arise is that the users and the development team may be a different group from those who gave the original requirements and developed the existing system. This may result in the software requirements of the new system being incomplete as the users may forget or even not know of certain functions or data that should be retained, and the new development team may not fully understand the business domain. Since incomplete software requirements will lead to incorrect design of the new system, the system analyst needs to verify that the gathered requirements for the new system are complete.

To help ensure that the new version of the system will be developed according to the correct change requirements, we present an approach to verifying software system redesign through a comparison of the conceptual UML class diagrams of the existing system and the new system. As a conceptual model, a conceptual UML class diagram captures important concepts and relationships as classes and their associations within a business domain [1]. We assume that, since the newly-designed conceptual UML class diagram captures initially the software requirements of the new system, comparing it with the conceptual class diagram of the existing system should help identify the similarities and differences between the two system versions. The system analyst can then review certain aspects of the software requirements (1) whether they are still needed but missing from the new model, (2) whether they are not needed but are still included in the new model, (3) whether they should be added but are missing from the new model, and (4) whether they really are changes that should be made in the new model. In other words, the system analyst can verify that certain requirements that should be retained in the new system are not missing and those that are changed or newly introduced are included. The comparison is done by an algorithm called S-UMLDiff, which is an extension to the UMLDiff algorithm proposed by Xing [2]. S-UMLDiff compares two versions of the conceptual UML class diagrams to analyze name changes and structural changes between subsequent versions. Unlike UMLDiff, the algorithm is enhanced with the capability to analyze semantic similarity between names in the two versions of the diagram using WordNet [3]. This paper is an extended version of the initial report of the approach [4]. It describes, in

addition, the comparison tool and an evaluation of the S-UMLDiff algorithm using a real-world case study of a bank. The evaluation gives a satisfactory result compared to class diagrams comparison using the original UMLDiff.

Section 2 discusses background and related work. Section 3 describes the S-UMLDiff algorithm, with a supporting tool presented in Sect. 4. An evaluation of S-UMLDiff performance is given in Sect. 5 and the paper concludes in Sect. 6.

2 Background and Related Work

UML class diagrams [1] are useful in many stages of software system design. In the analysis stage, a class diagram can help the system analyst to understand the requirements of the problem domain and to identify important elements, data, functions, and relationships between elements. The class diagram in this stage is conceptual, having no detailed design for the implementation of the software. A conceptual class diagram contains (1) classes that represent concepts in the business domain, (2) attributes of a class, (3) methods of a class, (4) associations, aggregations, compositions, and generalizations which represent different relationships between classes, and (5) packages that represent groups of related classes and their relationships. Here other details such as data types of attributes, method parameters, and visibility of attributes and methods are not of concern.

Many algorithms to compare UML class diagrams have been proposed for different purposes. There is a possibility to apply one of them to our problem, but the chosen algorithm has to be applicable to the conceptual class diagrams which leave out a number of design details and, at the same time, it should be able to accommodate different kinds of changes that could occur in real-world software. Among the algorithms that we consider is the one by Girschick [5] which detects differences between several modifications of a design class diagram for tracking changes during the development process. Matching of design elements are based on generic graph matching techniques, and the elements that are compared are packages, classes, generalizations, attributes, associations, and operations and their parameters. Detected changes are add, delete, rename, move, clone, and modify property (e.g., visibility, data type, multiplicity, stereotype). A color-coding scheme is used to present different kinds of changes in different colors. The algorithm by Auxepaules et al. [6] is also a graph matching method but is used in an object-oriented modeling learning environment. The algorithm compares a student's diagram with an expert's diagram to give the student relevant feedbacks in modeling exercises. It uses the graph matching algorithm of Sorlin et al. [7] and a string matching algorithm of Giunchiglia et al. [8] which also uses WordNet to determine similarity of names. Detected differences are insert, omit, transfer, replace, modify property, merge, split, and cluster.

We select the UMLDiff algorithm of Xing [2] as a basis for our work since the algorithm considers all model elements in a conceptual class diagram and the kinds of differences that are detectable, i.e., add, remove, rename, move, and modify

property of elements, are sufficient for class diagrams at a conceptual level. UMLDiff relies on lexical similarity and structure similarity for recognizing the conceptually same model elements in the two compared versions of the class diagram. It does not take semantics of names into account, and thus it cannot recognize when a model element changes to a different but semantically similar name. In addition, it cannot recognize when a model element changes its type, e.g., an attribute is changed to a class.

3 Conceptual Class Diagrams Comparison

We present the S-UMLDiff (or Semantic-UMLDiff) algorithm to compare the conceptual class diagram of a new system with that of the existing system. The S-UMLDiff algorithm shares with UMLDiff in that it considers lexical similarity of names and structural similarity of model elements. Lexical similarity refers to string similarity of names while structural similarity refers to similarity of containment (i.e., a parent element contains another element as its child) and other relationships (i.e., an element has an association, aggregation, composition, and generalization relationship with another element). S-UMLDiff also enhances UMLDiff by considering semantic similarity of names and change of model element types.

To explain the S-UMLDiff algorithm, we use two versions of a conceptual class diagram of a bank, shown partially in Figs. 1 and 2, as a case study. The existing version has 28 classes in 4 packages and the new version has 37 classes in 4 packages. In the figures, only a few differences between the packages *Shipping* and *Transportation* are circled as an example. That is, the package named *Shipping* is changed to a semantically similar name *Transportation*, and so is the case between the classes named *CarType* and *VehicleType*. The class *CarSchedule* is renamed to *Schedule*. The method *cancelShipment* is removed whereas the class *FeeType* is added. The attribute *fee* has its element type changed and becomes the class *Fee*.

3.1 Overview of Difference Analysis

The analysis of class diagrams differences consists of:

1. Transform the two conceptual class diagrams into the XML Metadata Interchange (XMI) format. We use the ArgoUML modeling tool [9] to draw the conceptual models and obtain their representation in the XMI format.
2. Extract model elements. Model elements in the two diagrams (i.e., packages, classes, attributes, and methods) are extracted, together with their names and the relationships that they have with other elements and that the other elements have with them. The relationships include containment, association, aggregation, composition, and generalization.

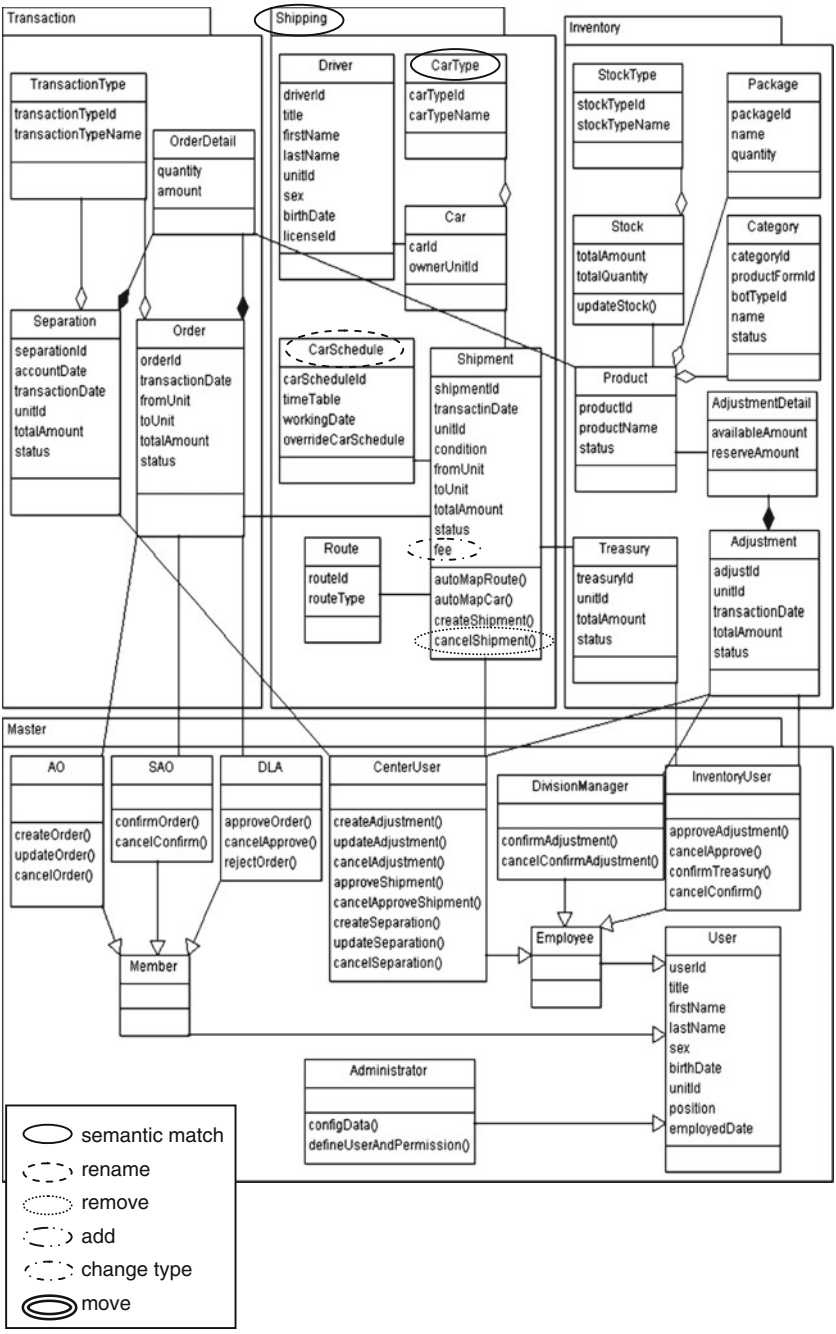


Fig. 1 Example of conceptual class diagram of existing system

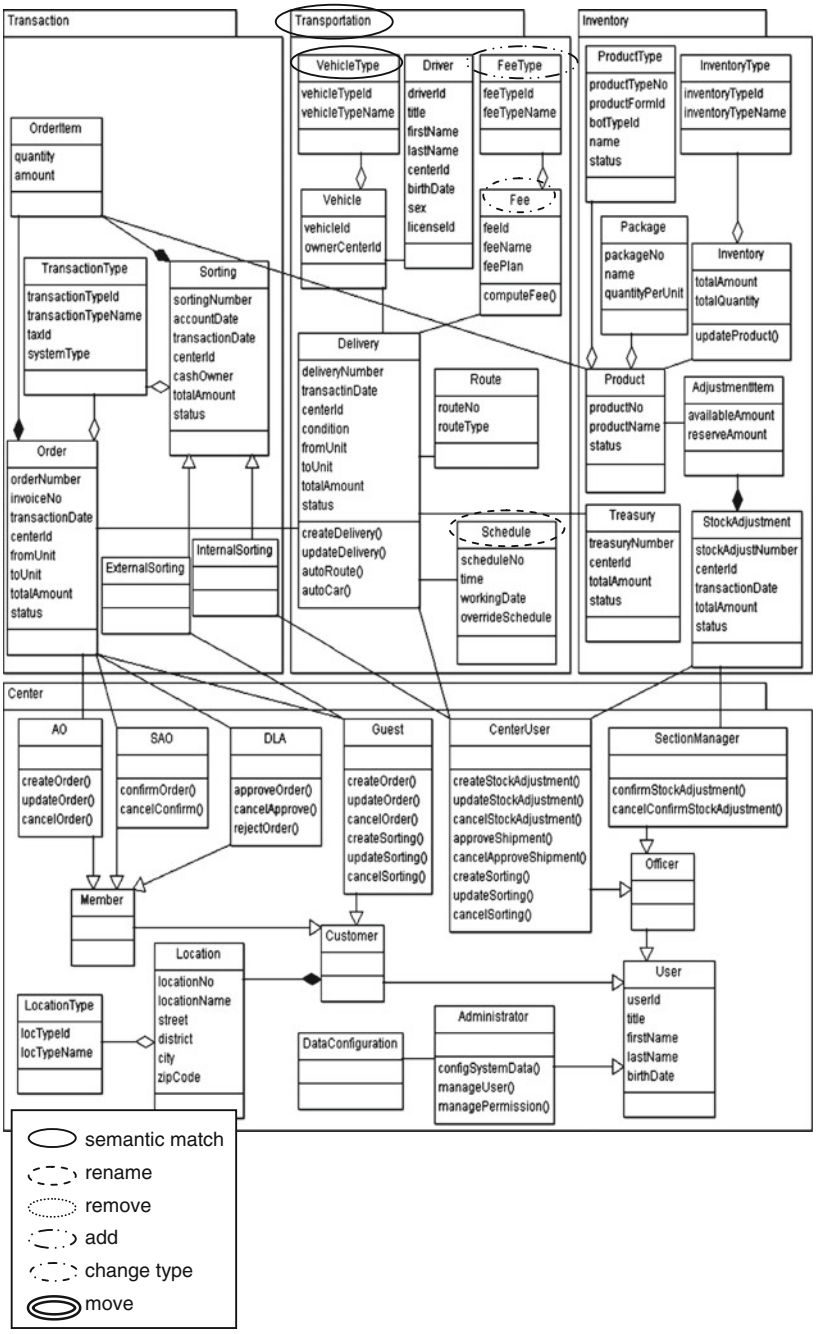


Fig. 2 Example of conceptual class diagram of new system

3. Build a directed graph $G(V, E)$ for each version of the class diagram, where the vertex set V contains the extracted model elements and the edge set E contains the relationships among them. An example of the vertices from the new diagram and relevant edges is shown in Table 1.
4. Map the two graphs $G_{existing}(V_{existing}, E_{existing})$ and $G_{new}(V_{new}, E_{new})$ by computing the intersection and margin sets between $(V_{existing}, V_{new})$ and $(E_{existing}, E_{new})$ to determine name similarity and structural similarity. That is, $(V_{existing} - V_{new})$ is computed for the removed model elements, $(V_{existing} \cap V_{new})$ for the mapped (i.e., matched, renamed, semantics-of-names-matched, moved, and element-type-changed) elements, $(V_{new} - V_{existing})$ for the added model elements, $(E_{existing} - E_{new})$ for the removed relationships, $(E_{existing} \cap E_{new})$ for the matched relationships, and $(E_{new} - E_{existing})$ for the added relationships. To be precise, the intersection and margin sets are computed by comparing the following in the two diagrams:

    4.1 Compare packages;
    4.2 Compare classes within the matched packages;
    4.3 Compare attributes within the matched classes;
    4.4 Compare methods within the matched classes;
    4.5 Compare removed class with added attribute, and;
    4.6 Compare removed attribute with added class.

For steps 4.1–4.4, S-UMLDiff identifies:

(a) Whether the model elements match by having the same name (i.e., *Match*);
(b) Whether the model elements with different names are the case of name change (i.e., *Rename* or *Semantic Match*) by determining the overall similarity including lexical, semantic, and structural similarities, and;
(c) Whether the model elements that are not identified as having a name change are moved (i.e., *Move*) by checking if there is a parent change. Otherwise it is the case of *Add* or *Remove*.

Table 1 Example of model elements and relationships from new diagram

Relationship	Source		Target	
	Name	Element type	Name	Element type
Contain	VirtualRoot[a]	Diagram	Transportation	Package
	Transportation	Package	Delivery	Class
	Delivery	Class	deliveryNumber	Attribute
	Delivery	Class	createDelivery	Method
Generalization	Customer	Class	Member	Class
Composition	Order	Class	OrderItem	Class
Aggregation	Order	Class	TransactionType	Class
Association	Delivery	Class	Fee	Class

[a]VirtualRoot is a default name for a conceptual class diagram

Table 2 Types of differences detected in new diagram

Difference type	Description
Match	Model element in new diagram is the same as model element in existing diagram
Semantic match	The name of model element in new diagram has semantic similarity to a name of model element in existing diagram
Rename	The name of model element in new diagram has lexical similarity to a name of model element in existing diagram
Change type	The name of model element in new diagram is the same as that of model element in existing diagram but has different element type
Move	The name of model element in new diagram is the same as that of model element in existing diagram but has different parent
Add	Model element is found in new diagram but not in existing diagram
Remove	Model element is found in existing diagram but not in new diagram

Steps 4.5–4.6 are added to S-UMLDiff to determine change of model element types (i.e., *Change Type*). Types of differences that will be reported by S-UMLDiff are shown in Table 2. Details of the comparison are discussed in the subsequent sections.

3.2 Name Similarity

To compute similarity of names, S-UMLDiff takes into account lexical similarity and semantic similarity. The model elements in the new conceptual class diagram may use different names for better modeling or due to change of concepts in the problem domain. S-UMLDiff first determines the semantic similarity *wScore* between the two words being compared, using the Wu-Palmer similarity measure that is implemented in the WordNet::Similarity package [3] where *wScore* is in [0, 1]. If the *wScore* is not less than a *Word Similarity Threshold* which is specified by the system analyst, the two words are considered similar by semantics. In the case that a string name is not a single word but a phrase (having dots, dashes, underscores and case switching as delimiters between words) and WordNet cannot determine similarity directly, we use a semantic similarity measure for phrases [10]. For phrases a and b comprising m and n words respectively, the phrase semantic similarity *pScore* is computed by

$$pScore(a,b) = \frac{\sum_{s=1}^{m} wpScore(a_s,b)}{m} \quad (1)$$

$$wpScore(a_s,b) = \max(wScore(a_s,b_1),\ldots,wScore(a_s,b_n)) \quad (2)$$

where

$wScore(a_s, b_n)$ = semantic similarity score between word s of phrase a and word n of phrase b by Wu-Palmer measure.

Similarly, if the *pScore* is not less than a *Phrase Similarity Threshold* which is specified by the system analyst, the two phrases are considered similar by semantics.

In the case that *wScore* (or *pScore*) is not greater than the corresponding threshold, name similarity is determined by lexical similarity using the Longest Common Subsequence (LCS) algorithm [11]. LCS is the longest subsequence (i.e., a set of characters that appear in left-to-right order but not necessarily consecutively) that appears in both string names a and b. The lexical similarity metric *lcsScore* is defined by

$$lcsScore(a, b) = \frac{2 * length(LCS(a, b))}{length(a) + length(b)}. \tag{3}$$

For example, using (1) and (2), we can compute the similarity score between the class names *CarType* in Fig. 1 and *VehicleType* in Fig. 2 by

$$\begin{aligned} &pScore(CarType, VehicleType) \\ &\quad = (wpScore(Car, VehicleType) + wpScore(Type, VehicleType))/2 \\ &\quad = (max(wScore(Car, Vehicle), wScore(Car, Type)) \\ &\qquad + max(wScore(Type, Vehicle), wScore(Type, Type)))/2 \\ &\quad = (max(0.9,\ 0.67)\ + max(0.7,\ 1))/2 = (0.9 + 1)/2 = 0.95 \end{aligned}$$

If the *Phrase Similarity Threshold* is 0.9, the two phrases are considered similar semantically. But if the threshold is higher than *pScore*, the two are not similar by semantics and S-UMLDiff will calculate their *lcsScore*.

Note that the part of speech of the two words has to be specified for WordNet:: Similarity to obtain *wScore* for them. Since names in the diagrams usually are nouns and verbs and if the two words can be both nouns and verbs, *wScore* for them will be an average of the similarity scores when they are nouns and when they are verbs. In addition, it is assumed that a method name starts with a verb and hence only verb will be used as the part of speech of the first word of a method name.

3.3 Structural Similarity

S-UMLDiff follows UMLDiff in checking for structural similarity by mapping the model elements of the same type and of the same name or similar names and then comparing names of the contained model elements. To compare structure of two packages, their classes are compared. To compare two classes, their attributes, methods, and relationships with other classes are compared. However, in the new

version of the diagram, there might be change of model element type such as the attribute *fee* of the class *Shipment* in Fig. 1 is changed to the class *Fee* in Fig. 2. S-UMLDiff therefore also checks names of class and attribute to see if it is the case of an attribute changing to a class or a class changing to an attribute, and not the case of removed and added model elements.

3.4 Overall Similarity

Computing overall similarity takes into account both names and structure of model elements in the two diagrams.

Let

MatchPoint = overall similarity score between model elements x and y
NamePoint = name similarity score between model elements x and y (see Sect. 3.2)
x = model element in the existing diagram
y = model element in the new diagram.

Adapted from UMLDiff, the overall similarity computation for each model element type is as follows.

3.4.1 Overall Similarity Between Packages

The following *MatchPoint* between two packages is calculated to determine if they match:

$$MatchPoint = \frac{NamePoint + ChildrenMatchCount}{NamePoint + (xChildrenCount + yChildrenCount - ChildrenMatchCount)} \quad (4)$$

where

ChildrenMatchCount = number of classes in package x whose names match those of classes in package y
xChildrenCount = number of classes in package x
yChildrenCount = number of classes in package y

3.4.2 Overall Similarity Between Classes

The following *MatchPoint* between two classes is calculated to determine if they match:

$$MatchPoint = \frac{NamePoint + ChildrenPoint + UsagePoint}{NamePoint + 3} \quad (5)$$

$$ChildrenPoint = \frac{ChildrenMatchCount}{xChildrenCount + yChildrenCount - ChildrenMatchCount} \quad (6)$$

where

ChildrenMatchCount = number of attributes in class *x* whose names match those of attributes in class *y* + number of methods in class *x* whose names match those of methods in class *y*
xChildrenCount = number of attributes and methods in class *x*
yChildrenCount = number of attributes and methods in class *y*
UsagePoint = similarity score between names of classes that have relationships with class *x* and names of classes that have relationships with class *y* (see Sect. 3.2).

3.4.3 Overall Similarity Between Attributes

The following *MatchPoint* between two attributes is calculated to determine if they match:

$$MatchPoint = \frac{ParentPoint * NamePoint}{ParentPoint * NamePoint + 2} \quad (7)$$

where

ParentPoint = similarity score between class name of attribute *x* and class name of attribute *y* (see Sect. 3.2).

3.4.4 Overall Similarity Between Methods

The following *MatchPoint* between two methods is calculated to determine if they match:

$$MatchPoint = \frac{ParentPoint * NamePoint}{ParentPoint * NamePoint + 2} \quad (8)$$

where

ParentPoint = similarity score between class name of method *x* and class name of method *y* (see Sect. 3.2).

3.4.5 Use of MatchPoint

To identify if the two model elements with different names match, i.e., there is a lexical or semantic change of name, the comparison, adapted from UMLDiff, is performed as in Fig. 3. Steps added by S-UMLDiff are shaded. First, semantic similarity of names of the two model elements is determined using *wScore* or *pScore*. As mentioned earlier, if the names are not considered as similar by semantics according to the respective *Word Similarity Threshold* or *Phrase Similarity Threshold*, lexical similarity is computed by *lcsScore*. After that, the overall similarity or *MatchPoint* is computed for the two model elements. In the same manner, the *Rename Threshold* is introduced to identify that the two model elements are a match as they are similar enough by name and structure even though there might be a change of name (i.e., *Semantic Match* or *Rename*). Otherwise, it is the case of *Not Rename* and the two are unmatched, i.e., they are different model

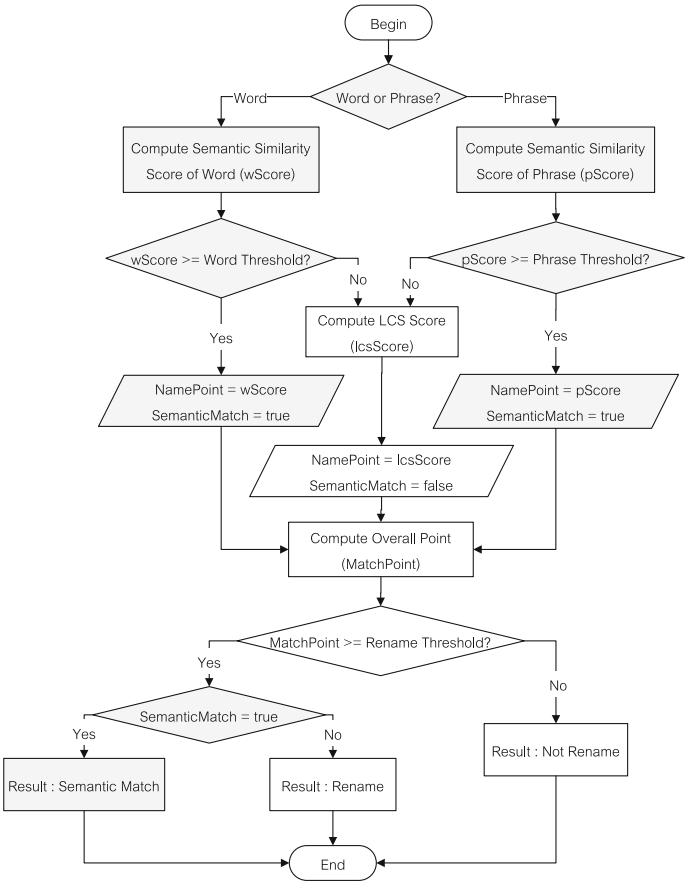


Fig. 3 Comparison to check if two model elements with different names match

Rename Threshold : 0.5 Semantic Word Similarity Threshold : 0.7

Semantic Phrase Similarity Threshold : 0.9

Original System	New System	Type	Match Point	Semantic Match Point
Virtual Root	Virtual Root	match	0.0	0.0
Transaction	Transaction	match	0.0	0.0
Shipping	Transportation	smatch	1.0	1.0
Inventory	Center	smatch	0.727273	0.727273
Order	Order	match	0.0	0.0
TransactionType	TransactionType	match	0.0	0.0
Route	Route	match	0.0	0.0
Driver	Driver	match	0.0	0.0
OrderDetail	OrderItem	smatch	1.0	1.0
Separation	Sorting	rename	0.823867	0.0
CarSchedule	Schedule	rename	0.570478	0.0
Car	Vehicle	smatch	0.669231	0.9

Fig. 4 Sample screenshot of comparison results

elements. For the unmatched elements, they will be checked further for the case of *Move*, *Add*, *Remove*, or *Change Type*.

4 Tool Implementation

A Java Web application is developed for system analysts to verify conceptual class diagram redesign. The inputs of this tool are the two versions of the conceptual class diagram in XMI format which can be exported from a UML modeling tool such as ArgoUML. The elements in XMI are extracted to build graphs to represent the two conceptual class diagrams and then the model elements in the two graphs are compared using the S-UMLDiff algorithm. In the comparison, the system analyst specifies relevant thresholds, i.e., *Word Similarity Threshold*, *Phrase Similarity Threshold*, and *Rename Threshold*. The tool reports matched and unmatched model elements with difference types as exemplified in Fig. 4.

5 Evaluation

Evaluation of S-UMLDiff is by measuring precision and recall [12] of diagram differences that are reported by the tool against the actual differences identified by system analysts with 9–12 years of experience.

Table 3 S-UMLDiff performance

Pair#	Precision	Recall	Pair#	Precision	Recall
1	0.91	0.91	2	1	1
3	1	1	4	0.74	0.89
5	0.95	0.95	6	0.97	0.97
7	0.73	0.92	8	0.73	0.92
9	0.82	0.9	10	0.97	0.97
11	0.74	0.84			
Average	0.87	0.93			

In the evaluation, we use 11 pairs of conceptual class diagrams, where pair# 11 is the case of Figs. 1 and 2. For each pair, the design of the two diagram versions and difference identification are done by different system analysts. We adjust the thresholds so that they give the best measurement results, i.e., *Word Similarity Threshold*, *Phrase Similarity Threshold*, and *Rename Threshold* are 0.7, 0.9, and 0.5 respectively.

As shown in Table 3, S-UMLDiff gives the average precision of 0.87 and average recall of 0.93 which are very satisfactory. In addition, S-UMLDiff is compared with the original UMLDiff. For example, for the pair# 11 of Figs. 1 and 2, precision and recall by UMLDiff are 0.28 and 0.53. The much lower performance of UMLDiff compared to S-UMLDiff is due to the large number of mismatches between the difference types reported by UMLDiff and those by the system analysts. This results from the fact that UMLDiff does not recognize semantic change of name and change of model element type. For example, by UMLDiff, the package *Shipping* in Fig. 1 is *Remove* and the package *Transportation* in Fig. 2 is *Add*. Thus the contained elements such as the class *Route* is considered *Move* from *Shipping* to *Transportation*. Likewise, the class *CarType* is also considered *Remove* and the class *VehicleType* is *Add*. On the contrary, S-UMLDiff and the system analyst consider *Shipping* and *Transportation* as *Semantic Match*. Thus the contained elements like the class *Route* can find its match, and the classes *CarType* and *VehicleType* are considered *Semantic Match*.

6 Conclusion

The S-UMLDiff algorithm can help identify the differences between the two versions of the conceptual diagram and therefore the system analyst can use the differences report to verify if the redesign is correct and complete. The analysis of semantic similarity of names and change of model element types enhances the algorithm and gives a more informative report of changes in the new version. However, the specified thresholds affect how S-UMLDiff classify changes. For future work, S-UMLDiff and the supporting tool can be improved by visualizing the comparison results and even supporting change impact analysis.

References

1. Object Management Group Unified modeling language [online]. Available: http://www.omg.org/spec/UML/2.4.1/
2. Xing Z (2008) Supporting object-oriented evolutionary development by design evolution analysis. Doctoral dissertation, Department of Computing Science, University of Alberta, Canada
3. Pedersen T (2005) WordNet:: similarity [online]. Available: http://wn-similarity.sourceforge.net/
4. Saisim P, Senivongse T (2014) Verifying software change requirements through conceptual class diagrams comparison. 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, pp 139–144
5. Girschick M (2006) Difference detection and visualization in UML class diagrams, TU Darmstadt, Germany, Technical report TUD-CS-2006-5
6. Auxepaules L, Py D, Lemeunier T (2008) A diagnosis method that matches class diagrams in a learning environment for object-oriented modeling. In: Proceedings of 8th IEEE international conference on advanced learning technologies, ICALT 2008, 1–5 July 2008, Santander, pp 26–30
7. Sorlin S, Solnon C, Jolion J-M (2007) A generic graph distance measure based on multivalent matchings. Applied Graph Theory Comp Vis Pattern Recogn Stud Comput Intell 52:151–181
8. Giunchiglia F, Yatskevich M, Shvaiko P (2007) Semantic matching: algorithms and implementation. J Data Semant 9:1–38
9. Tigris.org ArgoUML [online]. Available: http://argouml.tigris.org/
10. Gad W, Kamel M (2009) PH-SSBM: phrase semantic similarity based model for document clustering. In: Proceedings of 2nd international symposium on knowledge acquisition and modeling, KAM 2009, vol 2, 30 Nov–1 Dec 2009, Wuhan, pp 197–200
11. Stein C Longest common subsequence [online]. Available: http://www.columbia.edu/~cs2035/courses/csor4231.F11/lcs.pdf
12. Baeza-Yates R, Ribeiro-Neto B (2011) Modern information retrieval, 2nd edn. Addison Wesley, Essex

Transreal Limits and Elementary Functions

Tiago S. dos Reis and James A.D.W. Anderson

Abstract We extend all elementary functions from the real to the transreal domain so that they are defined on division by zero. Our method applies to a much wider class of functions so may be of general interest.

Keywords Transreal continuity · Transreal elementary functions · Transreal limits · Transreal numbers · Transreal sequences · Transreal series

1 Introduction

The set of real numbers, $\mathbb{R}$, is extended to the set of transreal numbers, $\mathbb{R}^T$, by the addition of three, definite, non-finite numbers: negative infinity, $-\infty = -1/0$; positive infinity, $\infty = 1/0$; and nullity, $\Phi = 0/0$. These three, non-finite numbers are called *strictly transreal* numbers. Nullity was introduced in [1]. Transreal arithmetic is axiomatised and proved consistent by machine proof in [4]. A construction of the transreal numbers from the real numbers is given in [5]. That construction provides a human proof of the consistency of transreal arithmetic. Transreal limits are given in [3]. Transreal derivatives are given in [6].

T.S. dos Reis (✉)
Federal Institute of Education, Science and Technology of Rio de Janeiro,
Rio de Janeiro 27215-350, Brazil
e-mail: tiago.reis@ifrj.edu.br

T.S. dos Reis
Program of History of Science, Technique and Epistemology;
Federal University of Rio de Janeiro, Rio de Janeiro 21941-916, Brazil

J.A.D.W. Anderson
School of Systems Engineering, University of Reading,
Whiteknights, Reading RG6 6AY, England
e-mail: j.anderson@reading.ac.uk

H.K. Kim et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-94-017-7236-5_15

Here we consider transreal limits and extend elementary functions of real numbers to transreal numbers. In [3] we define a topology on $\mathbb{R}^T$ which extends the topology of real numbers. Because of this we can define, in a rigorous way, limits of sequences, series and functions and continuity of functions on $\mathbb{R}^T$ so that wherever real numbers occur in real limits, they occur identically in transreal limits and wherever infinities occur as symbols in extended real limits, they occur identically in transreal limits but as definite numbers. This means that transreal topology agrees with the usual, real topology.

We extend every elementary function to the transreal domain. Some of these functions, such as the exponential and trigonometric functions, were defined, in [2], at strictly transreal numbers, motived by power series. However the convergence of series in [2] is taken in an intuitive way, without any rigorous definition of the limit of sequences and series on $\mathbb{R}^T$. That development also lacks a topology on $\mathbb{R}^T$. Here we also use power series but first we extended some results about series to $\mathbb{R}^T$ using its well-defined topology [3].

2 Topological Foundations

In [3] we define a topology, limit of sequences and limit and continuity of functions on $\mathbb{R}^T$. Below we summarise this content.

Definition 1 A set is open on $\mathbb{R}^T$ if and only if it is composed of arbitrarily many unions of finitely many intersections of the following four kinds of interval:

i. (a, b) where $a, b \in \mathbb{R}$,
ii. $[-\infty, b)$ where $b \in \mathbb{R}$,
iii. $(a, \infty]$ where $a \in \mathbb{R}$ and
iv. $\{\Phi\}$.

The reader can verify that these open sets do, in fact, make a topology on $\mathbb{R}^T$.

Proposition 1 *$\mathbb{R}^T$ is a Hausdorff, disconnected, separable, compact and completely metrisable space. For definitions of these terms and proofs see* [3] *and a forthcoming paper* [7].

Notice that Φ is the unique isolated point of $\mathbb{R}^T$.

Proposition 2 *The topology on $\mathbb{R}$, induced by the topology of $\mathbb{R}^T$, is the usual topology of $\mathbb{R}$. That is if $A \subset \mathbb{R}^T$ is open on $\mathbb{R}^T$ then $A \cap \mathbb{R}$ is open (in the usual sense) on $\mathbb{R}$ and if $A \subset \mathbb{R}$ is open (in the usual sense) on $\mathbb{R}$ then A is open on $\mathbb{R}^T$.*

Remember the definition of a sequence. A sequence in $\mathbb{R}^T$ is a function $x: \mathbb{N} \to \mathbb{R}^T$. We customarily write x_n in place of $x(n)$ and write $(x_n)_{n\in\mathbb{N}}$ in place of $x: \mathbb{N} \to \mathbb{R}^T$. We use the usual definition for the convergence of a sequence in a topological space. That is a sequence, $(x_n)_{n\in\mathbb{N}} \subset \mathbb{R}^T$, converges to $x \in \mathbb{R}^T$ if and

only if for each neighbourhood, $V \subset \mathbb{R}^T$ of x, there is $n_V \in \mathbb{N}$ such that $x_n \in V$ for all $n \geq n_V$. Notice that since $\mathbb{R}^T$ is a Hausdorff space, the limit of a sequence, when it exists, is unique.

Remark 1 Let $(x_n)_{n\in\mathbb{N}} \subset \mathbb{R}$ and let $L \in \mathbb{R}$. Notice that $\lim_{n\to\infty} x_n = L$ in $\mathbb{R}^T$ if and only if $\lim_{n\to\infty} x_n = L$ in the usual sense in $\mathbb{R}$. Furthermore, $(x_n)_{n\in\mathbb{N}}$ diverges, in the usual sense, to negative infinity if and only if $\lim_{n\to\infty} x_n = -\infty$ in $\mathbb{R}^T$. Similarly $(x_n)_{n\in\mathbb{N}}$ diverges, in the usual sense, to infinity if and only if $\lim_{n\to\infty} x_n = \infty$ in $\mathbb{R}^T$.

Remark 2 Let $(x_n)_{n\in\mathbb{N}} \subset \mathbb{R}^T$. Notice that $\lim_{n\to\infty} x_n = \Phi$ if and only if there is $k \in \mathbb{N}$ such that $x_n = \Phi$ for all $n \geq k$.

Proposition 3 *Every monotone sequence of transreal numbers is convergent.*

Proposition 4 (Transreal version of the Bolzano-Weierstrass Theorem) *Every sequence of transreal numbers has a convergent subsequence.*

3 Transreal Series

In this section we extend some results on series to the transreal domain.

Definition 2 Let $(x_n)_{n\in\mathbb{N}} \subset \mathbb{R}^T$. For each $n \in \mathbb{N}$, we define

$$s_n := \sum_{i=1}^{n} x_i = x_1 + \cdots + x_n.$$

The sequence $(s_n)_{n\in\mathbb{N}}$ is called a series and is denoted by $\sum x_n$, each s_n is called a partial sum of $\sum x_n$ and x_n is called the n-th term of $\sum x_n$. We say that $\sum x_n$ converges or is convergent if and only if there is the $\lim_{n\to\infty} s_n$. Otherwise, $\sum x_n$ diverges or is divergent. When $\sum x_n$ is convergent we denote

$$\sum_{n=1}^{\infty} x_n := \lim_{n\to\infty} s_n.$$

Definition 3 Let $(x_n)_{n\in\mathbb{N}} \subset \mathbb{R}^T$. We say that a series $\sum x_n$ converges absolutely or is absolutely convergent if and only if $\sum |x_n|$ is convergent.

Customarily, in a calculus course, the first example of a convergent series is the geometric series, $\sum r^n$. Recall that $\sum r^n$ converges in $\mathbb{R}$ if and only if $r \in (-1, 1)$ and, in this case, $\sum_{n=1}^{\infty} r^n = \frac{r}{1-r}$.

Example 1 Let $r \in \mathbb{R}^T$. The series $\sum r^n$ converges in $\mathbb{R}^T$ if and only if $r \in \{-\infty, \Phi\} \cup (-1, \infty]$. Indeed, if $r \in \{-\infty, \Phi\}$ then $\sum_{n=1}^{\infty} r^n = \Phi$, if $r \in (-1, 1)$ then $\sum_{n=1}^{\infty} r^n = \frac{r}{1-r}$, if $r \in [1, \infty]$ then $\sum_{n=1}^{\infty} r^n = \infty$ and if $r \in (-\infty, -1]$ then $s_{2n-1} < 0$ and $s_{2n} \geq 0$ for all $n \in \mathbb{N}$ whence $\sum r^n$ diverges.

Remark 3 In the real domain we know that if $\sum x_n$ is convergent then $\lim_{n\to\infty} x_n = 0$. In the transreal domain there is no similar result. That is, we can have a convergent series, $\sum x_n$, such that $(x_n)_{n\in\mathbb{N}}$ converges to anywhere in $\mathbb{R}^T$. Indeed, if $a \in \{\Phi, -\infty, \infty\}$ and $x_n = a$ for all $n \in \mathbb{N}$ then $\sum x_n$ is convergent (to a) and $\lim_{n\to\infty} x_n = a$. If $a \in \mathbb{R}$ is positive, let $x_n = a - \frac{1}{n^2}$ for each $n \in \mathbb{N}$. Remember that $\sum \frac{1}{n^2}$ converges increasingly to $\frac{\pi^2}{6}$ whence $\sum_{i=1}^{n} \frac{1}{i^2} < \frac{\pi^2}{6}$ for all $n \in \mathbb{N}$. Given an arbitrary positive $M \in \mathbb{R}$, there is $n_M \in \mathbb{N}$ such that $n_M a > M + \frac{\pi^2}{6}$. Hence if $n \geq n_M$ then $na \geq n_M a > M + \frac{\pi^2}{6} > M + \sum_{i=1}^{n} \frac{1}{i^2}$ whence $\sum_{i=1}^{n} x_n = na - \sum_{i=1}^{n} \frac{1}{i^2} > M$, that is, $\sum_{i=1}^{n} x_n \in (M, \infty]$. Thus $\sum x_n$ is convergent (to ∞) and $\lim_{n\to\infty} x_n = a$. If $a \in \mathbb{R}$ is negative, let $x_n = a + \frac{1}{n^2}$ for each $n \in \mathbb{N}$ then, by a similar argument, we have that $\sum x_n$ is convergent (to $-\infty$) and $\lim_{n\to\infty} x_n = a$.

Remark 4 In the real domain, we know that a series $\sum x_n$ is convergent if and only if for each positive $\varepsilon \in \mathbb{R}$ there is $n_\varepsilon \in \mathbb{N}$ such that

$$\left| \sum_{i=m}^{n} x_i \right| < \varepsilon \tag{1}$$

whenever $n \geq m \geq n_\varepsilon$. This is nothing more than the application of the Cauchy criterion. As $\mathbb{R}$ is a complete metric space, a sequence is convergent if and only if it is Cauchy. The inequality (1) is, in fact, $|s_n - s_{m-1}| < \varepsilon$. We can naturally enunciate that result in the transreal domain by using its complete metric, denoted as d. A series $\sum x_n$ is convergent if and only if for each positive $\varepsilon \in \mathbb{R}$ there is $n_\varepsilon \in \mathbb{N}$ such that $d(s_n, s_{m-1}) < \varepsilon$ whenever $n \geq m \geq n_\varepsilon$. This is correct but is not as helpful as in the real domain because, unlike in real domain, the fact that $d(s_n, s_{m-1})$ is less than an arbitrary, positive, real number is not equivalent to the fact that $\left|\sum_{i=m}^{n} x_i\right|$ is less than some positive, real number.

We can enunciate the Cauchy criterion, in the real domain, in another way. A series $\sum x_n$ is convergent if and only if for each neighbourhood V of zero there is $n_V \in \mathbb{N}$ such that $\sum_{i=m}^{n} x_i \in V$ whenever $n \geq m \geq n_V$. In this way we might imagine that there is a similar result for the transreal domain. For example, let $x_n = n^2$ for all $n \in \mathbb{N}$. The series $\sum x_n$ is convergent (to ∞) and for each neighbourhood V of infinity there is $n_V \in \mathbb{N}$ such that $\sum_{i=m}^{n} x_i \in V$ whenever $n \geq m \geq n_V$. However this is not always the case. Let $x_n = 1$ for all $n \in \mathbb{N}$. The series $\sum x_n$ is convergent (to ∞), but for all $n \in \mathbb{N}$, $\sum_{i=n}^{n} x_i = 1$ does not belong to neighbourhood of infinity $(M, \infty]$ where $M > 1$.

We might wish to enunciate, in the transreal domain, something like: if the series $\sum x_n$ is convergent then either $\sum_{i=m}^{n} x_i$ belongs to a neighbourhood of zero for n and m large enough or $\sum_{i=m}^{n} x_i$ belongs to a neighbourhood of infinity for n and m large enough or else $(x_n)_{n\in\mathbb{N}}$ is constant. But this is still not true. Let $x_n = \frac{1}{n}$. The series $\sum x_n$ is convergent (to ∞) but does not satisfy any of the three conditions above. In fact, let there be arbitrary, positive $\varepsilon, M \in \mathbb{R}$ such that $\varepsilon < M$. There is

$n \in \mathbb{N}$ such that $\sum_{i=n}^{n} x_i = \frac{1}{n} \in (-\varepsilon, \varepsilon)$ and there is $m > n$ such that $\sum_{i=n}^{m} x_i \in (M, \infty]$.

Remark 5 In the real domain, we know that if $\sum x_n$ is absolutely convergent then $\sum x_n$ is convergent. In the transreal domain this is not true. Indeed, according to Example 1, the series $\sum(-1)^n$ is absolutely convergent ($\sum|(-1)^n|$ *converges to* ∞), but it is not convergent.

Proposition 5 *In the transreal domain every series of non-negative terms is convergent. In other words, if* $(x_n)_{n\in\mathbb{N}} \in \mathbb{R}^T$ *and* $x_n \not< 0$ *for all* $n \in \mathbb{N}$ *then* $\sum x_n$ *is convergent.*

Proof As $x_n \not< 0$ for all $n \in \mathbb{N}$, for some index n_0, the sequence of partial sums of $\sum x_{n+n_0}$, $(s_{n+n_0})_{n\in\mathbb{N}}$ is monotone. By Proposition 3, $\sum x_{n+n_0}$ is convergent whence $\sum x_n$ is convergent.

Obviously if $(x_n)_{n\in\mathbb{N}} \in \mathbb{R}^T$ and $x_n \not> 0$ for all $n \in \mathbb{N}$ then $\sum x_n$ is convergent too. That is, every series of non-positive terms is convergent. □

Remark 6 Notice that, since in the transreal domain every series of non-negative terms is convergent, the Comparison Test, the Ratio Test and the Root Test are not appropriate for transreal series.

Definition 4 Let $(c_n)_{n\in\mathbb{N}} \in \mathbb{R}^T$. The series $c_0 + \sum c_n x^n$ is called a power series. A power series defines a function at the values x for which it is convergent. That is, if A is the set of all transreal numbers x such that $c_0 + \sum c_n x^n$ converges then

$$\begin{aligned} f\colon A &\to \mathbb{R}^T \\ x &\mapsto c_0 + \textstyle\sum_{n=1}^{\infty} c_n x^n \end{aligned}$$

is a well-defined function.

4 Elementary Functions

Recall that a real, elementary function is defined in the following way. Every polynomial, root, exponential, logarithmic, trigonometric and inverse trigonometric function is an elementary function; any finite composition of elementary functions is an elementary function; and any finite combination, using the four arithmetical operations, between elementary functions is an elementary function. We wish to extend the real elementary functions to the transreal numbers.

There are some choices which need to be made. The power series of some elementary functions have a value, at a strictly transreal number, that is different to the limit of the function when the argument tends to that strictly transreal number. Let us illustrate this with the exponential function. It is known that $e^x =$

$1+\sum_{n=1}^{\infty}\frac{x^n}{n!}$ for all $x\in\mathbb{R}$. We have that $\lim_{x\to-\infty}\mathrm{e}^x=0$ but $1+\sum_{n=1}^{\infty}\frac{(-\infty)^n}{n!}=\Phi$. How should we define $\mathrm{e}^{-\infty}$? By $\mathrm{e}^{-\infty}=0$ or $\mathrm{e}^{-\infty}=\Phi$? Geometrically speaking, it is intuitive to us that, in the graph of the exponential function, when x "hits" minus infinity, e^x "hits" zero. In this way we choose to define $\mathrm{e}^{-\infty}=0$. This has the advantage that the exponential function becomes continuous at $-\infty$. At ∞ there is no trouble because $\lim_{x\to\infty}\mathrm{e}^x=\infty=1+\sum_{n=1}^{\infty}\frac{\infty^n}{n!}$. Nullity, Φ, is an isolated point, whence it is nonsense to speak of the limit at Φ; because of this we choose to define e^{Φ} by way of the power series. We have that $1+\sum_{n=1}^{\infty}\frac{\Phi^n}{n!}=\Phi$ whence we define $\mathrm{e}^{\Phi}=\Phi$. However we do not always define a function by way of limits or power series. For example the function $\frac{\sin(x)}{x}$, calculated exactly at zero is $\frac{\sin(0)}{0}=\frac{0}{0}$ and $\frac{0}{0}=\Phi$ but $\lim_{x\to0}\frac{\sin(x)}{x}=1$. Again we have a choice to make. The transreal numbers were defined in order to allow division by zero. It seems to us counter-productive to ignore the fact that, in $\mathbb{R}^T$, $\frac{0}{0}$ is well-defined. Hence we choose to define $\frac{\sin(0)}{0}=\Phi$. Of course the function $\frac{\sin(x)}{x}$ becomes discontinuous at 0 but this is the price to pay for having transreal arithmetic.

In summary we adopt the following procedure to extend a function from the real domain to the transreal domain. If the expression of the function is lexically well-defined, at a transreal number, then we define the function by simply applying its expression at that transreal number. If the function f is not lexically well-defined at a transreal number, x_0, but there is a limit, $\lim_{x\to x_0}f(x)$, then we choose to define the function at x_0 by $\lim_{x\to x_0}f(x)$. Otherwise we choose to define the function by way of its power series if it converges. And if, nevertheless, its power series does not converge, we keep the function undefined.

Firstly we recall some results about the limit and continuity of functions obtained in [3]. Recall that if X is a topological space then $x_0\in A\subset X$ is a limit point of A if and only if for every neighbourhood V of x_0 it follows that $V\cap(A\backslash\{x_0\})=\emptyset$. The set of all limit points of A is denoted as A'. We use the usual definition of the limit of functions in a topological space. That is if A is a subset of $\mathbb{R}^T$, $f:A\to\mathbb{R}^T$ is a function, x_0 is a limit point of A and L is a transreal number, we say that $\lim_{x\to x_0}f(x)=L$ if and only if, for each neighbourhood V of L, there is a neighbourhood U of x_0 such that $f(A\cap U\backslash\{x_0\})\subset V$.

Remark 7 Notice that given $x_0,L\in\mathbb{R}$, $\lim_{x\to x_0}f(x)=L$, in the transreal sense, if and only if $\lim_{x\to x_0}f(x)=L$, in the real sense. The same can be said about $\lim_{x\to x_0}f(x)=-\infty$, $\lim_{x\to x_0}f(x)=\infty$, $\lim_{x\to-\infty}f(x)=L$, $\lim_{x\to-\infty}f(x)=-\infty$, $\lim_{x\to-\infty}f(x)=\infty$, $\lim_{x\to\infty}f(x)=L$, $\lim_{x\to\infty}f(x)=-\infty$ and $\lim_{x\to\infty}f(x)=\infty$.

Remark 8 Let $x_0\in\mathbb{R}^T$, notice that $\lim_{x\to x_0}f(x)=\Phi$ if and only if there is a neighbourhood U of x_0 such that $f(x)=\Phi$ for all $x\in U\backslash\{x_0\}$.

We use the usual definition of continuity in a topological space. That is if $A\subset\mathbb{R}^T$, $f:A\to\mathbb{R}^T$ is a function and $x_0\in A$, we say that f is continuous in x_0 if and only if, for each neighbourhood V of $f(x_0)$, there is a neighbourhood U of x_0 such

that $f(A \cap U) \subset V$. We say that f is continuous in A if and only if f is continuous in x for all $x \in A$.

Remark 9 Of course if x_0 is a limit point of A then f is continuous in x_0 if and only if $\lim_{x \to x_0} f(x) = f(x_0)$.

Remark 10 Notice that given $x_0 \in \mathbb{R}$, f is continuous in x_0, in the transreal sense, if and only if f is continuous in x_0, in the real sense.

Remark 11 Notice that if Φ belongs to the domain of f then f is continuous in Φ.

4.1 Polynomial Functions

A function, f, is a real, polynomial function if and only if there is $n \in \mathbb{N}$ and $a_0, \ldots, a_n \in \mathbb{R}$ such that $f(x) = a_n x^n + \cdots + a_1 x + a_0$ for all $x \in \mathbb{R}$. Of course, f can be written $f(x) = a_m x^m + \cdots + a_1 x + a_0$ for any $m > n$ provided that $a_{n+1} = \cdots = a_m = 0$. And if $a_k = 0$ for some $k \in \{1, \ldots, n\}$ then f can also be written $f(x) = a_n x^n + \cdots + a_{k+1} x^{k+1} + a_{k-1} x^{k-1} + \cdots + a_1 x + a_0$. This multiplicity, in the representation of f, is a problem in the transreal domain. In the real domain, $0 \times x^k = 0$ for all real x but $0 \times x^k = 0$ does not hold for all transreal x. We have $0 \times (-\infty)^k = 0 \times \infty^k = 0 \times \Phi^k = \Phi \neq 0$. In order to avoid this problem we define that a function f is a real, polynomial function if and only if there is $n, k \in \mathbb{N}$, $n_1, \ldots, n_k \in \{1, \ldots, n-1\}$ and $a_0, a_{n_1} \ldots, a_{n_k}, a_n \in \mathbb{R}$ such that $a_{n_1} \ldots, a_{n_k}, a_n$ are different from zero and

$$\begin{aligned} f\colon \mathbb{R} &\to \mathbb{R} \\ x &\mapsto a_n x^n + a_{n_k} x^{n_k} + \cdots + a_{n_1} x^{n_1} + a_0. \end{aligned} \tag{2}$$

By following our procedure, as every arithmetical operation is well-defined in transreal numbers, we extend the function f to $\mathbb{R}^T$ naturally in the following way.

Definition 5 A function f is a transreal, polynomial function if and only if there is $n, k \in \mathbb{N}$, $n_1, \ldots, n_k \in \{1, \ldots, n-1\}$ and $a_0, a_{n_1} \ldots, a_{n_k}, a_n \in \mathbb{R}$ such that $a_{n_1} \ldots, a_{n_k}, a_n$ are different from zero and

$$\begin{aligned} f\colon \mathbb{R}^T &\to \mathbb{R}^T \\ x &\mapsto a_n x^n + a_{n_k} x^{n_k} + \cdots + a_{n_1} x^{n_1} + a_0 \end{aligned}.$$

Remark 12 For every non-constant, transreal, polynomial function, f, we have that $f(\Phi) = \Phi$.

Notice that, in the transreal domain, some polynomial functions are not continuous.

Example 2 Let $n \in \mathbb{N}$ such that $n \geq 2$ and let non-zero numbers $a_0, \ldots, a_n \in \mathbb{R}$ such that $a_n a_{n-1} < 0$ and $f(x) = a_n x^n + \cdots + a_1 x + a_0$ then $f(\infty) = \Phi$ but $\lim_{x \to \infty} f(x) = sgn(a_n) \times \infty$, whence f is not continuous at ∞.

Example 3 Let $n \in \mathbb{N}$ such that $n \geq 2$ and let non-zero numbers $a_0, \ldots, a_n \in \mathbb{R}$ such that $a_n a_{n-1} > 0$ and $f(x) = a_n x^n + \cdots + a_1 x + a_0$ then $f(-\infty) = \Phi$ but $\lim_{x \to -\infty} f(x) = sgn(a_n) \times (-\infty)^n$, whence f is not continuous at $-\infty$.

If the reader wishes to extend the function in (2) continuously to $\mathbb{R}^T$ then boundary conditions are needed at $-\infty$ and ∞:

$$f: \mathbb{R}^T \to \mathbb{R}^T$$
$$x \mapsto \begin{cases} a_n x^n & , \quad \text{if } x \in \{-\infty, \infty\} \\ a_n x^n + a_{n_k} x^{n_k} + \cdots + a_{n_1} x^{n_1} + a_0 & , \quad \text{otherwise} \end{cases}.$$

Notice that just by using Definition 5, some particular, transreal, polynomial functions become continuous—as exemplified next.

Example 4 Let $a, b \in \mathbb{R}$ such that $a \neq 0$ and $f(x) = ax + b$ for all $x \in \mathbb{R}^T$. Notice that $\lim_{x \to -\infty} f(x) = f(-\infty)$ and $\lim_{x \to \infty} f(x) = f(\infty)$ whence f is continuous in $\mathbb{R}^T$.

4.2 Exponential Functions

A function, f, is a real, exponential function if and only if there is a positive $a \in \mathbb{R}$ such that $a \neq 1$ and

$$\begin{aligned} f: \mathbb{R} &\to \mathbb{R} \\ x &\mapsto a^x. \end{aligned} \tag{3}$$

The natural, exponential function is $f(x) = \mathrm{e}^x$ where e is Euler's number. We know that every exponential function $f(x) = a^x$ can be written $f(x) = \mathrm{e}^{\ln(a)x}$ where $\ln a$ is the natural logarithm of a.

Let

$$\begin{aligned} f: \mathbb{R} &\to \mathbb{R} \\ x &\mapsto \mathrm{e}^x. \end{aligned}$$

By following our procedure, as the function f is not defined by way of arithmetical operations, we define $f(-\infty) := \lim_{x \to -\infty} \mathrm{e}^x$ and $f(\infty) := \lim_{x \to \infty} f(x)$

whence $f(-\infty) = 0$ and $f(\infty) = \infty$. Furthermore, we know that $e^x = 1 + \sum_{n=1}^{\infty} \frac{x^n}{n!}$ for all $x \in \mathbb{R}$. Following our method, in the absence of limit, we define $f(\Phi) = 1 + \sum_{n=1}^{\infty} \frac{\Phi^n}{n!}$, whence $f(\Phi) = \Phi$. Thus we extend f to $\mathbb{R}^T$ by defining

$$f: \mathbb{R}^T \to \mathbb{R}^T$$
$$x \mapsto \begin{cases} 0 & , \text{ if } x = -\infty \\ 1 + \sum_{n=1}^{\infty} \frac{x^n}{n!} & , \text{ otherwise} \end{cases}.$$

Now we extend the function in (3) to $\mathbb{R}^T$ in the following way.

Definition 6 A function, f, is a transreal, exponential function if and only if there is a positive $a \in \mathbb{R}$ such that $a \neq 1$ and

$$f: \mathbb{R}^T \to \mathbb{R}^T$$
$$x \mapsto a^x = e^{\ln(a)x}.$$

Proposition 6 *(Properties of exponential functions) Let f be the transreal, exponential function, $f(x) = a^x$. We have that, for all $x, y \in \mathbb{R}^T$:*

1. *f is continuous.*
2. *The image of f is $[0, \infty] \cup \{\Phi\}$.*
3. *f is strictly monotone.*
4. *f is injective.*
5. $a^{x+y} = a^x a^y$.
6. $a^{x-y} = \frac{a^x}{a^y}$.

Remark 13 In [6] we define the transreal derivative in the following way. Let $A \subset \mathbb{R}^T$, $f : A \to \mathbb{R}^T$ and $x_0 \in A$. If $x_0 \in \mathbb{R} \cap A'$ and f is differentiable at x_0, in the real sense, then the transreal derivative of f at x_0 is the real derivative $f'(x_0)$. If $x_0 \in \{-\infty, \infty\} \cap D'$ (where D denotes the set of points in A at which f is differentiable, in the real sense) then the transreal derivative of f at x_0 is $f'(x_0) = \lim_{x \to x_0} f'(x)$ if this limit exist. And if $x_0 \notin A'$ then the transreal derivative of f at x_0 is $f'(x_0) := \Phi$.

Remark 14 According to Remark 13, we have that $\frac{d}{dx} a^x = \ln(a) a^x$ for all $x \in \mathbb{R}^T$. In particular, $\frac{d}{dx} e^x = e^x$ for all $x \in \mathbb{R}^T$.

4.3 Logarithmic Functions

A function, f, is a real, logarithmic function if and only if there is a positive $a \in \mathbb{R}$ such that $a \neq 1$ and

$$\begin{aligned} f\colon (0,\infty) &\to \mathbb{R} \\ x &\mapsto \log_a(x) \end{aligned} \tag{4}$$

where $\log_a(x)$ denotes the unique, real number y, for which $a^y = x$.

By following our procedure, as the transreal exponential is injective and has image $[0,\infty]\cup\{\Phi\}$, the function f in (4) is lexically well-defined in $[0,\infty]\cup\{\Phi\}$. Because of this we extend the function f to $\mathbb{R}^T$ in the following way.

Definition 7 A function, f, is a transreal, logarithmic function if and only if there is a positive $a \in \mathbb{R}$ such that $a \neq 1$ and

$$\begin{aligned} f\colon [0,\infty]\cup\{\Phi\} &\to \mathbb{R}^T \\ x &\mapsto \log_a(x) \end{aligned}$$

where $\log_a(x)$ denotes the unique, transreal number y, for which $a^y = x$.

Like in the real domain, we call $\log_e(x)$ the transreal, natural logarithm of x and denote it as $\ln(x)$. In particular we have that $\ln(0) = -\infty$, $\ln(\infty) = \infty$ and $\ln(\Phi) = \Phi$.

Proposition 7 (Properties of logarithmic functions) *Let f be the transreal, logarithmic function, $f(x) = \log_a(x)$. We have that for all $x, y \in \mathbb{R}^T$:*

1. *f is continuous.*
2. *The image of f is $\mathbb{R}^T$.*
3. *f is strictly monotone.*
4. *f is injective.*
5. $\log_a(x \times y) = \log_a(x) + \log_a(y)$.
6. $\log_a\left(\frac{x}{y}\right) = \log_a(x) - \log_a(y)$.

Remark 15 We have already said that the derivative of a function f at ∞, if $\infty \in D'$, is $f'(\infty) := \lim_{x\to\infty} f'(x)$ if this limit exists. Fortunately this definition has sense for all $x_0 \in D'$ so we define, for all $x_0 \in D'$, $f'(x_0) := \lim_{x\to x_0} f'(x)$, if this limit exists.

Remark 16 According to Remark 15, $\frac{d}{dx}\log_a(x)|_0 = \lim_{x\to 0} \frac{1}{\ln(a)} \times \frac{1}{x} = \lim_{x\to 0^+} \frac{1}{\ln(a)} \times \frac{1}{x} = \frac{1}{\ln(a)} \times \infty = \begin{cases} -\infty, & \text{if } 0<a<1 \\ \infty \ , & \text{if } a > 1 \end{cases}$. Therefore we have that $\frac{d}{dx}\log_a(x) = \frac{1}{\ln(a)} \times \frac{1}{x}$ for all $x \in \mathbb{R}^T$. In particular, $\frac{d}{dx}\ln(x) = \frac{1}{x}$ for all $x \in \mathbb{R}^T$.

Remark 17 The Definition 6 defines powers x^y when $x \in \mathbb{R}$ is positive and $x \neq 1$. The transreal logarithm allows us to define x^y for all non-negative transreal x. Let $x \in [0,\infty]\cup\{\Phi\}$ and $y \in \mathbb{R}^T$, we define $x^y := \mathrm{e}^{\ln(x)y}$. In particular,

(a) $0^y = \infty$ for all $y < 0$.
(b) $0^0 = \Phi$.
(c) $0^y = 0$ for all $y > 0$.
(d) $0^\Phi = \Phi$.
(e) $1^y = 1$ for all $y \in \mathbb{R}$.
(f) $1^y = \Phi$ for all $y \in \{-\infty, \infty, \Phi\}$.
(g) $\infty^y = 0$ for all $y < 0$.
(h) $\infty^0 = \Phi$.
(i) $\infty^y = \infty$ for all $y > 0$.
(j) $\infty^\Phi = \Phi$.
(k) $\Phi^\Phi = \Phi$.

4.4 Trigonometric Functions

We know that the real, trigonometric functions can be defined as follows.

(a)
$$\sin : \mathbb{R} \to \mathbb{R}$$
$$x \mapsto \sin(x) = \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{(2n-1)!} x^{2n-1},$$

(b)
$$\cos : \mathbb{R} \to \mathbb{R}$$
$$x \mapsto \cos(x) = 1 + \sum_{n=1}^{\infty} \frac{(-1)^{n}}{(2n)!} x^{2n},$$

(c)
$$\tan : \mathbb{R} \backslash \left\{ \frac{\pi}{2} + k\pi;\ k \in \mathbb{Z} \right\} \to \mathbb{R}$$
$$x \mapsto \tan(x) = \frac{\sin(x)}{\cos(x)},$$

(d)
$$\csc : \mathbb{R} \backslash \{k\pi;\ k \in \mathbb{Z}\} \to \mathbb{R}$$
$$x \mapsto \csc(x) = \frac{1}{\sin(x)},$$

(e)
$$\sec : \mathbb{R} \backslash \left\{ \frac{\pi}{2} + k\pi;\ k \in \mathbb{Z} \right\} \to \mathbb{R}$$
$$x \mapsto \sec(x) = \frac{1}{\cos(x)} \quad \text{and}$$

(f)
$$\cot : \mathbb{R} \backslash \{k\pi;\ k \in \mathbb{Z}\} \to \mathbb{R}$$
$$x \mapsto \cot(x) = \frac{\cos(x)}{\sin(x)}.$$

If f is a trigonometric function, that is, if f is one of the six functions above then there are no limits $\lim_{x \to -\infty} f(x)$ and $\lim_{x \to \infty} f(x)$. So by following our procedure,

we define $\sin(-\infty) := \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{(2n-1)!}(-\infty)^{2n-1}$, $\cos(-\infty) := 1 + \sum_{n=1}^{\infty} \frac{(-1)^n}{(2n)!}(-\infty)^{2n}$, $\tan(-\infty) := \frac{\sin(-\infty)}{\cos(-\infty)}$, $\csc(-\infty) := \frac{1}{\sin(-\infty)}$, $\sec(-\infty) := \frac{1}{\cos(-\infty)}$ and $\cot(-\infty) := \frac{\cos(-\infty)}{\sin(-\infty)}$. And similarly for ∞ and Φ. Hence $\sin(-\infty) = \cos(-\infty) = \tan(-\infty) = \csc(-\infty) = \sec(-\infty) = \cot(-\infty) = \sin(\infty) = \cos(\infty) = \tan(\infty) = \csc(\infty) = \sec(\infty) = \cot(\infty) = \sin(\Phi) = \cos(\Phi) = \tan(\Phi) = \csc(\Phi) = \sec(\Phi) = \cot(\Phi) = \Phi$.

Notice that tan, csc, sec and cot are lexically well-defined at $\frac{\pi}{2} + k\pi$ and $k\pi$ for all $k \in \mathbb{Z}$. In this way we extend the trigonometric functions to $\mathbb{R}^T$ in the following way.

Definition 8 A function is a transreal, trigonometric function if and only if it is one of the functions below.

(a) $\sin : \mathbb{R}^T \to \mathbb{R}^T$

$$x \mapsto \sin(x) = \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{(2n-1)!} x^{2n-1},$$

(b) $\cos : \mathbb{R}^T \to \mathbb{R}^T$

$$x \mapsto \cos(x) = 1 + \sum_{n=1}^{\infty} \frac{(-1)^n}{(2n)!} x^{2n},$$

(c) $\tan : \mathbb{R}^T \to \mathbb{R}^T$

$$x \mapsto \tan(x) = \frac{\sin(x)}{\cos(x)},$$

(d) $\csc : \mathbb{R}^T \to \mathbb{R}^T$

$$x \mapsto \csc(x) = \frac{1}{\sin(x)},$$

(e) $\sec : \mathbb{R}^T \to \mathbb{R}^T$

$$x \mapsto \sec(x) = \frac{1}{\cos(x)} \quad \text{and}$$

(f) $\cot : \mathbb{R}^T \to \mathbb{R}^T$

$$x \mapsto \cot(x) = \frac{\cos(x)}{\sin(x)}.$$

Proposition 8 (Properties of trigonometric functions)

1. *Every transreal, trigonometric function is discontinuous.*
2. *The image of* sin *and* cos *is* $[-1, 1] \cup \{\Phi\}$.
3. *The image of* tan *and* cot *is* $\mathbb{R}^T$.
4. *The image of* csc *and* sec *is* $(-\infty, -1] \cup [1, \infty] \cup \{\Phi\}$.
5. *Every transreal, trigonometric function has period* 2π.

6. sin *restricted to* $[-\pi, \pi] \cup \{\Phi\}$ *is injective.*
7. cos *restricted to* $[0, 2\pi] \cup \{\Phi\}$ *is injective.*
8. tan *restricted to* $\left[-\frac{\pi}{2}, \frac{\pi}{2}\right] \cup \{\Phi\}$ *is injective.*
9. csc *restricted to* $\left[-\frac{\pi}{2}, \frac{\pi}{2}\right] \cup \{\Phi\}$ *is injective.*
10. sec *restricted to* $[0, \pi] \cup \{\Phi\}$ *is injective.*
11. cot *restricted to* $[0, \pi] \cup \{\Phi\}$ *is injective.*
12. $\sin^2(x) + \cos^2(x) = 1^x$ *for all* $x \in \mathbb{R}^T$.
13. $\tan^2(x) + 1^x = \sec^2(x)$ *for all* $x \in \mathbb{R}^T$.
14. $\cot^2(x) + 1^x = \csc^2(x)$ *for all* $x \in \mathbb{R}^T$.

Remark 18 We have already defined the derivative of a function f at x_0, if $x_0 \in D'$, by $f'(x_0) := \lim_{x \to x_0} f'(x)$ if this limit exists. Regrettably if f is a trigonometric function and $x_0 \in \{-\infty, \infty\}$ then there is no limit $\lim_{x \to x_0} f(x)$. If we appeal to lexical application of the derivative, we can adopt the following. If $x_0 \in D'$ and there is no limit $\lim_{x \to x_0} f'(x)$ and the expression of f' is lexically well-defined at x_0 then we define the derivative of f at x_0 as $f'(x_0)$.

Remark 19 For example $\frac{d}{dx}\sin(x) = \cos(x)$ for all $x \in \mathbb{R}$, $-\infty$ and ∞ are limit points of the set where sin is differentiable, in the real sense, and there are no $\lim_{x \to -\infty} \cos(x)$ and $\lim_{x \to \infty} \cos(x)$. As $\cos(x)$ is lexically well-defined for $x = -\infty$ and $x = \infty$, according to Remark 18, $\frac{d}{dx}\sin(x)|_{-\infty} = \cos(-\infty)$ and $\frac{d}{dx}\sin(x)|_{\infty} = \cos(\infty)$. And since Φ is not a limit point of the domain of sin, according to Remark 13, $\frac{d}{dx}\sin(x)|_{\Phi} = \Phi = \cos(\Phi)$. Because of this, $\frac{d}{dx}\sin(x) = \cos(x)$ for all $x \in \mathbb{R}^T$.

As another example $\frac{d}{dx}\tan x = \sec^2(x)$ for all $x \in \mathbb{R} \setminus \left\{\frac{\pi}{2} + k\pi;\ k \in \mathbb{Z}\right\}$. As, for every $k \in \mathbb{Z}$, $\frac{\pi}{2} + k\pi$ is a limit point of the set where tan is differentiable, in the real sense, and there is $\lim_{x \to \frac{\pi}{2} + k\pi} \sec^2(x)$, according to Remark 15, we define $\frac{d}{dx}\tan(x)|_{\frac{\pi}{2} + k\pi} = \lim_{x \to \frac{\pi}{2} + k\pi} \sec^2(x) = \infty = \sec^2\left(\frac{\pi}{2} + k\pi\right)$. Since $-\infty$ and ∞ are limit points of the set where tan is differentiable, in the real sense, and there are no limits $\lim_{x \to -\infty} \sec^2(x)$ and $\lim_{x \to \infty} \sec^2(x)$ but $\sec^2(x)$ is lexically well-defined for $x = -\infty$ and $x = \infty$ so, according to Remark 18, $\frac{d}{dx}\tan(x)|_{-\infty} = \sec^2(-\infty)$ and $\frac{d}{dx}\tan(x)|_{\infty} = \sec^2(\infty)$. And since Φ is not a limit point of the domain of tan, according to Remark 13, $\frac{d}{dx}\tan(x)|_{\Phi} = \Phi = \sec^2(\Phi)$. Because of this, $\frac{d}{dx}\tan(x) = \sec^2(x)$ for all $x \in \mathbb{R}^T$.

Generally, we have, for all $x \in \mathbb{R}^T$:

1. $\frac{d}{dx}\sin(x) = \cos(x)$.
2. $\frac{d}{dx}\cos(x) = -\cos(x)$.
3. $\frac{d}{dx}\tan(x) = \sec^2(x)$.
4. $\frac{d}{dx}\csc(x) = -\csc(x)\cot(x)$.

5. $\frac{d}{dx}\sec(x) = \sec(x)\tan(x)$.
6. $\frac{d}{dx}\cot(x) = -\csc^2(x)$.

At this point the reader can already define the inverse trigonometric functions and deduce their properties.

5 The Limit and Value of a Function at a Point

We tabulate both the values and the limits of the above functions so that we can see whether a function is extended to $\mathbb{R}^T$ continuously or not.

$a_2x^2 + a_1x$ where $a_2 > 0$ and $a_1 > 0$			
x_0	$a_2x^2 + a_1x$	$\lim_{x\to x_0}(a_2x^2 + a_1x)$	Continuity at x_0
$-\infty$	Φ	∞	Not continuous
∞	∞	∞	Continuous
Φ	Φ	Not applicable	Continuous

$a_2x^2 + a_1x$ where $a_2 > 0$ and $a_1 < 0$			
x_0	$a_2x^2 + a_1x$	$\lim_{x\to x_0}(a_2x^2 + a_1x)$	Continuity at x_0
$-\infty$	∞	∞	Continuous
∞	Φ	∞	Not continuous
Φ	Φ	Not applicable	Continuous

e^x			
x_0	e^{x_0}	$\lim_{x\to x_0} e^x$	Continuity at x_0
$-\infty$	0	0	Continuous
∞	∞	∞	Continuous
Φ	Φ	Not applicable	Continuous

$\ln(x)$			
x_0	$\ln(x_0)$	$\lim_{x\to x_0}\ln(x)$	Continuity at x_0
0	$-\infty$	$-\infty$	Continuous
∞	∞	∞	Continuous
Φ	Φ	Not applicable	Continuous

$\sin(x)$			
x_0	$\sin(x_0)$	$\lim_{x\to x_0}\sin(x)$	Continuity at x_0
$-\infty$	Φ	None	Not continuous
∞	Φ	None	Not continuous
Φ	Φ	Not applicable	Continuous

$\cos(x)$			
x_0	$\cos(x_0)$	$\lim_{x\to x_0}\cos(x)$	Continuity at x_0
$-\infty$	Φ	None	Not continuous

(continued)

$\cos(x)$			
∞	Φ	None	Not continuous
Φ	Φ	Not applicable	Continuous

$\tan(x)$			
x_0	$\tan(x_0)$	$\lim_{x\to x_0}\tan(x)$	Continuity at x_0
$-\infty$	Φ	None	Not continuous
∞	Φ	None	Not continuous
Φ	Φ	Not applicable	Continuous
$\frac{\pi}{2}+(2k+1)\pi$	$-\infty$	None	Not continuous
$\frac{\pi}{2}+2k\pi$	∞	None	Not continuous

$\tan(x)$ restricted to $\left[-\frac{\pi}{2},\frac{\pi}{2}\right]$			
x_0	$\tan(x_0)$	$\lim_{x\to x_0}\tan(x)$	Continuity at x_0
$-\frac{\pi}{2}$	$-\infty$	$-\infty$	Continuous
$\frac{\pi}{2}$	∞	∞	Continuous

$\csc(x)$			
x_0	$\csc(x_0)$	$\lim_{x\to x_0}\csc(x)$	Continuity at x_0
$-\infty$	Φ	None	Not continuous
∞	Φ	None	Not continuous
Φ	Φ	Not applicable	Continuous
$k\pi$	∞	None	Not continuous

$\csc(x)$ restricted to $\left[-\frac{\pi}{2},\frac{\pi}{2}\right]$			
x_0	$\csc(x_0)$	$\lim_{x\to x_0}\csc(x)$	Continuity at x_0
0	∞	None	Not continuous

$\sec(x)$			
x_0	$\sec(x_0)$	$\lim_{x\to x_0}\sec(x)$	Continuity at x_0
$-\infty$	Φ	None	Not continuous
∞	Φ	None	Not continuous
Φ	Φ	Not applicable	Continuous
$\frac{\pi}{2}+k\pi$	∞	None	Not continuous

$\sec(x)$ restricted to $[0,\pi]$			
x_0	$\sec(x_0)$	$\lim_{x\to x_0}\sec(x)$	Continuity at x_0
$\frac{\pi}{2}$	∞	None	Not continuous

$\cot(x)$			
x_0	$\cot(x_0)$	$\lim_{x\to x_0}\cot(x)$	Continuity at x_0
$-\infty$	Φ	None	Not continuous
∞	Φ	None	Not continuous
Φ	Φ	Not applicable	Continuous
$2k\pi$	∞	None	Not continuous
$(2k+1)\pi$	$-\infty$	None	Not continuous

$\cot(x)$ restricted to $[0,\pi]$			
x_0	$\cot(x_0)$	$\lim_{x\to x_0}\cot(x)$	Continuity at x_0

(continued)

$\cot(x)$ restricted to $[0, \pi]$			
0	∞	∞	Continuous
π	$-\infty$	$-\infty$	Continuous

$\frac{\sin(x)}{x}$			
x_0	$\frac{\sin(x_0)}{x_0}$	$\lim_{x \to x_0} \frac{\sin(x)}{x}$	Continuity at x_0
0	Φ	1	Not continuous
$-\infty$	Φ	0	Not continuous
∞	Φ	0	Not continuous
Φ	Φ	Not applicable	Continuous

6 Conclusion

We extend all elementary functions from the real domain to the transreal domain so that they are defined on division by zero. If the expression of the function is lexically well-defined, at a transreal number, then we define the function by simply applying its expression at that transreal number. If the function f is not lexically well-defined at a transreal number, x_0, but there is a limit, $\lim_{x \to x_0} f(x)$, then we choose to define the function at x_0 by $\lim_{x \to x_0} f(x)$. Otherwise we choose to define the function by way of its power series if it converges. And if, nevertheless, its power series does not converge, we keep the function undefined. This method for extending functions from the real domain to the transreal domain clearly works for a much wider class of functions so it may be of general interest.

Acknowledgments The authors would like to thank the members of Transmathematica for many helpful discussions. The first author's research was financially supported, in part, by the Federal Institute of Education, Science and Technology of Rio de Janeiro, campus Volta Redonda and by the Program of History of Science, Technique and Epistemology, Federal University of Rio de Janeiro. The second author's research was financially supported, in part, by the School of Systems Engineering at the University of Reading and by a Research Travel Grant from the University of Reading Research Endowment Trust Fund (RETF).

References

1. Anderson JADW (1997) Representing geometrical knowledge. Phil Trans Roy Soc Lond Series B 352(1358):1129–1139
2. Anderson JADW (2007) Perspex machine ix: transreal analysis. In: Lateki LJ, Mount DM, Wu AY (eds) Vision geometry XV of proceedings of SPIE, vol 6499, pp J1–J12
3. Anderson JADW, dos Reis TS (2004) Transreal limits expose category errors in ieee 754 floating-point arithmetic and in mathematics. In: Lecture notes in engineering and computer science: proceedings of the world congress on engineering and computer science 2014, WCECS 2014, vol 1, 22–24 Oct 2014, San Francisco, pp 86–91

4. Anderson JADW, Völker N, Adams AA (2007) Perspex machine viii: axioms of transreal arithmetic. In Lateki LJ, Mount DM, Wu AY (eds) Vision geometry XV of proceedings of SPIE, vol 6499, pp 2.1–2.12
5. dos Reis TS, Anderson JADW (2014) Construction of the transcomplex numbers from the complex numbers. In: Lecture notes in engineering and computer science: proceedings of the world congress on engineering and computer science 2014, WCECS 2014, vol 1, 22–24 Oct 2014, San Francisco, pp 97–102
6. dos Reis TS, Anderson JADW (2014) Transdifferential and transintegral calculus. In: lecture notes in engineering and computer science: proceedings of the world congress on engineering and computer science 2014, WCECS 2014, vol 1, 22–24 Oct 2014, San Francisco, pp 92–96
7. dos Reis TS, Anderson JADW (2015) Transreal calculus. IAENG Int J Appl Math 45(1):51–63

Transreal Logical Space of All Propositions

Walter Gomide, Tiago S. dos Reis and James A.D.W. Anderson

Abstract Transreal numbers provide a total semantics containing classical truth values, dialetheaic, fuzzy and gap values. A paraconsistent Sheffer Stroke generalises all classical logics to a paraconsistent form. We introduce logical spaces of all possible worlds and all propositions. We operate on a proposition, in all possible worlds, at the same time. We define logical transformations, possibility and necessity relations, in proposition space, and give a criterion to determine whether a proposition is classical. We show that proofs, based on the conditional, infer gaps only from gaps and that negative and positive infinity operate as bottom and top values.

Keywords All possible worlds · Logical spaces · Multi-valued logics · Paraconsistent logics · Transreal numbers · Total semantics

W. Gomide (✉)
Philosophy Department, Institute of Humanities and Social Sciences,
Federal University of Mato Grosso, Cuiabá 78060-900, Brazil
e-mail: waltergomide@yahoo.com

W. Gomide · T.S. dos Reis
Program of History of Science, Technique and Epistemology,
Federal University of Rio de Janeiro, Rio de Janeiro 21941-916, Brazil
e-mail: tiago.reis@ifrj.edu.br

T.S. dos Reis
Federal Institute of Education, Science and Technology of Rio de Janeiro,
Rio de Janeiro 27215-350, Brazil

J.A.D.W. Anderson
School of Systems Engineering, University of Reading, Whiteknights,
Reading RG6 6AY, England, UK
e-mail: j.anderson@reading.ac.uk

H.K. Kim et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-94-017-7236-5_16

1 Introduction

We rehearse our development of total semantics [8] by considering paraconsistent logics. These were explicitly introduced in the second half of the twentieth century as non-classical logics that can reason about inconsistent axioms [12, 21]. In a classical logic, inconsistent axioms *explode*, allowing any theorem to be proved in a trivial way [11, 18]. Paraconsistent logics do not explode, they allow only limited conclusions to be drawn from inconsistent axioms. Some admit *dialetheias*, that is propositions that are both False and True [20], and some admit Gap values with no degree of falsity or truthfulness [23]. Gap values are usually treated absorptively so that any logical combination with a Gap produces a Gap as result. This behaviour is consistent with one reading [17] of Frege's principle of compositionallity so that a compound proposition lacks reference if any component of it lacks reference. It should be added that paraconsistent logics are also capable of classical reasoning so they provide a robust generalisation of classical logic. This makes them interesting both from a theoretical and a practical perspective [10, 21, 28].

Paraconsistent logics are often formalised in advanced mathematics [12, 23, 24]. We take the simpler approach of expressing them arithmetically. We use transreal arithmetic, which is a generalisation of real arithmetic. Transreal arithmetic was originally developed [1, 2] from a subset of the algorithms used in the arithmetic of fractions. It has been axiomatised and a machine proof of consistency has been given [7]. A human proof of consistency is given in [13]. The algorithms of transreal arithmetic are explained, particularly clearly, in a tutorial in [4].

In Sect. 2 we generalise all classical logics to a paraconsistent form by expressing the Sheffer Stroke [9] in transreal arithmetic. In Sect. 3 we introduce a logical space in which we operate on a proposition in all possible worlds at the same time. The idea of logical space is inspired by Wittgenstein's conception that the world's logical form is given by a picture that is a "configuration of objects." See [26, 27] Sects. 2, 3 and, especially 3.4. Thus, just as physical objects are arranged in physical space, so logical objects are arranged in a "logical space" [15]. Wittgenstein did not define precisely his notion of logical space. However, by following the intuitive idea that the elements of this space are propositions and the interactions between them are connectives, we establish a logical space as a well-defined mathematical structure, something like a vector space, where the propositions are "vectors" and the connectives are "vector" transformations. In this space we give a mathematical sense to the notion of logical transformation and of possibility and necessity relations. The mathematical treatment ends by establishing a criterion for determining whether a proposition is or is not classical in a given possible world.

2 Paraconsistent Logic

We use the entire set of transreal numbers to supply the semantic values, that is the truth values, of paraconsistent logic. In this section we exploit the intuition that a conclusion departs no further from being equally false and true than the most extreme of its antecedents. This is sufficient to make a logic non-explosive. We define a paraconsistent version of the Sheffer Stroke via the minimum value of its antecedents. This paraconsistent operator may be used to generalise all classical logics to a paraconsistent form.

2.1 Truth Values

The transreal numbers are just the real numbers augmented with three non-finite numbers: negative infinity ($-\infty$), positive infinity (∞) and nullity (Φ). Nullity is absorptive over the elementary arithmetical operations so that when it is involved in a sum, difference, product or quotient, the result is nullity. However nullity is not universally absorptive, it may be an element of arbitrary mappings. Nullity is the only unordered number in transreal arithmetic [6, 7]. Nullity's absorptive properties make it a good candidate for a Gap value that has no degree of falsity or truthfulness [8, 25]. We define that negative infinity is classical False and positive infinity is classical True. This has the merit that we have now used up all of the non-finite, transreal numbers, leaving all of the real numbers to convey dialetheaic degrees of falsity and truthfulness. Here we use arithmetical negation (unary subtraction) to model logical negation. Alternative encodings are discussed in [8, 16].

Returning now to our paraconsistent logic, we define that the real numbers encode degrees of both falsity and truthfulness. The negative real numbers are more False than True, the positive, real numbers are more True than False, zero is equally False and True. We relate the degree of falsity and truthfulness monotonically to the number modelling the truth value so that negative infinity is entirely False, that is classically False, and positive infinity is entirely True, that is classically True.

2.2 Sheffer Stroke

It is known that the truth functional (Boolean) operators for logical negation (not, $\neg$), logical conjunction (and, &), and logical disjunction (or, $\vee$) are functionally complete [9] (See entry "Sheffer Stroke"), [22] (p. 29) so that any truth functional operators can be derived from these three. In fact it is known that the sets $\{\neg, \&\}$ and $\{\neg, \vee\}$ are each functionally complete but it serves our purpose better to consider the wider set of operators $\{\neg, \&, \vee\}$. We use the transreal minimum and maximum functions to define paraconsistent versions of the classical negation,

conjunction and disjunction operators. We use negative infinity ($-\infty$) to model classical False (F) and positive infinity (∞) to model classical True (T). We use nullity (Φ) to model the logical Gap value (G). Note that only the real numbers model dialetheaic truth values. The three non-finite numbers each model a single truth value. We then prove that the paraconsistent operators contain the classical ones. With a little extra work we prove that the paraconsistent operators are well defined for all transreal arguments when we assume that the finite, truth values are arranged monotonically with the real numbers that model them. We then define a paraconsistent version of the Sheffer Stroke ($|$). There are three, well known, identities that relate the classical Sheffer Stroke to classical negation, conjunction and disjunction. We show that these identities hold when we substitute the paraconsistent Sheffer Stroke and the paraconsistent negation, conjunction and disjunction. Thus we prove that the paraconsistent operators are defined everywhere and are consistent with their classical counterparts.

We begin by defining the binary, transreal, minimum and maximum functions so that the minimum of two transreal numbers is the least, ordered one of them or else is nullity. Similarly the maximum of two transreal numbers is the greatest, ordered one of them or else is nullity. These definitions rely on the three transreal relations less-than, equal-to, greater-than as axiomatised in [7], explicated in [5] and corrected in [14]. It is sufficient for the reader to know that: nullity is the uniquely unordered, transreal number so it is the only transreal number that compares not-less-than, not-equal-to and not-greater than any other distinct number; negative infinity is the least, ordered, transreal number; positive infinity is the greatest, ordered, transreal number.

Definition 1 Transreal minimum,

$$\min(a,b) = \begin{cases} a: & a<b \\ a: & a=b \\ a: & b=\Phi \\ b: & b<a \\ b: & a=\Phi \end{cases}.$$

Definition 2 Transreal maximum,

$$\max(a,b) = \begin{cases} a: & a>b \\ a: & a=b \\ a: & b=\Phi \\ b: & b>a \\ b: & a=\Phi \end{cases}.$$

The minimum and maximum functions, just defined, treat nullity non-absorptively but we chose to treat the logical Gap value absorptively.

Definition 3 Paraconsistent conjunction,

$$a \,\&\, b = \begin{cases} \Phi : & a = \Phi \text{ or } b = \Phi \\ \min(a, b) : & \text{otherwise} \end{cases}$$

Definition 4 Paraconsistent disjunction,

$$a \vee b = \begin{cases} \Phi : & a = \Phi \text{ or } b = \Phi \\ \max(a, b) : & \text{otherwise} \end{cases}$$

We now define the paraconsistent, logical negation as transarithmetical negation.

Definition 5 Paraconsistent negation, $\neg a = -a$.

Transreal arithmetic has $-0 = 0$, $-\Phi = \Phi$ and in all other cases, the negation is distinct so that $-a \neq a$.

The Sheffer Stroke (|) may be defined as an infix operator but we follow the more modern practice of taking it as a post-fix operator so that no bracketing is needed. This leads to shorter and clearer formulas.

Definition 6 Paraconsistent Sheffer Stroke, $ab| = \neg(a \,\&\, b)$, with all symbols read paraconsistently.

We now prove that the paraconsistent negation, conjunction and disjunction contain their classical counterparts and that the paraconsistent operators are well defined for all transreal arguments.

Theorem 1 *Paraconsistent negation contains classical negation.*

Proof Classical negation has $\neg \mathrm{F} = \mathrm{T}$ and $\neg \mathrm{T} = \mathrm{F}$. Equivalently paraconsistent negation has $\neg(-\infty) = -(-\infty) = \infty$ and $\neg\infty = -\infty$. □

Theorem 2 *Paraconsistent conjunction contains classical conjunction.*

Proof Classical conjunction has F & F = F; F & T = F; T & F = F; T & T = T. Equivalently paraconsistent conjunction has $-\infty \,\&\, -\infty = \min(-\infty, -\infty) = -\infty$; $-\infty \,\&\, \infty = \min(-\infty, \infty) = -\infty$; $\infty \,\&\, -\infty = \min(\infty, -\infty) = -\infty$; $\infty \,\&\, \infty = \min(\infty, \infty) = \infty$. □

Theorem 3 *Paraconsistent disjunction contains classical disjunction.*

Proof Classical disjunction has $\mathrm{F} \vee \mathrm{F} = \mathrm{F}$; $\mathrm{F} \vee \mathrm{T} = \mathrm{T}$; $\mathrm{T} \vee \mathrm{F} = \mathrm{T}$; $\mathrm{T} \vee \mathrm{T} = \mathrm{T}$. Equivalently paraconsistent disjunction has $-\infty \vee -\infty = \max(-\infty, -\infty) = -\infty$; $-\infty \vee \infty = \max(-\infty, \infty) = \infty$; $\infty \vee -\infty = \max(\infty, -\infty) = \infty$; $\infty \vee \infty = \max(\infty, \infty) = \infty$. □

Theorem 4 *Paraconsistent negation, conjunction, and disjunction are well defined for all transreal arguments.*

Proof Paraconsistent negation, conjunction, and disjunction are defined for all transreal arguments. It remains only to show that these operators are monotonic.

Firstly nullity is absorptive in these operators so that if any argument is nullity the result is nullity. Nullity is disjoint from all other transreal numbers because it is the uniquely isolated point of transreal space [6], therefore nullity results are disjoint from all other transreal results and cannot contradict them. Secondly the preceding three theorems show that the paraconsistent operators are well defined at the boundaries $-\infty$ and ∞ but, by definition, the non-nullity, paraconsistent, truth values are monotonic so the operators just defined are monotonic for all transreal t in the range $-\infty \leq t \leq \infty$. This completes the proof for all transreal arguments.

We now derive the paraconsistent negation, conjunction and disjunction from formulas involving the paraconsistent Sheffer Stroke. This proves that the paraconsistent Sheffer Stroke is functionally complete both for classical truth values and for the paraconsistent truth values defined here. □

Theorem 5 *$pp| = \neg p$ for all transreal p.*

Proof $pp| = \neg(p \,\&\, p) = \neg p$, with all symbols read paraconsistently. □

Theorem 6 *$pq|pq|| = p \,\&\, q$ for all transreal p, q.*

Proof $pq|pq|| = (\neg(p \,\&\, q))(\neg(p \,\&\, q))| = \neg(\neg(p \,\&\, q)) = p \,\&\, q$, with all symbols read paraconsistently. □

Theorem 7 *$pp|qq|| = p \vee q$ for all transreal p, q.*

Proof $pp|qq|| = (\neg p)\,(\neg q)\,| = \neg((\neg p) \,\&\, (\neg q)) = p \vee q$ by the classical de Morgan's Law, generalised to all transreal numbers by monotonicity and the absorptiveness of nullity, with all symbols read paraconsistently. □

3 Proposition Space

We define *logical space*, very generally, as a scalar space whose axes are logical elements and whose scalar values are semantic values. This allows us to apply many mathematical methods to logic. We begin with an orthogonal co-ordinate frame where each axis is a copy of the transreal number line. This gives us a trans-Cartesian co-ordinate frame. The more abstract *Proposition space* has each possible world as an axis and each point is a proposition whose co-ordinates are the semantic values of that proposition in each possible world. This allows us to apply mathematical and logical operations, simultaneously, to propositions in all possible worlds.

We define *logical transformations*, very generally, as a transformations in logical space. In Sect. 2, above, we use a paraconsistent Sheffer Stroke. This transformation can be summarised by the side condition that its conclusion departs no further from being equally False and True than its antecedents. We now use a generalisation of the classical conditional which has the property that its conclusion is at least as true

as its antecedents. We say that a proposition, or other point, p, is *derived* from a proposition, or other point, q, if and only if there is a chain of logical transformations that maps q onto p. In particular the chain of conditionals is a *proof path*.

3.1 Transreal Functions of All Propositions

We will define atomic propositions as objects in our transreal model. However, to begin, we assume that the set of atomic propositions is a countable set. Hence the set of atomic propositions can be written in the form $\{P_1, P_2, P_3, \ldots\}$, where $P_i \neq P_j$ whenever $i \neq j$. Note that we are not defining the set of atomic propositions yet. This is just a representation of the set to be defined.

A possible world is a binding of the atomic propositions to its semantic values. That is, at a given possible world, each atomic proposition takes on a semantic value in $\mathbb{R}^T$. Thus we can interpret a possible world as a function from $\{P_1, P_2, P_3, \ldots\}$ to $\mathbb{R}^T$. But this is nothing more than an infinite sequence of elements from $\mathbb{R}^T$. In this way, every possible world is an element from $(\mathbb{R}^T)^{\mathbb{N}}$. Conversely every element from $(\mathbb{R}^T)^{\mathbb{N}}$ is a possible world.

To facilitate understanding of possible worlds let us introduce a simplified model with just two semantic values F and T and just two atomic propositions P_1 and P_2. An example of a possible world is the world where the atomic proposition P_1 has semantic value F and the atomic proposition P_2 has semantic value T. Another example is the world where both the atomic propositions P_1, P_2 have semantic value F. The first world can be represented by the pair (F, T) and the second world by the pair (F, F), where the first co-ordinate of the pair represents the semantic value of the proposition P_1 and the second co-ordinate represents the semantic value of the proposition P_2. In this simplified model there are four possible worlds: (F, T), (F, F), (T, T) and (T, F). These pairs can be viewed geometrically as "vectors" (Fig. 1).

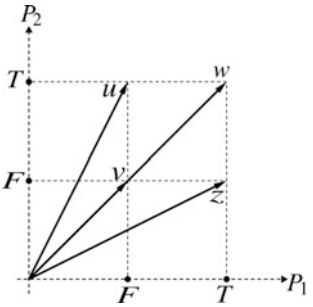


Fig. 1 Possible worlds are vectors with a co-ordinate in each proposition. Here $u = (F, T)$, $v = (F, F)$, $w = (T, T)$, $z = (T, F)$ are vectors

Returning to our transreal model, possible worlds are also "vectors" but with infinitely many co-ordinates, not just two, and these co-ordinates take values in $\mathbb{R}^T$ not in $\{F, T\}$. Possible worlds are points in $(\mathbb{R}^T)^{\mathbb{N}}$ whose axes are atomic propositions and whose co-ordinates are the semantic value of the underlying atomic proposition in that possible world. Given a possible world $w = (w_i)_{i \in \mathbb{N}} \in (\mathbb{R}^T)^{\mathbb{N}}$, we have that w_i corresponds to the semantic value of P_i in w, for each $i \in \mathbb{N}$.

Generic table (cells)

	P_1	P_2
u	F	T
v	F	F
w	T	T
z	T	F

Possible worlds (rows)

	P_1	P_2
u	F	T
v	F	F
w	T	T
z	T	F

Atomic propositions (columns)

	P_1	P_2
u	F	T
v	F	F
w	T	T
z	T	F

We now define atomic propositions in the simplified model, before generalising them in our transreal model. We can represent all possible worlds in a table.

The generic table associates propositions with worlds. A possible world is a row of the table which gives the semantic values of successive propositions. An atomic proposition is uniquely determined by its semantic values in all possible worlds. That is, if we know the semantic values of an atomic proposition in each possible world, we know this atomic proposition. In other words, an atomic proposition is completely determined by a column of the table. Thus $P_1 = (F, F, T, T)$ and $P_2 = (T, F, T, F)$. Here the atomic propositions are 4-tuples, which is to say they are "vectors" of four co-ordinates. Of course we cannot have a picture of the atomic propositions as vectors, because this figure would be in four dimensions, but we can draw the projections in three dimensions, ignoring the fourth co-ordinate. Thus the projections of P_1 and P_2, in three dimensions, are (F, F, T) and (T, F, T) (Fig. 2).

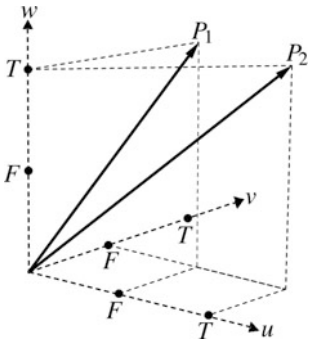


Fig. 2 Atomic propositions are vectors with a co-ordinate in each possible world. Here the projections $P_1 = (F, F, T)$ and $P_2 = (T, F, T)$ are vectors

We now extend the simplified model to the transreal model. For each $i \in \mathbb{N}$, let p_i be the co-ordinate function $p_i : (\mathbb{R}^T)^{\mathbb{N}} \to \mathbb{R}^T$ where $p_i((w_j)_{j\in\mathbb{N}}) = w_i$. Given $i \in \mathbb{N}$, notice that for each possible world $w = (w_j)_{j\in\mathbb{N}}$, we interpret $p_i(w)$ as the semantic value of the i-th atomic proposition, P_i, in the possible world w. Hence for each $i \in \mathbb{N}$, $(p_i(w))_{w\in(\mathbb{R}^T)^{\mathbb{N}}}$ is an $(\mathbb{R}^T)^{\mathbb{N}}$-tuple which expresses the semantic value of the atomic proposition P_i in all possible worlds. In this way, each atomic proposition, P_i, is represented by $(p_i(w))_{w\in(\mathbb{R}^T)^{\mathbb{N}}}$. This motivates the following definition.

Definition 7 Let $\mathcal{P} := \left\{ \left(p_1(w)\right)_{w\in(\mathbb{R}^{\mathbb{T}})^{\mathbb{N}}}, \left(p_2(w)\right)_{w\in(\mathbb{R}^{\mathbb{T}})^{\mathbb{N}}}, \left(p_3(w)\right)_{w\in(\mathbb{R}^{\mathbb{T}})^{\mathbb{N}}}, \ldots \right\}$. We call each element from $\mathcal{P}$ an **atomic proposition**, hence $\mathcal{P}$ is the **set of atomic propositions**.

The set $\mathcal{P}$ is infinite. Because the $(\mathbb{R}^T)^{\mathbb{N}}$-tuples

$$\left(p_1(w)\right)_{w\in(\mathbb{R}^{\mathbb{T}})^{\mathbb{N}}}, \left(p_2(w)\right)_{w\in(\mathbb{R}^{\mathbb{T}})^{\mathbb{N}}}, \left(p_3(w)\right)_{w\in(\mathbb{R}^{\mathbb{T}})^{\mathbb{N}}}, \ldots$$

are pairwise distinct. Indeed for each $i,j \in \mathbb{N}$, such that $i \neq j$, there is $u \in (\mathbb{R}^T)^{\mathbb{N}}$ such that $p_i(u) \neq p_j(u)$. Thus $\left(p_i(w)\right)_{w\in(\mathbb{R}^{\mathbb{T}})^{\mathbb{N}}} \neq \left(p_j(w)\right)_{w\in(\mathbb{R}^{\mathbb{T}})^{\mathbb{N}}}$ whenever $i \neq j$.

By Definition 7, each atomic proposition is a point within the space $(\mathbb{R}^T)^{(\mathbb{R}^T)^{\mathbb{N}}}$, the set of the functions whose domain is the space of sequences of transreal numbers and whose counter-domain is the set of transreal numbers. Further, for each $i \in \mathbb{N}$, $\left(p_i(w)\right)_{w\in(\mathbb{R}^{\mathbb{T}})^{\mathbb{N}}}$ corresponds to i-th atomic proposition, that is, $\left(p_i(w)\right)_{w\in(\mathbb{R}^{\mathbb{T}})^{\mathbb{N}}}$ corresponds to P_i. And, for each $w \in (\mathbb{R}^T)^{\mathbb{N}}$, $p_i(w)$ corresponds to the semantic value of P_i in the possible world w.

Definition 8 Let $\neg : (\mathbb{R}^T)^{(\mathbb{R}^T)^{\mathbb{N}}} \to (\mathbb{R}^T)^{(\mathbb{R}^T)^{\mathbb{N}}}$,

$$\neg(p_w)_{w\in(\mathbb{R}^T)^{\mathbb{N}}} = (\neg p_w)_{w\in(\mathbb{R}^T)^{\mathbb{N}}}, \tag{1}$$

$\vee : (\mathbb{R}^T)^{(\mathbb{R}^T)^{\mathbb{N}}} \times (\mathbb{R}^T)^{(\mathbb{R}^T)^{\mathbb{N}}} \to (\mathbb{R}^T)^{(\mathbb{R}^T)^{\mathbb{N}}}$

$$(p_w)_{w\in(\mathbb{R}^T)^{\mathbb{N}}} \vee (q_w)_{w\in(\mathbb{R}^T)^{\mathbb{N}}} = (p_w \vee q_w)_{w\in(\mathbb{R}^T)^{\mathbb{N}}} \tag{2}$$

and $\& : (\mathbb{R}^T)^{(\mathbb{R}^T)^{\mathbb{N}}} \times (\mathbb{R}^T)^{(\mathbb{R}^T)^{\mathbb{N}}} \to (\mathbb{R}^T)^{(\mathbb{R}^T)^{\mathbb{N}}}$

$$(p_w)_{w\in(\mathbb{R}^T)^{\mathbb{N}}} \& (q_w)_{w\in(\mathbb{R}^T)^{\mathbb{N}}} = (p_w \& q_w)_{w\in(\mathbb{R}^T)^{\mathbb{N}}}. \tag{3}$$

We abuse notation, above, but the reader will perceive that, in (1), the symbol $\neg$, on the left hand side of the equality, refers to a function which is being defined on $(\mathbb{R}^T)^{(\mathbb{R}^T)^{\mathbb{N}}}$, while the symbol $\neg$, on the right hand side of the equality, refers to a function already defined on $\mathbb{R}^T$. Similarly for $\vee$ in (2) and for $\&$ in (3).

Definition 9 Let $A \subset (\mathbb{R}^T)^{(\mathbb{R}^T)^{\mathbb{N}}}$ and let $\mathcal{L}_A$ be defined as the class of all sets $X_A \subset (\mathbb{R}^T)^{(\mathbb{R}^T)^{\mathbb{N}}}$, where X_A has the following properties:

(i) $A \subset X_A$ and
(ii) if $p, q \in X_A$ then $\neg p, p \vee q, p\&q \in X_A$.

Define $L_A := \bigcap_{X_A \in \mathcal{L}_A} X_A$. Let $p \in (\mathbb{R}^T)^{(\mathbb{R}^T)^{\mathbb{N}}}$ then we say that p is a **logical combination of elements from** A if and only if $p \in L_A$. Given $B \subset (\mathbb{R}^T)^{(\mathbb{R}^T)^{\mathbb{N}}}$, we say that B is a **logically independent** set, if and only if, for all $p \in B$, p is not a logical combination of elements from $B \backslash \{p\}$.

Proposition 1 *The set $\mathcal{P}$ is logically independent.*

Proof Suppose $\mathcal{P}$ is not logically independent. This means there is an element from $\mathcal{P}$ which is a logical combination of some other elements from $\mathcal{P}$. Without loss of generality, suppose $\big(p_1(w)\big)_{w\in(\mathbb{R}^{\mathbb{T}})^{\mathbb{N}}}$ is a logical combination of some elements from $\mathcal{P} \setminus \Big\{\big(p_1(w)\big)_{w\in(\mathbb{R}^{\mathbb{T}})^{\mathbb{N}}}\Big\} = \Big\{\big(p_2(w)\big)_{w\in(\mathbb{R}^{\mathbb{T}})^{\mathbb{N}}}, \big(p_3(w)\big)_{w\in(\mathbb{R}^{\mathbb{T}})^{\mathbb{N}}}, \ldots\Big\}$. So $\big(p_1(w)\big)_{w\in(\mathbb{R}^{\mathbb{T}})^{\mathbb{N}}}$ is the result of a determinate composition, T, between the functions $\neg$, $\vee$ and $\&$, applied to some elements from $\Big\{\big(p_2(w)\big)_{w\in(\mathbb{R}^{\mathbb{T}})^{\mathbb{N}}}, \big(p_3(w)\big)_{w\in(\mathbb{R}^{\mathbb{T}})^{\mathbb{N}}}, \ldots\Big\}$. For some $m \in \mathbb{N}$, $\big(p_{j_1}(w)\big)_{w\in(\mathbb{R}^{\mathbb{T}})^{\mathbb{N}}}$, $\big(p_{j_2}(w)\big)_{w\in(\mathbb{R}^{\mathbb{T}})^{\mathbb{N}}}$, $\ldots$, $\big(p_{j_m}(w)\big)_{w\in(\mathbb{R}^{\mathbb{T}})^{\mathbb{N}}}$ are elements from $\Big\{\big(p_2(w)\big)_{w\in(\mathbb{R}^{\mathbb{T}})^{\mathbb{N}}}, \big(p_3(w)\big)_{w\in(\mathbb{R}^{\mathbb{T}})^{\mathbb{N}}}, \ldots\Big\}$ where T is applied. Now $\big(p_1(w)\big)_{w\in(\mathbb{R}^{\mathbb{T}})^{\mathbb{N}}} = T\Big(\big(p_{j_1}(w)\big)_{w\in(\mathbb{R}^{\mathbb{T}})^{\mathbb{N}}}, \big(p_{j_2}(w)\big)_{w\in(\mathbb{R}^{\mathbb{T}})^{\mathbb{N}}}, \ldots, \big(p_{j_m}(w)\big)_{w\in(\mathbb{R}^{\mathbb{T}})^{\mathbb{N}}}\Big) = = \Big(T\big(p_{j_1}(w), p_{j_2}(w), \ldots, p_{j_m}(w)\big)\Big)_{w\in(\mathbb{R}^{\mathbb{T}})^{\mathbb{N}}}$. That is,

$$p_1(w) = T\big(p_{j_1}(w), p_{j_2}(w), \ldots, p_{j_m}(w)\big) \text{for all } w \in (\mathbb{R}^{\mathbb{T}})^{\mathbb{N}}. \tag{4}$$

Let arbitrary $(u_1, u_2, \ldots) \in (\mathbb{R}^T)^{\mathbb{N}}$, $u' \neq T(u_{j_1}, u_{j_2}, \ldots, u_{j_m})$, $v = (v_1, v_2, v_3 \ldots) := (u', u_2, u_3, \ldots)$. We have $v \in (\mathbb{R}^T)^{\mathbb{N}}$ and $v_1 \neq T(v_{j_1}, v_{j_2}, \ldots, v_{j_m})$. Hence $p_1(v) \neq T\big(p_{j_1}(v), p_{j_2}(v), \ldots, p_{j_m}(v)\big)$. This contradicts (4). Hence $\mathcal{P}$ is logically independent. □

Remark 1 Proposition 1 justifies us in calling the propositions in $\mathcal{P}$ atomic.

Definition 10 Let $\Pi := L_{\mathcal{P}}$, that is,

$$\Pi = \Big\{q \in (\mathbb{R}^T)^{(\mathbb{R}^T)^{\mathbb{N}}}; \text{q is logical combination of elements from } \mathcal{P}\}.$$

We call Π a **proposition space** and each element from Π a **proposition**.

Proposition 2 *The set Π is countable.*

Proof Each element from Π can be written as a finite sequence of symbols from $S := \mathcal{P} \cup \{\neg, \vee, \&, (,)\}$. Hence an element from Π can be seen as an element from S^n for some $n \in \mathbb{N}$. Thus Π can be seen as a subset of $\bigcup_{n\in\mathbb{N}} S^n$. Since S is countable, $\bigcup_{n\in\mathbb{N}} S^n$ is countable, whence Π is countable. □

Corollary 1 *Proposition space, Π, is a proper subset of $(\mathbb{R}^T)^{(\mathbb{R}^T)^{\mathbb{N}}}$. Thus Π is different to $(\mathbb{R}^T)^{(\mathbb{R}^T)^{\mathbb{N}}}$.*

Proof Denote the cardinality of a set S by $|S|$ and let $\mathfrak{c}$ be the cardinality of the continuum. Note that, using Cantor's transfinite arithmetic, $|\Pi| = \aleph_0$ but $\left|(\mathbb{R}^T)^{(\mathbb{R}^T)^{\mathbb{N}}}\right| = \mathfrak{c}^{(\mathfrak{c}^{\aleph_0})} = \mathfrak{c}^{\mathfrak{c}} = (2^{\aleph_0})^{\mathfrak{c}} = 2^{\aleph_0 \times \mathfrak{c}} = 2^{\mathfrak{c}} > \aleph_0$. Hence $|\Pi| < \left|(\mathbb{R}^T)^{(\mathbb{R}^T)^{\mathbb{N}}}\right|$ whence $\Pi \neq (\mathbb{R}^T)^{(\mathbb{R}^T)^{\mathbb{N}}}$. □

Remark 2 Notice that every proposition, in the ordinary sense, lies in Π but Π is within the bigger set $(\mathbb{R}^T)^{(\mathbb{R}^T)^{\mathbb{N}}}$. Hence there are points from $(\mathbb{R}^T)^{(\mathbb{R}^T)^{\mathbb{N}}}$ that are not logical combinations of atomic propositions. These points are not expressible in any language but this does not require that they are meaningless. This issue is taken up in Sect. 4. We call the whole space, $(\mathbb{R}^T)^{(\mathbb{R}^T)^{\mathbb{N}}}$, **hyper-proposition space**.

Remark 3 We emphasise that in our transreal model there are three distinct sets of propositions:

- $\mathcal{P}$ is the set of all atomic propositions.
- Π, called proposition space, is the set of all propositions (atomic or molecular).
- $(\mathbb{R}^T)^{(\mathbb{R}^T)^{\mathbb{N}}}$, called hyper-proposition space, is the whole set of all function whose domain is the space of sequences of transreal numbers and whose counter-domain is the set of transreal numbers.

We have the proper inclusion: $\mathcal{P} \subsetneq \Pi \subsetneq (\mathbb{R}^T)^{(\mathbb{R}^T)^{\mathbb{N}}}$.

Definition 11 We say that a proposition $(p_w)_{w\in(\mathbb{R}^T)^{\mathbb{N}}} \in \Pi$ is necessary in a possible world u if and only if $p_w > 0$ for all $w \in (\mathbb{R}^T)^{\mathbb{N}}$, such that u accesses w. And we say that $(p_w)_{w\in(\mathbb{R}^T)^{\mathbb{N}}} \in \Pi$ is possible in a possible world u if and only if there is a $w \in (\mathbb{R}^T)^{\mathbb{N}}$ such that u accesses w and $p_w > 0$. A world u accesses a world w if and only if there is a linear transformation on $(\mathbb{R}^T)^{\mathbb{N}}$ that maps u onto w.

3.2 Logical Transformations

In classical logic the connective *conditional*, $\rightarrow$, is defined as follows [19]:

$$\begin{aligned}&\rightarrow: \{F,T\} \times \{F,T\} \rightarrow \{F,T\}\\ &F \rightarrow F = T, \quad F \rightarrow T = T\\ &T \rightarrow F = F, \quad T \rightarrow T = T.\end{aligned}$$

This means that:

(i) if the antecedent is false then the conditional is true, regardless of the value of the consequent and
(ii) if the consequent is true then the conditional is true, regardless of the value of the antecedent.

In non-classical logics, the conditional is defined in various ways. However, the above observation gives us the familiar intuition that the conditional is true if and only if the degree of truth of the consequent is greater than or equal to the degree of truth of the antecedent. Motivated by this, we propose the following definition.

Definition 12 Let $T : \Pi \rightarrow \Pi$ and, for each $(p_w)_{w\in(\mathbb{R}^T)^{\mathbb{N}}} \in \Pi$, write a transformation as $\big(T(p_w)\big)_{w\in(\mathbb{R}^T)^{\mathbb{N}}} := T\Big((p_w)_{w\in(\mathbb{R}^T)^{\mathbb{N}}}\Big)$. We call T a **logical transformation** if and only if $p_w \leq T(p_w)$ for all $w \in (\mathbb{R}^T)^{\mathbb{N}}$.

It is well known that, in classical, propositional calculus, one can derive any proposition from bottom [11, 18], $\perp$, that is:

$$\text{for all } p \text{ within the system}, \perp \vdash p. \tag{5}$$

Consider $\Gamma := \Big\{(p_w)_{w\in(\mathbb{R}^T)^{\mathbb{N}}} \in \Pi;\ p_w \in \{-\infty, \infty\}\Big\}$.

In Γ, the meta-theorem in (5) is equivalent to an affirmative answer to the question: *is there any point in Γ from which we can derive, by means of a logical transformation, any point in Γ*? One such point is $(-\infty)_{w\in(\mathbb{R}^T)^{\mathbb{N}}}$. That is, for all $p \in \Gamma$, there is a logical transformation, T, such that $T\Big((-\infty)_{w\in(\mathbb{R}^T)^{\mathbb{N}}}\Big) = p$.

In an analogous way we can read the meta-theorem: one can derive top, $\top$, from any proposition, as: for all p within the system, $p \vdash \top$. Now $(\infty)_{w\in(\mathbb{R}^T)^{\mathbb{N}}}$ is a point which can be derived from any other point, that is: or all $p \in \Gamma$, there is a logical transformation T such that $T(p) = (\infty)_{w\in(\mathbb{R}^T)^{\mathbb{N}}}$.

If we consider a propositional calculus in which one allows continuous degrees of truth and falseness, and, furthermore, propositions that can be both true and false then the "bottom-point" still is the point from where every point can be reached and the "top-point" still is the point to which every point can derive. The verification of this is analogous to the classical case but now we consider $\Sigma := \Big\{(p_w)_{w\in(\mathbb{R}^T)^{\mathbb{N}}} \in \Pi;\ p_w \in [-\infty, \infty]\Big\}$ instead of Γ. And so, for all $p \in \Sigma$, there is a

logical transformation T such that $T\left((-\infty)_{w\in(\mathbb{R}^T)^{\mathbb{N}}}\right) = p$ and for all $p \in \Sigma$, there is a logical transformation T such that $T(p) = (\infty)_{w\in(\mathbb{R}^T)^{\mathbb{N}}}$.

If we extend propositional calculus to Π then the "bottom-point" is no longer the "privileged place" from which we can derive any point of Π. Let $q = (q_w)_{w\in(\mathbb{R}^T)^{\mathbb{N}}} \in \Pi$ such that, for some $u \in (\mathbb{R}^T)^{\mathbb{N}}$, $q_u = \Phi$. The u-th co-ordinate of the "bottom-point" is $-\infty$ and, since $-\infty \leq \Phi$ does not hold then we can not derive q from $(-\infty)_{w\in(\mathbb{R}^T)^{\mathbb{N}}}$ by a logical transformation. Thus q is an inaccessible point of Π by means of a logical transformation that has as initial points those whose co-ordinates belong to $[-\infty, \infty]$. Points like q, that have Φ as the value of some co-ordinate, are only derivable from points whose corresponding co-ordinate is also Φ.

3.3 A Criterion to Distinguish Classical and Non-classical Propositions

Since a proposition is a point in proposition space and since an axis, u, of this space is a possible world, if a proposition behaves classically, its numerical value at u is $-\infty$ (classical false) or ∞ (classical true). Hence its contradictory is a point in the proposition space that has u-co-ordinate $\neg(-\infty) = \infty$ or $\neg(\infty) = -\infty$. Thus if, in a given possible world u, a proposition is classical then the absolute difference between the u-co-ordinates of the proposition and of its contradictory is $|(-\infty) - (\infty)|$ or $|\infty - (-\infty)|$, whichever case holds $|(-\infty) - (\infty)| = |\infty - (-\infty)| = \infty$. Conversely if the absolute difference between the u-co-ordinates of a proposition $p = (p_w)_{w\in(\mathbb{R}^T)^{\mathbb{N}}}$ and of its contradictory is ∞ then $\infty = |p_u - (\neg p_u)| = |p_u - (-p_u)| = 2|p_u|$ whence $p_u = -\infty$ or $p_u = \infty$. Hence p is classical in the possible world u. This demonstrates the following proposition.

Proposition 3 *Let $p = (p_w)_{w\in(\mathbb{R}^T)^{\mathbb{N}}} \in \Pi$ and $(q_w)_{w\in(\mathbb{R}^T)^{\mathbb{N}}} = \neg p$. The proposition p is classical in the possible world $u \in (\mathbb{R}^T)^{\mathbb{N}}$ if and only if $|p_u - q_u| = \infty$.*

Thus in our proposition space, which instantiates a total semantics, we stipulate a criterion to distinguish classical from non-classical propositions. Usually there is no way to distinguish atomic propositions because they have no inner structure and, therefore, no feature that can be used as a criterion for the distinction, which is taken arbitrarily. But, as propositions are points located at co-ordinates in proposition space, atomic propositions are elements of a structured space and this structure offers a criterion for distinguishing classical and non-classical atomic propositions.

4 Discussion

Proposition space, Π, is a geometrical version of the usual propositional calculus. It has all the expected, logical properties of paraconsistent, propositional calculi and offers a classical structure when it operates on positive and negative infinity, which represent the classical truth values, False and True respectively.

Proposition space, Π, is a small part of the entire hyper-proposition space, $\left(\mathbb{R}^T\right)^{\left(\mathbb{R}^T\right)^{\mathbb{N}}}$. The cardinality of the former is $\aleph_0$; the cardinality of the latter is greater than the cardinality of the continuum. Thus we can infer that there are proposition points that will never be expressed by logical combinations of atomic propositions: these points are essentially non-logical, since we understand by "logical" that a point belongs to proposition space. But does this imply that *complementary* elements, in hyper-proposition space but outside proposition space, are meaningless? We think not.

Complementary points represent what is logically inexpressible: they contain information, since they lie in the entire $\left(\mathbb{R}^T\right)^{\left(\mathbb{R}^T\right)^{\mathbb{N}}}$ space, but this information cannot be accessed through ordinary, logical reasoning, expressed in any language. Perhaps we should associate, at least some of them, with metaphysical or philosophical statements that cannot be proved by logical apparatus?

On this point let us emphasise the role of logical transformations. In Sect. 3.2 a logical transformation is defined as a certain transformation in proposition space. If we extend this definition, by allowing transformations in the entire hyper-proposition space, then the concept of a continuous proof path appears. Recall that a *proof path* is a chain of compositions of logical transformations. If we restrict ourselves to the enumerable proposition space then a proof path has a finite or denumerably infinite length but, more generally speaking, a proof path is a geometrical translation of the notion of demonstration that is used in logic: a list of propositions that start with premises, supposed to be true, and a conclusion that is true in virtue of the soundness or correctness of the rules of inference. In the entire hyper-proposition space, this proof path can be continuous. This fact is very significant: it gives us a geometrical entity, a continuous path, that has logical meaning. This path must have a non-denumerable infinitude of hidden propositions —let us call them subatomic propositions. Hence we see the need to expand the notion of a discrete demonstration to a continuous demonstration in which, between two proof steps indexed with finite numbers, there is a continuum of steps that cannot be expressed in language.

A finite simulation (not an emulation) of a machine that operates on a continuum of propositions is given in [3].

5 Conclusion

We develop a model for a total semantics, for possible worlds, for proposition space and for hyper-proposition space. We define the set of semantic values as the set of transreal numbers, $\mathbb{R}^T$. This is sufficient to model classical truth values, paraconsistent, fuzzy and gap values. We give a paraconsistent version of the Sheffer Stroke which is sufficient to extend all classical logics to a paraconsistent form. We then turn our attention to logical spaces. We define each possible world, in world space, as a sequence of transreal numbers. We define the set of propositions, Φ, as a certain subset of $\left(\mathbb{R}^T\right)^{\left(\mathbb{R}^T\right)^{\mathbb{N}}}$. That is each proposition is a tuple, $\left(p_w\right)_{w\in\left(\mathbb{R}^T\right)^{\mathbb{N}}}$, where each axis is one possible world, w, and each co-ordinate, p_w, of p is the transreal semantic value of p in the possible world w. This allows us to operate on a proposition in all possible worlds at the same time. We define in Φ the concepts of possibility, necessity and logical transformation and we define a criterion which distinguishes whether a proposition is or is not classical. A proposition is classical in a determined, possible world if and only if the absolute difference between the semantic value of the proposition and its negation, in the underlying possible world, is infinite. We show that proofs based on the conditional can infer gap values only from gap values, that the proposition which is classically False ($-\infty$), in all possible worlds, is a bottom point from which all non-gap propositions can be derived and that the proposition which is classically True (∞), in all possible worlds, is a top point entailed by all non-gap propositions. We discuss the need for continuous versions of logic that capture inferences and proofs in our very high cardinality hyper-proposition space, $\left(\mathbb{R}^T\right)^{\left(\mathbb{R}^T\right)^{\mathbb{N}}}$.

Acknowledgments The authors would like to thank the members of Transmathematica for many helpful discussions. The third author's research was financially supported, in part, by the School of Systems Engineering at the University of Reading and by a Research Travel Grant from the University of Reading Research Endowment Trust Fund (RETF).

References

1. Anderson JADW (1997) Representing geometrical knowledge. Phil Trans Roy Soc Lond Ser B 352(1358):1129–1139
2. Anderson JADW (2002) Exact numerical computation of the rational general linear transformations. In: Lateki LJ, Mount DM, Wu AY (eds) Vision geometry XI, proceedings of SPIE, vol 4794, pp 22–28
3. Anderson JADW (2005) Perspex machine iii: continuity over the turing operations. In: Lateki LJ, Mount DM, Wu AY (eds) Vision geometry XIII, proceedings of SPIE, vol l5675, pp 112–123
4. Anderson JADW (2011) Evolutionary and revolutionary effects of transcomputation. In: 2nd IMA conference on mathematics in defence. Institute of Mathematics and its Applications, Oct 2011

5. Anderson JADW (2014) Trans-floating-point arithmetic removes nine quadrillion redundancies from 64-bit ieee 754 floating-point arithmetic. In: Lecture notes in engineering and computer science: proceedings of the world congress on engineering and computer science 2014, WCECS 2014, San Francisco, USA, 22–24 Oct 2014, vol 1, pp 80–85
6. Anderson JADW, dos Reis TS (2014) Transreal limits expose category errors in ieee 754 floating-point arithmetic and in mathematics. In: Lecture notes in engineering and computer science: proceedings of the world congress on engineering and computer science 2014, WCECS 2014, San Francisco, USA, 22–24 Oct 2014, vol 1, pp 86–91
7. Anderson JADW, Völker N, Adams AA (2007) Perspex machine viii: axioms of transreal arithmetic. In: Lateki LJ, Mount DM, Wu AY (eds) Vision geometry XV, proceedings of SPIE, vol 6499, pp 2.1–2.12
8. Anderson JADW, Gomide W (2014) Transreal arithmetic as a consistent basis for paraconsistent logics. In: lecture notes in engineering and computer science: proceedings of the world congress on engineering and computer science 2014, WCECS 2014, San Francisco, USA, 22–24 Oct 2014, vol 1, pp 103–108
9. Audi R (1997) The Cambridge dictionary of philosophy, 2nd edn. Cambridge University Press, Cambridge
10. Bellman RE, Zadeh LA (1977) Local and fuzzy logics. In: Dunn JM, Epstein G (eds) Modern uses of multiple-valued logic. D. Reidel Publishing Co., Dordrecht, pp 103–165
11. Carnelli W, Marcos J (2001) Ex contradictione non sequitur quodlibet. In: Proceedings of 2nd conference on reasoning and logic, Bucharest
12. da Costa NCA (1982) The philosophical import of paraconsistent logic. J Non-Classical Logic 1:1–19
13. dos Reis TS, Anderson JADW (2014) Construction of the transcomplex numbers from the complex numbers. In: Lecture notes in engineering and computer science: proceedings of the world congress on engineering and computer science 2014, WCECS 2014, San Francisco, USA, 22–24 Oct 2014, vol 1, pp 97–102
14. dos Reis TS, Anderson JADW (2015) Transreal calculus. IAENG Int J Appl Mathe 45(1):51–63
15. Floyd J (2005) Wittgenstein on philosophy of logic and mathematics. In: Shapiro S (ed) Oxford handbook of philosophy of logic and mathematics, chapter 4. Oxford University Press, Oxford, pp 75–128
16. Gomide W (2013) O princpio de não-contradição e sua tradução para a aritmética transreal. Investigação Filosófica 4(1)
17. Jansen TMV (2001) Frege, contextuality and composionality. J Logic Lang Inf 10
18. Martin-Löf P (1996) On the meanings of the logical constants and the justification of the logical laes. Nordic J Philos Logic 1:11–60
19. Nolt J, Rohatyn D (1988) Logic. McGraw-Hill, New York
20. Priest G, Beall JC, Gab BA (2004) The law of non-contradiction. Oxford University Press, Oxford
21. Priest G, Tanaka K, Weber Z (2013) Paraconsistent logic. The Stanford encyclopedia of philosophy, http://plato.stanford.edu/archives/fall2013/entries/logic-paraconsistent
22. Rosen KH (2007) Discrete mathematics and its appications, 6th edn. McGraw-Hillition, New York
23. Shramko Y, Wansing H (2007) Entailment relations and truth values. Bull Sect Logic 36:131–143
24. Shramko Y, Wansing H (2007) Truth and falsehood. an inquiry into generalised logical values. Trends Logic 36
25. van Fraasen BC (1966) Singular terms, truth values, gaps, and frege logic. J Philos 63:481–495
26. Wittgenstein L (1921) Logisch-philosophische abhandlung. Unesma
27. Wittgenstein L (2010) Tractatus Logico-Philasophicus. Gutenberg
28. Zadeh LA (1975) Fuzzy logic and approximate reasoning. Synthese 30:407–428

The Adaptive80 Round Robin Scheduling Algorithm

Christopher McGuire and Jeonghwa Lee

Abstract Many improved versions of the Round Robin CPU scheduling algorithm have been created to fix the shortcomings of the Standard Round Robin algorithm. This research presents a new Round Robin algorithm, named *Adaptive80 Round Robin*, and compares it with several improved Round Robin algorithms. The results show that even among several improved Round Robin algorithms, Adaptive80 Round Robin will generally perform better than or equal to them for the average turnaround time, average waiting time, average response time, and number of context switches.

Keywords Context switch · Response time · Round robin · Scheduling algorithm · Turnaround time · Waiting time

1 Introduction

The Round Robin (RR) CPU scheduling algorithm, referred to hereafter as Standard RR, is commonly used in time sharing and real time operating systems [1–6] because it keeps response time low [1, 7], gives each process a fair share of time on the CPU, and is starvation free [8]. Despite these advantages, it is well known that Standard RR suffers from several disadvantages; those being low throughput, high turnaround time, high waiting time, and a high amount of context switches [1, 4, 7, 9, 10]. Other researchers have proposed improved versions of Standard RR to minimize these shortcomings [11].

C. McGuire · J. Lee (✉)
Department of Computer Science and Engineering, Shippensburg University, Shippensburg, PA 17257, USA
e-mail: jlee@ship.edu
URL: http://www.cs.ship.edu/∼jlee

C. McGuire
e-mail: cm7602@cs.ship.edu

H.K. Kim et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-94-017-7236-5_17

This research presents a new RR algorithm, named *Adaptive80 RR* and compares its effectiveness with five other improved RR algorithms by measuring their average turnaround times, average waiting times, average response times, and number of context switches. Turnaround time is the time from the submission of a process to its completion. Waiting time is the amount of time a process spends waiting in the ready queue. Response time is the time from the submission of a process until it first receives the CPU. A context switch is when the CPU switches from one running process to another. These other improved RR algorithms to be compared are A New RR (AN RR) [1], Optimized RR [2], Priority-Based RR [4], Adaptive RR [10], and Efficient RR [6].

As described in further detail later, Adaptive80 RR gets its name because the time quantum is set equal to the burst time of the process at the 80th percentile. To further test the effectiveness of Adaptive80 RR, the algorithm was also tested using other percentile values. When referring to a set of Adaptive RR algorithms that each use the burst time of the process at a different percentile, this paper will refer to them as *Adaptive Percentile algorithms*.

Section 2 discusses current related work in the field of RR algorithms. Section 3 describes each of the RR algorithms used. Section 4 shows a simulation of each algorithm for a hypothetical set of processes. Section 5 explains the numerical experiment used to analyze the effectiveness of Adaptive80 RR. Conclusion remarks are in Sect. 6.

2 Related Work and Current Trends

In 2009, Matarneh proposed the Self-Adjustment Time Quantum Round Robin algorithm (SARR) [12]. This algorithm introduced the concept of using a dynamic time quantum based on the burst times of the processes rather than using a static time quantum [1, 13–16].

Since that time, other researchers have expanded on the concept of using a dynamic time quantum. Recent algorithms that utilize a dynamic time quantum include Dynamic Quantum with Re-Adjusted RR (DQRRR) [14], Average Max RR (AMRR) [17], Shortest Remaining Burst RR (SRBRR) [15], Shortest Remaining Burst RR (SRBRR) using a finest time quantum [18], Min-Max RR (MMRR) [16], and Adaptive RR [10]. These algorithms also make use of sorting the processes by burst time in order to remove shorter processes from the system more quickly as this has been shown to reduce the average turnaround time and average waiting time.

Another recent trend involves combining elements of other traditional CPU scheduling algorithms with Standard RR. As described in further detail later, Priority-Based RR combines priority scheduling with RR Scheduling [4], and Efficient RR combines the Shortest Remaining Time algorithm with RR scheduling [6].

3 Descriptions of the Round Robin Algorithms

3.1 AN RR

AN RR focuses on calculating an ideal time quantum [1]. It is like Standard RR with the following exception. Each time a process moves in or out of the ready queue, the time quantum is recalculated. If the ready queue is empty, then the time quantum equals the burst time of the running process. Otherwise, the time quantum equals the average burst time of the processes in the ready queue [11].

3.2 Optimized RR

Optimized RR is like Standard RR with the following exceptions. Optimized RR consists of two phases. During phase 1, processes are executed in order just like they are in Standard RR, and each process runs for one time slice. During phase 2, the time quantum is doubled, and processes are executed in the order of their remaining burst times with shorter times running before longer times. After each process has run for one time slice, the phase shifts back to phase one [2, 11]. In [2], no information was given as to what would happen if a process arrived mid-phase. This paper assumes that processes that arrive mid-phase will not get a chance to run until the next phase. This paper also assumes the time quantum resets to its initial value after the second phase.

3.3 Priority-Based RR

Priority-Based RR combines elements of priority scheduling and RR scheduling. Priority-Based RR consists of two phases. During phase 1, processes are executed in RR fashion in the order of their default priorities, and each process runs for one time slice [4]. During phase 2, processes are assigned new priorities based on their remaining burst times with shorter remaining burst times receiving higher priorities. Processes are executed in the order of their new priorities, and each process runs to completion [4, 11]. This paper assumes that processes that arrive mid-phase will not get a chance to run until the next phase.

3.4 Adaptive RR

Like AN RR, Adaptive RR focuses on calculating an ideal time quantum [10]. Adaptive RR is like Standard RR with the following exceptions. First, processes are

sorted by their burst times with the shorter processes at the front of the ready queue. Next, the adaptive time quantum is calculated. If the number of processes in the ready queue is even, then the time quantum equals the average burst time of all the processes. Otherwise, the time quantum equals the burst time of the process in the middle of the ready queue. Any processes that arrive in the middle of the algorithms execution are added at the end of the queue and do not run during the current round. After each of the initial processes have had a chance to run, the process repeats [11].

3.5 *Efficient RR*

Efficient RR combines elements of the Shortest Remaining Time (SRT) algorithm and Standard RR. In the SRT algorithm, the process with the shortest remaining burst time is always selected to run, and preemption can occur whenever a new process arrives. One of the downsides to SRT is that processes with long remaining burst times can suffer from starvation [1, 3, 5, 11].

Efficient RR is just like the SRT algorithm, but instead of preemption occurring whenever a new process arrives, preemption only occurs at the end of the time slice. At the end of the time slice, when it comes time to select a process to run, the process with the shortest remaining burst time is always selected [6, 19]. Long processes can suffer from starvation in Efficient RR just like in SRT [11].

3.6 *Adaptive80 RR*

Several sources state that, as a rule of thumb, 80 % of the processes' burst times should be shorter than the time quantum [3, 9, 13, 20], while other sources state that using the median burst time for a set of processes for the time quantum yields good results [1, 12]. Additionally, [10, 13, 14, 18] actually make use of the median process burst time to calculate the time quantum in their own improved RR algorithms. These differing statements drive the reasoning behind the alteration of Adaptive RR to create the new Adaptive80 RR algorithm.

In Adaptive80 RR, processes are sorted by their burst times with the shortest processes at the front of the ready queue just like in Adaptive RR. Next, the adaptive time quantum is calculated. Instead of using the median burst time of the processes or the average burst time for the time quantum as Adaptive RR does, Adaptive80 RR sets the time quantum equal to the burst time of the process at the 80th percentile. This ensures that at least 80 % of the processes' burst times are less than the time quantum. For other Adaptive Percentile algorithms, they set the time quantum equal to the burst time of the process at their given percentiles. For example, Adaptive20 would use the process at the 20th percentile. After each process has had a chance to run during the current round, the process repeats. Any

processes that arrive in the middle of the algorithms execution are added at the end of the ready queue and do not run during the current round.

Algorithm of the Adaptive80 RR Pseudo Code

1. While (ready queue is not empty or a process is still running or processes are still arriving):
   (a) If a new process arrived, add it to the ready queue but do not count it as part of the current round
   (b) If all of the processes in the current round have had a chance to run:
      (i) All processes in the ready queue are included as part of the current round
      (ii) Sort the processes in the ready queue in increasing order by their burst times
      (iii) *index* = size of ready queue * 0.8
      (iv) *time quantum* = burst time of process at *index*
   (c) Select a process to run:
      (i) If the time slice has expired, preempt the running process and run the next process in the ready queue
      (ii) Else If no process is currently running, run the next process in the ready queue
   (d) Increment the current time by 1
   (e) If the running process finished, record its turnaround time and waiting time, and remove it as the running process
2. End of While
3. Calculate average turnaround time, average waiting time, average response time, and number of context switches

4 Simulation

The following simulation, also presented in [11], gives an example of how each algorithm would schedule a given set of processes for a given time quantum (when applicable to the algorithm). For all algorithms assume the following list of processes given in Table 1. In the case where two processes arrive at the same time, the process listed first is assumed to have arrived just slightly before the next process.

Standard Round Robin

Gantt Chart

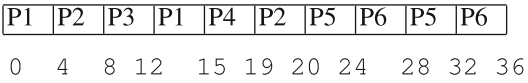
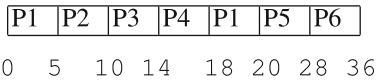


Table 1 Simulated list of processes

Process	Arrival time	Burst time	Priority
P1	0	7	2
P2	0	5	1
P3	3	4	6
P4	5	4	4
P5	10	8	3
P6	13	8	5

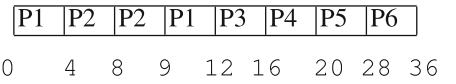
Time Quantum = 4
Average Turnaround Time = 17.17
Average Waiting Time = 11.17
Average Response Time = 6.67
Number of Context Switches = 9

AN RR
Gantt Chart



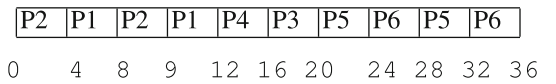
Time Quantum calculated by algorithm
Average Turnaround Time = 15.83
Average Waiting Time = 9.83
Average Response Time = 7.67
Number of Context Switches = 6

Optimized RR
Gantt Chart



Time Quantum = 4 (Phase One), 8 (Phase Two)
Average Turnaround Time = 15.00
Average Waiting Time = 9.00
Average Response Time = 8.17
Number of Context Switches = 7

Priority-Based RR
Gantt Chart

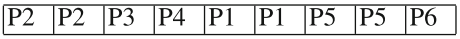


Time Quantum = 4
Average Turnaround Time = 15.67
Average Waiting Time = 9.67
Average Response Time = 7.50
Number of Context Switches = 9

Adaptive RR
Gantt Chart

Time Quantum calculated by algorithm
Average Turnaround Time = 16.67
Average Waiting Time = 10.67
Average Response Time = 7.17
Number of Context Switches = 7

Efficient RR
Gantt Chart



Time Quantum = 4
Average Turnaround Time = 13.33
Average Waiting Time = 7.33
Average Response Time = 7.33
Number of Context Switches = 8

Adaptive80 RR
Gantt Chart

Time Quantum calculated by algorithm
Average Turnaround Time = 14.33
Average Waiting Time = 8.33
Average Response Time = 8.33
Number of Context Switches = 5

5 Numerical Experiments

5.1 Hardware Specs

The hardware specs for the machine used to run this experiment were as follows:

- Operating System: Windows 8
- Processor: Intel Core i7-3630QM CPU @ 2.4 GHz
- Installed memory (RAM): 8.00 GB (7.89 GB usable)
- System type: 64-bit Operating System, x64-based processor

5.2 Sample Process Sets

Seven different sets of processes were randomly generated. Each set contained 1000 processes. Each process's burst time was randomly generated between 1 and 20. Because very few processes in a typical system have burst times greater than 8 ms [3], burst time values from 1 to 8 were weighted to occur more frequently in the set while values above 8 were weighted to occur less frequently. Each process's arrival time was offset from 0 to 9 from the previous process's arrival time with 0 to 4 occurring more frequently and 5 to 9 occurring less frequently. All process's priorities were 0.

5.3 Adaptive Percentile Algorithms Test

Each set of processes was simulated under the Adaptive Percentile algorithms at each multiple of 10 from 0 to 100. This led to 7 different simulations with 11 Adaptive Percentile algorithms used in each simulation. For each simulation ran, the Adaptive Percentile algorithms were ranked according their average turnaround time using the fractional ranking method. The lowest value was ranked with a 1, the next lowest value with a 2, and so on. The same was done for the average waiting time, average response time, and the number of context switches.

After all of the ranks were calculated, a Friedman statistical test was used to determine if the Adaptive Percentile algorithms differed in their rankings [21, 22]. The null and alternative hypotheses were as follows:

Average Turnaround Time:

- $\mathbf{H_0}$: There are no Adaptive Percentile algorithms which are ranked differently based on their average turnaround time compared to the other algorithms.
- $\mathbf{H_A}$: There is at least one Adaptive Percentile algorithm which is ranked differently based on its average turnaround time compared to at least one other algorithm.

Average Waiting Time:

- $\mathbf{H_0}$: There are no Adaptive Percentile algorithms which are ranked differently based on their average waiting time compared to the other algorithms.
- $\mathbf{H_A}$: There is at least one Adaptive Percentile algorithm which is ranked differently based on its average waiting time compared to at least one other algorithm.

Average Response Time:

- $\mathbf{H_0}$: There are no Adaptive Percentile algorithms which are ranked differently based on their average response time compared to the other algorithms.

- $\mathbf{H_A}$: There is at least one Adaptive Percentile algorithm which is ranked differently based on its average response time compared to at least one other algorithm.

Number of Context Switches:

- $\mathbf{H_0}$: There are no Adaptive Percentile algorithms which are ranked differently based on their number of context switches compared to the other algorithms.
- $\mathbf{H_A}$: There is at least one Adaptive Percentile algorithm which is ranked differently based on its number of context switches compared to at least one other algorithm.

After determining whether the null hypotheses should be accepted or rejected, the sum of the ranks and the mean rank for each Adaptive Percentile algorithm in these four categories were calculated. The differences of the mean ranks between each algorithm and the critical value for the difference of the mean ranks were calculated in order to determine which algorithms' rankings differed.

5.4 Improved RR Algorithms Test

Each set of processes was run under each improved RR algorithm and the Standard RR algorithm using four different time quanta. The choice of the time quanta was standardized based on the burst times of the processes in the sets. The first time quantum was the smallest number that would make Standard RR degenerate to the FCFS algorithm. The second, third, and fourth time quanta were three-fourths, one-half, and one-fourth of this number, rounded to the nearest whole number. Because 20 was the smallest number that would make Standard RR degenerate to FCFS for all of the process sets, the 4 time quanta were 5, 10, 15, and 20.

With 7 sets of processes and 4 time quanta per set, this led to 28 different simulations. For each simulation ran, the RR algorithms were ranked according their average turnaround time using the fractional ranking method. The same was done for the average waiting time, average response time, and the number of context switches.

After all of the ranks were calculated, a Friedman statistical test was used to determine if the algorithms differed in their rankings [21, 22]. The null and alternative hypotheses were as follows:

Average Turnaround Time:

- $\mathbf{H_0}$: There are no RR algorithms which are ranked differently based on their average turnaround time compared to the other algorithms.
- $\mathbf{H_A}$: There is at least one RR algorithm which is ranked differently based on its average turnaround time compared to at least one other algorithm.

Average Waiting Time:

- $\mathbf{H_0}$: There are no RR algorithms which are ranked differently based on their average waiting time compared to the other algorithms.
- $\mathbf{H_A}$: There is at least one RR algorithm which is ranked differently based on its average waiting time compared to at least one other algorithm.

Average Response Time:

- $\mathbf{H_0}$: There are no RR algorithms which are ranked differently based on their average response time compared to the other algorithms.
- $\mathbf{H_A}$: There is at least one RR algorithm which is ranked differently based on its average response time compared to at least one other algorithm.

Number of Context Switches:

- $\mathbf{H_0}$: There are no RR algorithms which are ranked differently based on their number of context switches compared to the other algorithms.
- $\mathbf{H_A}$: There is at least one RR algorithm which is ranked differently based on its number of context switches compared to at least one other algorithm.

After determining whether the null hypotheses should be accepted or rejected, the sum of the ranks, the mean ranks, the differences of the mean ranks, and the critical value for the difference of the mean ranks were calculated just as in the Adaptive Percentile algorithms test.

5.5 *Results for Adaptive Percentile Algorithms*

Table 2 shows the sum of ranks and the mean rank for the Adaptive Percentile algorithms for the average turnaround time. For any two algorithms, the closer the sums and the closer the averages, the more closely the algorithms ranked. The Chi-squared (χ^2 value yielded by the Friedman test for the average turnaround time was 68.34. The critical χ^2 value at a 95 % confidence level for 10 degrees of freedom is 18.307. The differences in the mean ranks between the Adaptive Percentile algorithms and Adaptive80 RR's mean rank are also given in Table 2. The critical value for the difference of mean ranks was 0.58. The critical value serves as the cutoff line; any difference that is greater than the critical value indicates that the two algorithms ranked differently.

Table 3 shows the sum of ranks and the mean rank for the Adaptive Percentile algorithms for the average waiting time as well as the differences in the mean ranks between the Adaptive Percentile algorithms and Adaptive80 RR's mean rank. The χ^2 value yielded by the Friedman test for the average waiting time was 68.34. The critical value for the difference of mean ranks was 0.58.

Table 4 shows the sum of ranks and the mean rank for the Adaptive Percentile algorithms for the average response time as well as the differences in the mean

Table 2 Rank summary for average turnaround time for adaptive percentile algorithms

Adaptive percentile algorithm	Sum of ranks	Mean rank	Difference in mean rank with Adaptive80
Adaptive0	77	11	8.43[a]
Adaptive10	70	10	7.43[a]
Adaptive20	63	9	6.43[a]
Adaptive30	56	8	5.43[a]
Adaptive40	45	6.43	3.86[a]
Adaptive50	46	6.57	4[a]
Adaptive60	33	4.71	2.14[a]
Adaptive70	20	2.86	0.29
Adaptive80	18	2.57	n/a
Adaptive90	27	3.86	1.29[a]
Adaptive100	7	1	1.57[a]

[a]Value is greater than the critical value

Table 3 Rank summary for average waiting time for adaptive percentile algorithms

Adaptive percentile algorithm	Sum of ranks	Mean rank	Difference in mean rank with Adaptive80
Adaptive0	77	11	8.43[a]
Adaptive10	70	10	7.43[a]
Adaptive20	63	9	6.43[a]
Adaptive30	56	8	5.43[a]
Adaptive40	45	6.43	3.86[a]
Adaptive50	46	6.57	4[a]
Adaptive60	33	4.71	2.14[a]
Adaptive70	20	2.86	0.29
Adaptive80	18	2.57	n/a
Adaptive90	27	3.86	1.29[a]
Adaptive100	7	1	1.57[a]

[a]Value is greater than the critical value

ranks between the Adaptive Percentile algorithms and Adaptive80 RR's mean rank. The χ^2 value yielded by the Friedman test for the average response time was 70.00. The critical value for the difference of mean ranks was 0.

Table 5 shows the sum of ranks and the mean rank for the Adaptive Percentile algorithms for the number of context switches as well as the differences in the mean ranks between the Adaptive Percentile algorithms and Adaptive80 RR's mean rank. The χ^2 value yielded by the Friedman test for the average response time was 70.00. The critical value for the difference of mean ranks was 0.

Table 4 Rank summary for average response time for adaptive percentile algorithms

Adaptive percentile algorithm	Sum of ranks	Mean rank	Difference in mean rank with Adaptive80
Adaptive0	7	1	8[a]
Adaptive10	14	2	7[a]
Adaptive20	21	3	6[a]
Adaptive30	28	4	5[a]
Adaptive40	35	5	4[a]
Adaptive50	42	6	3[a]
Adaptive60	49	7	2[a]
Adaptive70	56	8	1[a]
Adaptive80	63	9	n/a
Adaptive90	70	10	1[a]
Adaptive100	77	11	2[a]

[a]Value is greater than the critical value

Table 5 Rank summary for number of context switches for adaptive percentile algorithms

Adaptive percentile algorithm	Sum of ranks	Mean rank	Difference in mean rank with Adaptive80
Adaptive0	77	11	8[a]
Adaptive10	70	10	7[a]
Adaptive20	63	9	6[a]
Adaptive30	56	8	5[a]
Adaptive40	49	7	4[a]
Adaptive50	42	6	3[a]
Adaptive60	35	5	2[a]
Adaptive70	28	4	1[a]
Adaptive80	21	3	n/a
Adaptive90	14	2	1[a]
Adaptive100	7	1	2[a]

[a]Value is greater than the critical value

5.6 Results for Improved RR Algorithms

Table 6 shows the sum of ranks and the mean rank for the improved RR algorithms for the average turnaround time. The Chi-squared (χ^2) value yielded by the Friedman test for the average turnaround time was 124.39. The critical χ^2 value at a 95 % confidence level for 6 degrees of freedom is 12.592. The differences in the mean ranks between the improved RR algorithms and Adaptive80 RR's mean rank are also given in Table 6. The critical value for the difference of mean ranks was 0.56.

Table 6 Rank summary for average turnaround time for improved RR algorithms

RR algorithm	Sum of ranks	Mean rank	Difference in mean rank with Adaptive80
Standard	167	5.96	3.36[a]
AN	152	5.43	2.82[a]
Adaptive	93	3.32	0.71[a]
Efficient	29	1.04	1.57[a]
Optimized	105	3.75	1.14[a]
Priority-based	165	5.89	3.29[a]
Adaptive80	73	2.61	n/a

[a]Value is greater than the critical value

Table 7 shows the sum of ranks and the mean rank for the improved RR algorithms for the average waiting time as well as the differences in the mean ranks between the improved RR algorithms and Adaptive80 RR's mean rank. The χ^2 value yielded by the Friedman test for the average waiting time was 121.90. The critical value for the difference of mean ranks was 0.58.

Table 8 shows the sum of ranks and the mean rank for the improved RR algorithms for the average response time as well as the differences in the mean ranks between the improved RR algorithms and Adaptive80 RR's mean rank. The

Table 7 Rank summary for average waiting time for improved RR algorithms

RR algorithm	Sum of ranks	Mean rank	Difference in mean rank with Adaptive80
Standard	166	5.93	3.32[a]
AN	151	5.39	2.79[a]
Adaptive	93	3.32	0.71[a]
Efficient	29	1.04	1.57[a]
Optimized	108	3.86	1.25[a]
Priority-based	164	5.86	3.25[a]
Adaptive80	73	2.61	n/a

[a]Value is greater than the critical value

Table 8 Rank summary for average response time for improved RR algorithms

RR algorithm	Sum of ranks	Mean rank	Difference in mean rank with Adaptive80
Standard	170	6.07	2.36[a]
AN	145.5	5.20	1.48[a]
Adaptive	73	2.61	1.11[a]
Efficient	30	1.07	2.64[a]
Optimized	90	3.21	0.50
Priority-based	171.5	6.13	2.41[a]
Adaptive80	104	3.71	n/a

[a]Value is greater than the critical value

Table 9 Rank summary for number of context switches for improved RR algorithms

RR algorithm	Sum of ranks	Mean rank	Difference in mean rank with Adaptive80
Standard	109.5	3.91	0.09
AN	145	518	1.18[a]
Adaptive	176	6.29	2.29[a]
Efficient	74.5	2.66	1.34[a]
Optimized	63.5	2.27	1.73[a]
Priority-based	103.5	3.70	0.30
Adaptive80	112	4	n/a

[a]Value is greater than the critical value

χ^2 value yielded by the Friedman test for the average response time was 128.72. The critical value for the difference of mean ranks was 0.53.

Table 9 shows the sum of ranks and the mean rank for the improved RR algorithms for the number of context switches as well as the differences in the mean ranks between the improved RR algorithms and Adaptive80 RR's mean rank. The χ^2 value yielded by the Friedman test for the number of context switches was 69.05. The critical value for the difference of mean ranks was 0.84.

6 Conclusion Remarks

For each of the eight tests, based on the χ^2 results, there is enough evidence to reject each null hypothesis at the 95 % confidence level meaning that there must be at least one algorithm in each test which is ranked differently compared to at least one algorithm.

6.1 Average Turnaround Time

The results in Table 2 show that Adaptive80 performed better than Adaptive0 through Adaptive60 and Adaptive90, equal to Adaptive70, and worse than Adaptive100 for the average turnaround time. The results in Table 6 show that Adaptive80 performed better than Standard RR, AN RR, Adaptive RR, Optimized RR, and Priority-based RR and worse than Efficient RR for the average turnaround time.

6.2 Average Waiting Time

The results in Table 3 show that Adaptive80 performed better than Adaptive0 through Adaptive60 and Adaptive90, equal to Adaptive70, and worse than

Adaptive100 for the average waiting time. The results in Table 7 show that Adaptive80 performed better than Standard RR, AN RR, Adaptive RR, Optimized RR, and Priority-based RR and worse than Efficient RR for the average waiting time.

6.3 Average Response Time

The results in Table 4 show that Adaptive80 performed better than Adaptive90 and Adaptive100 and worse than Adaptive0 through Adaptive70 for the average response time. The results in Table 8 show that Adaptive80 performed better than Standard RR, AN RR and Priority-based RR, equal to Optimized RR, and worse than Adaptive RR and Efficient RR for the average response time.

6.4 Number of Contact Switches

The results in Table 5 show that Adaptive80 performed better than Adaptive0 through Adaptive70 and worse than Adaptive90 and Adaptive100 for the number of context switches. The results in Table 9 show that Adaptive80 performed better than AN RR and Adaptive RR, equal to Standard RR and Priority-based RR, and worse than Efficient RR and Optimized RR for the number of context switches.

6.5 Summary

Among all of the RR algorithms tested in both groups, Adaptive80 RR performed better than the vast majority of them for the average turnaround time and average waiting time. It performed worse than roughly two-thirds of them for the average response time, and it performed better than roughly two-thirds of them for the number of context switches. These results indicate that if average response time is not a high priority while the average turnaround time, average waiting time, and number of context switches are high priorities, then Adaptive80 is a very appealing and effective RR algorithm to use. If average response time is the highest priority, then Adaptive80 is not the most effective algorithm to use.

Acknowledgments This work was supported in part by Shippensburg University under the Student/Faculty Research Engagement (SFRE) Grant.

References

1. Noon A, Kalakech A, Kadry S (2011) A new round robin based scheduling algorithm for operating systems: dynamic quantum using the mean average. Int J Comput Sci 8(1):224–229
2. Singh A, Goyal P, Batra S (2010) Optimized round robin scheduling algorithm for CPU scheduling. Int J Comput Sci Eng 02(07):2383–2385
3. Silberschatz A, Galvin PB, Gagne G (2011) Operating systems concepts essentials, 7th edn. Wiley, USA
4. Rajput IS, Gupta D (2012) A priority based round robin CPU scheduling algorithm for real time systems. Int J Innovations Eng Technol 1(3):1–11
5. Sindhu M, Rajkamal R, Vigneshwaran P (2010) An optimum multilevel CPU scheduling algorithm. In: International conference on advances in computer engineering, 2010
6. Sukumar Babu B, Neelima Priyanka N, Sunil Kumar B (2012) Efficient round robin CPU scheduling algorithm. Int J Eng Res Dev 4(9):36–42
7. Helmy T, Dekdouk A (2007) Burst round robin as a proportional-share scheduling algorithm. In: 4th IEEE GCC conference & exhibition: towards techno-industrial innovations, Nov 2007
8. Mostafa SM, Rida SZ, Hamad SH (2010) Finding time quantum of round robin CPU scheduling algorithm in general computing systems using integer programming. Int J Res Rev Appl Sci 5(1):64–71
9. Behera HS, Mohanty R, Sahu S, Bhoi SK (2011) Design and performance evaluation of multi cyclic round roubin (MCRR) algorithm using dynamic time quantum. J Global Res Comput Sci col 2(2):48–53
10. Hiranwal S, Roy KC (2011) Adaptive round robin scheduling using shortest burst approach based on smart time slice. Int J Comput Sci Commun 2(2):319–323
11. McGuire C, Lee J (2014) Comparisons of improved round robin algorithms. In: Lecture notes in engineering and computer science: proceedings of the world congress on engineering and computer science 2014, WCECS 2014, San Francisco, USA, 22–24 Oct 2014, pp 158–161
12. Matarneh RJ (2009) Self-adjustment time quantum in round robin algorithm depending on burst time of the now running processes. Am J Appl Sci 6(10):1831–1837
13. Behera HS, Mohanty R, Sahu S, Bhoi SK (2011) Comparative performance analysis of multi-dynamic time quantum round robin (mdtqrr) algorithm with arrival time. Indian J Comput Sci Eng 2(2)
14. Behera HS, Mohanty R, Nayak D (2010) A new proposed dynamic quantum with re-adjusted round robin scheduling algorithm and its performance analysis. Int J Comput Appl 5(5):10–15
15. Mohanty R et al (2010) Design and performance evaluation of a new proposed shortest remaining burst round robin (SRBRR) scheduling algorithm. Institute of Engineering and Technology
16. Panda SS, Bhoi SK (2012) An effective round robin algorithm using min-max dispersion measure. Int J Comput Sci Eng 4(01)
17. Banerjee P, Banerjee P, Dhal SS (2012) Comparative performance analysis of average max round robin scheduling algorithm (AMRR) using dynamic time quantum with round robin scheduling algorithm using static time quantum. Int J Innovative Technol Exploring Eng 1 (3):56–62
18. Surendra Varma P (2013) A finest time quantum for improving shortest remaining burst round robin (SRBRR) algorithm. J Global Res Comput Sci 4(3):10–15
19. Sukumar Babu B, Neelima Priyanka N, Suresh Varma P (2012) Optimized round robin CPU scheduling algorithm. Global J Comput Sci Technol 12(11):21–25
20. Kumar N, Nirvikar (2013) Performance improvement using CPU scheduling algorithm-SRT. Int J Emerging Trends Technol Comput Sci 2(2):110–113
21. Stricker D (2014) Friedman. BrightStat.com (online) 2014. Accessed 18 June 2014
22. Lowry R (2014) Subchapter 15a. The Friedman test for 3 or more correlated samples. Concepts and applications of inferential statistics, (online) 2013. Accessed 18 June 2014

Intentional Agents

Nandan Parameswaran and Pani N. Chakrapani

Abstract Agents have mental attitudes such as belief, goal, commitment, plan, intention, etc. In this chapter, we propose a type of agent called Intentional Agents and discuss complex intention structures as the mental attitude of these agents. We consider intention structures for several forms of actions such as simple action, conditional action, repetitive action, concurrent actions, etc. We also extend this idea to mental actions in the form of meta intentions. Intentions are useful in reasoning with actions in future once they are committed to, particularly while (a) responding to unexpected events in the world with regard to an intended action, and (b) executing the action at the best available opportunity. However, complex intention structures result in computational overhead called mental effort for an agent. We measure the mental effort of an agent in terms of its intentions over time in different shopping scenarios and observe that repetitive (loop based) actions result in larger mental effort compared to simple and conditional actions.

Keywords Agents · Commitment · Intention · Intention structure · Mental effort · Mental states · Meta intention · Option

1 Introduction

Agents perform actions to achieve their goals either in response to an event in the world or based on the content of their mental states. The former is referred to as reactive problem solving while the latter is often classified as deliberative problem solving.

N. Parameswaran
School of Computer Science and Engineering, The University of New South Wales, Sydney 2052, Australia
e-mail: paramesh@cse.unsw.edu.au

P.N. Chakrapani (✉)
Department of Mathematics and Computer Science, University of Redlands, Redlands, CA 92373, USA
e-mail: pani_chakrapani@redlands.edu

H.K. Kim et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-94-017-7236-5_18

The use of intention in deliberative problem solving has been well researched in the AI literature, for example, under the sub area of agent architecture as BDI agents [1]. Intention intuitively refers to "a course of action that one intends to follow" [2] where the term "intend" refers to the mental disposition of having a plan or a goal for the specific purpose of achieving it over the course of time. For example, the statement "I intend to go to a restaurant" refers to the current mental disposition of the agent where the agent will initiate the action of going to the restaurant over the course of time. In this chapter, we suggest that intentions are of different types and may have complex structures.

In Sect. 2 we introduce the notion of intention and define Intentional Agents discussing the role of intention in their action execution behavior. In Sect. 3 we present several intention structures and discuss how they are maintained over time as actions are executed. In Sect. 5 we discuss the intention structure specific to a shopping scenario, and propose a measure for the overhead that occurs while maintaining complex intention structures during an agent's shopping activities. Section 7 is related work where we review similar work in the literature. Section 8 is suggestion for future work and conclusion.

2 Intention

Mental state attributes of an agent often include beliefs, goals, plans, intentions, choices, etc. Intention refers to the mental disposition of the agent that is committed to performing a task. The following definition of intention provides a simple characterization of intention which is adequate for our treatment.

Definition 1 An agent is said to have an intention I(a1) at time t if the agent is committed at time t to perform the action a1 sometime in the future.

Figure 1 shows the states of the world and how transitions among states occur when actions are performed. Suppose the world to begin with is at state s0. At this point, the agent has two options: perform action a0 or perform action a4. The agent that is committed to performing the action a0 is said to have the intention I(a0). The intention I(a0) also means that the agent is committed to performing a0 and soon will be selecting the option a0 and executing a0. In this case, the execution of a0 commences at the next time instant.

In Fig. 1 above, we also show several options at a few selected states where C0, C1, etc. refer to agent commitments. If the agent adopts the intention I(a4) at t = 0, it then results in the commitment C0 where the agent is committed to the action a4. Similarly, intention I($\langle$a1; a2; a3$\rangle$) at s0 results in committing to the sequence of actions in C1. Sometimes the agent can commit to actions in more than one option. For example, the commitment C2 corresponds to an intention of the conditional action of the form *if c do a1*. C3 depicts a commitment to a sequence of such conditional actions where action a1 is repeatedly performed as long as the condition c is true, and when c is not true, the action *nil* is executed once. This corresponds to the intention I(*while c do a1*). These examples illustrate the point that an agent can

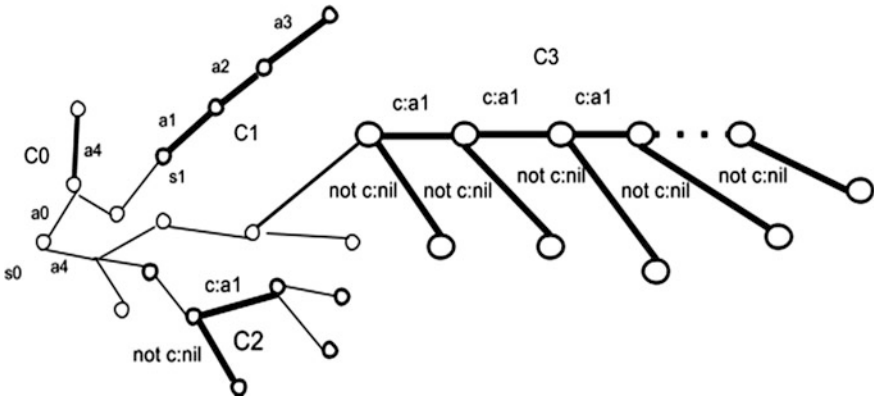


Fig. 1 Commitments in state diagram. C1, C2, and C3 show some of the commitments

commit to complex actions and the corresponding intentions can have complex structures. Once a commitment is made, the agent would take efforts to fulfil its commitment within a reasonable time. Intention exists always at the present; however, an agent can have an intention to perform an action at a specific point of time in the future.

2.1 Intention Cycle

Figure 2 shows a model of action execution by an agent A1 based on intention in the world W. To start with, the agent has no intention, that is, it has a null intention, denoted as I_{null}; at t = t0, the agent adopts the intention I(a1). At this point, the agent has made a commitment to itself that it will attempt to execute a1 (within a reasonable time). After sometime, at t = t4, the agent finds it possible to execute the

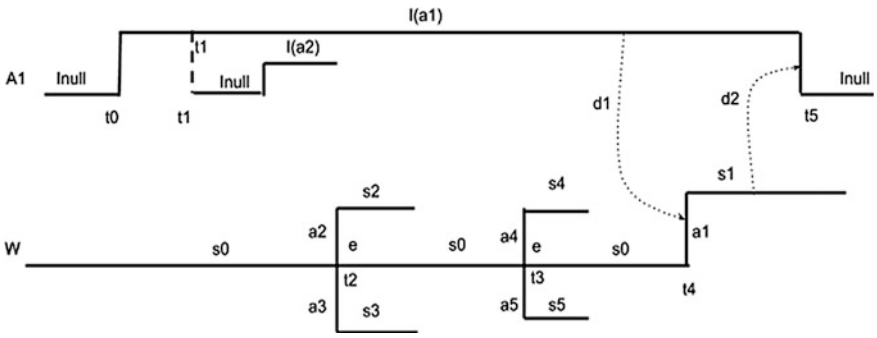


Fig. 2 Intention cycle

action a1 and executes a1 which changes the world state from s0 to s1. Since the action has been successfully executed, the agent drops its intention as the world stabilizes at state s1 at time t = t5. We call this an intention cycle. Note that at time t = t1, the agent had the option of choosing to adopt I(a1) or I(a2) after dropping I (a1); but, it chose I(a1) and completed its cycle at t = t5. At t = t2, the world W would have changed its state from s0 to s2 or s3 if the agent had executed a2 or a3, but since the agent chose not to execute a2 or a3, the world continued to exist in the same state s0 (corresponding to an empty event e). Similarly at t = t3, since no actions were performed by the agent, the world continued to exist in its old state s0 until t = t4.

2.2 Forced Intention and Autonomous Intention

Often intention is based upon choice and such intentions are called autonomous intentions. For example, the intention I(a0) at s0 in Fig. 1 is an autonomous intention since the agent could have chosen to execute a4 and formed I(a4) instead of I(a0). However, consider the state s1. At this state, the only option the agent has is to choose and execute a1, and thus the only intention the agent can form at s1 is I (a1). Such an intention is called a *forced intention*.

3 Intentional Agents

An intentional agent is an agent that performs each one of its actions based on its intentions only; that is, it does not perform any action that it is not committed to earlier via its intentions. Thus our agent is not capable of reactive actions.

3.1 Intention Structures

In complex activities, agents may need to adopt complex intention structures. We discuss below some of the intention structures that are useful in building complex intentions. An intention structure is built using intentions and sub-intentions.

Sub-Intention

Let x and y be two sequences of actions and let w_x and w_y be the corresponding sequences of states that are traced during the execution of the actions in x and y, respectively. The intention I(x) is a sub-intention of the intention I(y), denoted I (x) ≤ I(y), if w_x is a subsequence of w_y.

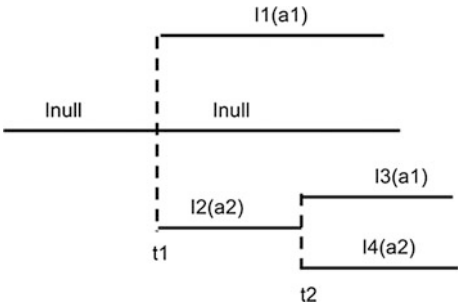


Fig. 3 Intention structure with options

3.1.1 Intention Structure for Single Action

Figure 2 shows the intention structure for a single action a1 where the agent's initial intention is I_{null}. In this, the agent adopts an intention I(a1) at time t0 and initiates the execution of the action a1. Once the action a1 is executed successfully at t4 and the world state stabilizes in s1, the agent drops its intention I(a1) at t5 and its intention becomes I_{null}. Arcs d1 and d2 denote instances of dependency relations.

3.1.2 Intention Structures with Options

Figure 3 shows an intention structure where the agent has an option to drop its current intention and choose another intention. At t = t1, the agent can change its intention or retain its current intention I_{null}. At t = t2, the agent does not have the option of retaining its current intention. Options in intention structures can be used to model intentions for conditional actions.

3.1.3 Intention Structure for Concurrent Actions

Figure 4 shows the intention structure model for a pair of concurrent actions a1 and a2 denoted as a1‖ a2. I(a1‖a2) gives rise to two concurrent intentions I(a1) and I(a2)

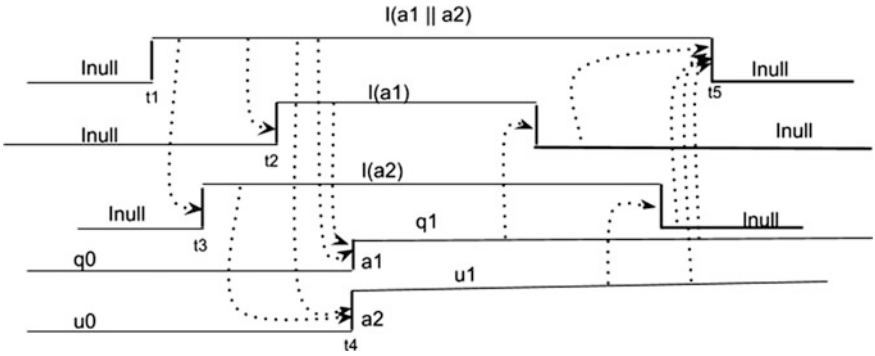


Fig. 4 Intention structure for concurrent actions a1 and a2

which cause the execution of the actions a1 and a2 in the *context* of I(a1‖a2). After dropping I(a1) and I(a2), the parent intention I(a1‖a2) itself is dropped once a1 and a2 are executed successfully.

3.1.4 Intention Structure for Repetitive Actions

Figure 5 shows the intention structure for an action a1 executed n times. To begin with, the intention structure shows two options: adopt the intention I(repeat n do a1) if $n > 0$ or adopt I_{null} if $n = 0$. Rewriting I(repeat n do a1) when $n > 0$ as I(a1; repeat $n - 1$ do a1) gives rise to two distinct sub-intentions: I(a1) and I(repeat $n - 1$ do a1) if $n - 1 > 0$ and I_{null} if $n - 1 = 0$. Once I(a1) results in the successful execution of a1, the agent proceeds to realize its pending intentions until no more intentions are left to realize. At this point the agent drops all intentions of the form I(repeat k do a1) where $n \leq k \leq 1$.

3.1.5 Exception Handling in Intention Structures

An exception is said to have occurred when an intended action fails to execute successfully or when an intention in the intention structure is terminated before the action it was intended for was successfully completed. The exception is propagated upwards to the parent intentions.

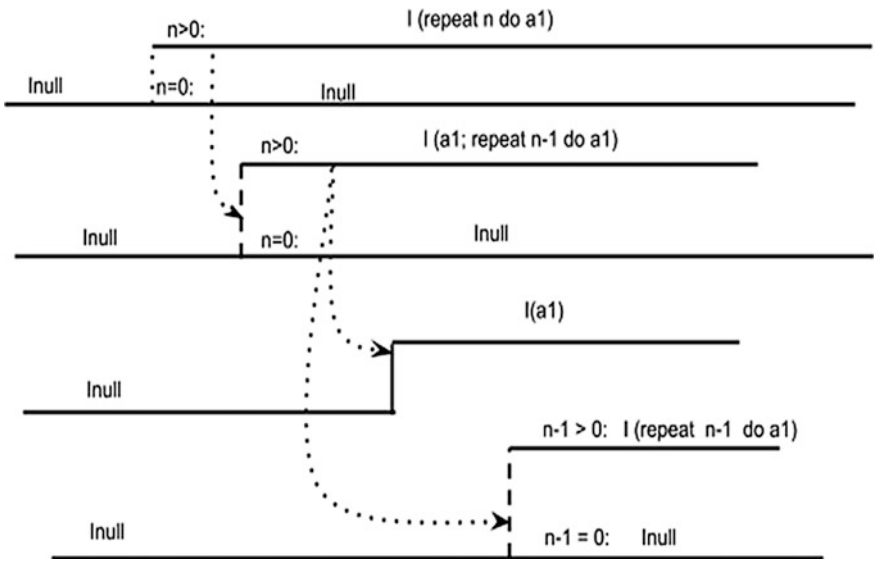


Fig. 5 Intention structure for repetitive execution of action a1

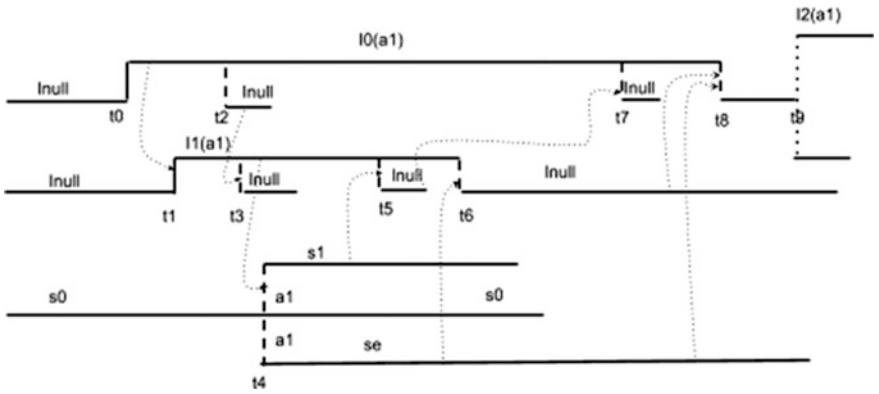


Fig. 6 Exception handling at different levels of intention structure in a complex action execution scenario

Figure 6 shows how an exception is handled at different levels of an intention structure; it shows three scenarios where intentions are propagated from lower levels to upper levels when an exception arises. The agent which holds a null intention for t < t0, adopts the intention I0(a1) at t0 to execute the action a1. It chooses to execute a1 by adopting another intention I1(a1) at t1. In scenario 1, the agent abruptly drops I0(a1) at t2 which results in dropping I1(a1) at t3, and the world continues to exist in its initial state s0. In scenario 2, the agent in order to realize I1(a1) attempts to execute a1 at t4 which successfully changes the world state from s0 to s1. This results in dropping I1(a1) at t5 and subsequently dropping I0(a1) at t7, and the world stabilizes in s1. In scenario 3, the agent while attempting to realize I1(a1) attempts to execute a1 at t4 but an exception occurs, and the world state changes from s0 to an error state se. In response to this, the agent drops I1(a1) at t6 and subsequently I0(a1) at t8, and the world continues to remain in the error state se. Later at t9, the agent has an option to attempt to execute a1 again by adopting another intention I2(a1).

3.2 Meta Intentions

Meta intentions refer to intentions about intentions. In this, an agent intends to intend to perform an action. Thus, in Fig. 7, I0(I1(a1)) is a meta intention where the agent intends to perform an object intention I1(a1) to perform the action a1. Meta intentions may be viewed as providing an awareness about mental action executional behavior. For example, the statement "I intend to buy a car this month," depicts the agent's intention I(buy_car_this_month) to buy a car this month. This implies that the agent is now committed to actively looking for opportunities to buy a car (perhaps) as early as possible within this month. Now consider the meta

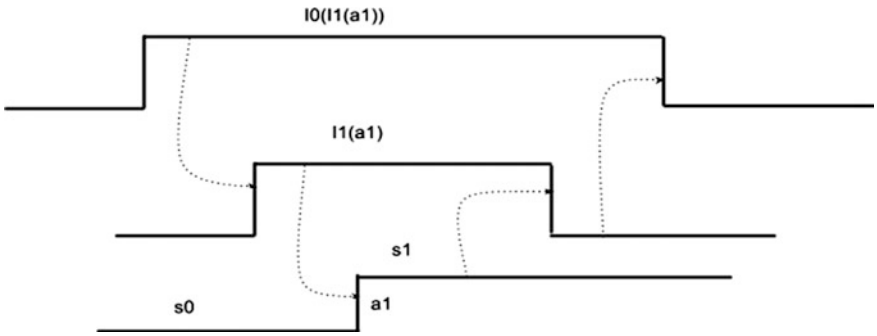


Fig. 7 Meta intention I0 and its object intention I1

intention in "I intend to intend to buy a car this month" which is denoted as I0(I1 (buy_car_this_month)). In this, the meta intention provides an opportunity to reason about the object level intention I(buy_car_this_month) and manipulate it in response to changing situations.

3.3 Incomplete Intention Cycles

Sometimes an intention cycle may not be as complete as the one shown in Fig. 2. Figure 8 shows three scenarios in which the intention cycle is not complete. In Fig. 8a the agent may save some time by assuming that the execution it initiated at t1 will be successful and dropping I(a1) at t2 without waiting for the successful completion of a1. Figure 8b depicts a scenario where the agent does not even bother to drop its intention after initiating a1. This situation may be acceptable only when the agent soon terminates all its parent and meta intentions so that I(a1) does not

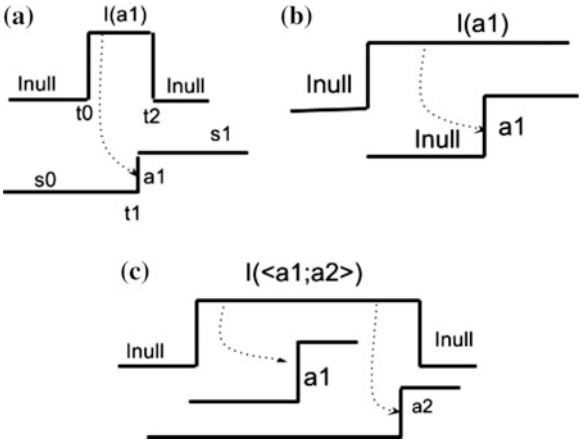


Fig. 8 Incomplete intention cycle: three scenarios

result in executing a1 once again. Figure 8c depicts the scenario where the agent is committed to executing a plan $\langle a1; a2\rangle$ where the agent saves time by only initiating actions and not waiting for their completion of execution. Strategies such as this will be useful when a master agent uses its slave agents to achieve the execution of its plan.

4 Intentions and mental effort

The presence of intentions in an agent's mental state demands that the agent do more work. Since the presence of an intention results in a commitment, the agent has to do extra work to honor the commitment. This typically includes choosing appropriate options at every point of reasoning so as to maximize the chances of successfully executing the intended action. We measure the mental effort em(I(x), τ) due to an intention I(x), where x is an action, by an amount proportional to the duration τ over which the intention was maintained.

4.1 Mental Effort Due to Single Intention

Let I(a1) be an intention which was adopted at t1 and dropped at t2. Thus, the life time of the intention τ is given by $\tau = t2 - t1$. The mental effort due to this intention is given as em(I(a1), τ) = k*τ where k is some constant. This model assumes that the reasoning that the agent performs at any time throughout the life time of the intention is constant and does not vary due to any exceptions. (This is a strong assumption, but nevertheless we use it to simplify the analysis.)

4.2 Mental Effort in Concurrent Intentions

For a sequence of two actions $\langle a1; a2\rangle$ the mental effort can be defined as:em(I($\langle a1; a2\rangle$), τ) = em(I(a1), τ1) + em(I(a2), τ2), where τ, τ1 and τ2 are the life times of the intentions I($\langle a1; a2\rangle$), I(a1) and I(a2) respectively. We can now generalize this result to n actions as follows.

$$\mathrm{em}(\mathrm{I}(\langle a_1; \ldots; a_n\rangle), \tau) = \sum (\mathrm{em}(\mathrm{I}(a_i), \tau_i), \ 1 \leq i \leq n.$$

Assuming that em(I(a_i), τ_i) = p for some constant p for all i, em(I($\langle a1; \ldots a_n\rangle$), τ) can be shown to be $O(n^2)$ for concurrent intentions strategy.

4.3 Mental Effort in Sequential Intention

While it is not clear when exactly a human forms her intention in the course of her problem solving activity, we can design agents that use different types of strategies for adopting and dropping intentions.

In the cases we presented above, the agents we used adopted all their intentions and held them concurrently until they were dropped. Concurrent intention strategy is expensive in terms of overhead since the agent has to perform intention related reasoning for more than one intention at any time. In order to reduce the overhead, the agent can sequentialize the concurrent intentions. In such a case, the total mental effort for a sequence of actions $\langle a1; \ldots an \rangle$ can be reduced to a constant; that is, $em(I(\langle a1; \ldots an \rangle), \tau)$ can be reduced to $O(1)$. Thus, agents with sequential intentions are more efficient. However, this efficiency comes with a cost. In the sequential intention case, say one intention I1(ai) at a time, the agent can only perform the intention related reasoning for the action ai. Thus, for example, during this time, if another intention I2(aj) has to be dropped, the agent would not be able to do it since the agent's current mental state does not hold the intention I2(aj).

4.4 Mental Effort in Repetitive Intention

As we have discussed above, the number of intentions the agent has to maintain at any time will depend on the intention generation strategy the agent employs: either concurrent or sequential where the concurrent strategy results in excessive overhead and sequential strategy results in preventing the agent from reasoning with all intentions. For a repetitive action, we can use a combination of these two strategies. At the beginning of each iteration, the agent holds three intentions out of which one will be dropped after the execution of the action and two will be retained (see Fig. 5). To start with, the agent has one intention I_{null}. After some time, the agent adopts one intention I(repeat n do a1), assuming $n > 0$, which results in another intention $I(\langle a1;\ \text{repeat}\ n - 1\ \text{do}\ a1 \rangle)$. This eventually leads to two more intentions I (a1) and I(repeat $n - 1$ do a1). Thus, at this point, the agent has three intentions active of which the agent realizes one intention I(a1), and drops it after the successful execution of a1, leaving the agent two intentions I(repeat n do a1) and I (repeat $n - 1$ do a1). Similarly, I(repeat $n - 1$ do a1) results in the execution of a1 and another new intention I(repeat $n - 2$ do a1), and thus the agent has three intentions now. This process continues until the condition c becomes false. Assuming that c becomes false after n iterations, at the nth iteration, there will be n intentions, I(repeat $n - i$ do a1), $n \leq i \leq 0$, in the agent's mental state. Thus, the mental effort em(I(repeat n do a1, τ) where $\tau = n$, is proportional to n^2, which simplifies to $O(n^2)$. If we consider nested repetitions, nesting m times, then the mental effort em(I(repeat n do (repeat n do (…))),τ) is estimated as $O(n^{2m})$. Thus, nested loops are more expensive even in sequential intention model.

5 Mental Effort in a Shopping Scenario

In the simulation of the shopping scenario, we consider multiple agents performing shopping activities by moving from one shop to another. Each agent starts its shopping activity with a shopping list. The following figures show the mental efforts spent for different shopping scenarios (in NETLOGO [3] simulation). We consider three types of actions of buying an item: simple (Fig. 9), conditional (Fig. 10), and repetitive (Fig. 11), all using the sequential intention strategy [4]. The shopping lists were created randomly and the shops sold items from a randomized list of products. An agent begins with a large set of intentions, looking for opportunities to realize its immediate intentions first. The mental effort increases as time passes. The agent drops an intention when the item corresponding to that intention has been bought, and this results in the reduction of its mental effort at that point. Mental effort consumption behavior due to conditional actions is similar to

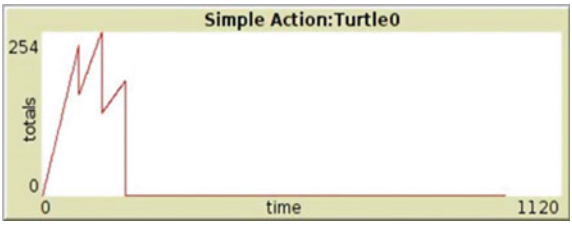


Fig. 9 Mental effort spent on a simple action [4]

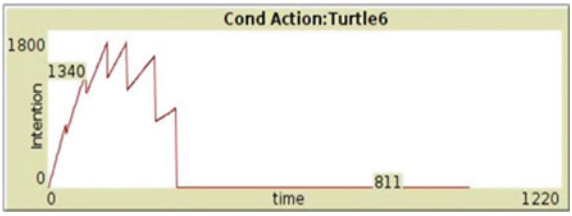
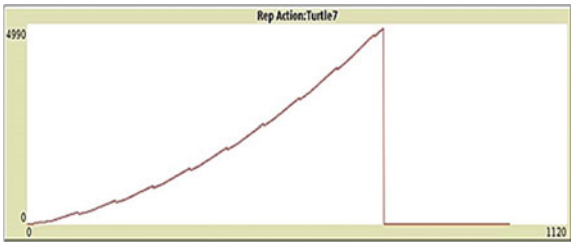


Fig. 10 Mental effort spent on a conditional action [4]

Fig. 11 Mental effort spent on repetitive action [4]



simple actions, but the rate is higher due to the condition c which results in more than one option. Repetitive actions require increasing mental effort as time passes since several intentions are needed to be maintained until the loop condition c becomes false.

6 Discussion

When an agent adopts an intention, the adoption typically occurs in a situation where the agent also has an option not to adopt the intention. Intention is about doing an action in future however close it is to the present. It is, however, useful to ask how early an intention should be formed in an agent's mental state. Adopting intentions too early can cause excessive overhead since typically an agent in a dynamic world performing complex tasks may have several hundred intentions at any time. It is thus advisable to delay the adoption of an intention. However, delaying the adoption of an intention can be problematic sometimes since the agent may risk the opportunity of executing the action in the right context where mental states of other agents and the world state might be favorable to action execution. When exactly an intention should be formed and how long should it exist in the mental state are domain dependent issues. An intention that is similar to meta intention was proposed by Grosz [5] and it is known as *intend that*. For example, "I intend that Mary buy a car" is an example of the intention *intend that*. In this example, Mary may form an intention of her own to buy a car. If Mary is a robot that is not capable cognitively forming an intention, the action will be executed without an intention formed in the mental state. The type of intentions we discussed so far in this chapter may be termed as *explicit intentions* which may be distinguished from *implicit intentions*. An agent is said to perform an action a1 with an implicit intention when the action execution is not directed by the agent's intention (represented in the mental state) cycle. This distinction is particularly important in a multi-agent scenario since an agent with an implicit intention may not be able to share its mental state with the other cooperating agents.

6.1 Uses of Intention

An action may be executed at various levels in an agent. For example, an agent may execute an action in response to an event in the world in a reactive mode. Sometimes, the execution of a complex action composed of several sub actions may be performed by a lower level module of an agent, such as a mechanical device. In a group of agents, a group action may be performed by an individual agent without a group intention. These are some of the examples where actions are not driven by an underlying explicit intention. In dynamic situations with rich options presenting themselves, an agent would find it profitable to let the actions be driven by

intentions. This is particularly so in a team activity where the intentions of the agents must be explicitly known to each other. An intention models the agent's executional awareness for an action and it is useful in communicating with the other agents during cooperation. It also provides a source of persistence in a dynamic world.

7 Related Work

Richard Sheer defines intention as a course of action which an agent has adopted for possible execution [6]. Bratman, perhaps for the first time, discussed the role of intention as the mental state agents hold for performing actions in future. He argues that intending to act is different from acting intentionally [7]. Groz extends this idea and proposes another type of intention called *intention-that*. She argues that it is possible for an agent to intend that some agent intend to perform a physical task [5]. Pollack has suggested in her seminal paper on plans as complex mental attitudes [8] how agents use plans not only as a recipe to achieve their goals, but also to reason about situations and cooperate with other agents. Hunsberger and Ortiz [9] present a theory of intention representation that provides solutions to representational problems in order to fill an important gap in existing theories of agents, planning and collaborative planning. Stone has attempted to formalize natural language grammatical knowledge using intentional structures [10] in discourses. In our work, we have used intention to model a restricted form of temporal awareness; this helps the agent to choose an appropriate time to perform actions based on its intentions.

8 Conclusion

In this chapter, we argued the need for explicitly modelling intentions in action execution. We also demonstrated that the procedural control abstractions can be used to derive complex yet well-defined intention structures. Tasks can be as simple as a set of primitive actions or a complex structure specified as a BPMN process [11]. To our knowledge, intentions in the context of interactive task execution has remained largely an unexplored area. One natural extension of our work will be to consider the characterization of cooperative tasks performed by multiple agents. In this domain, individual agents may spend considerable amount of mental effort in maintaining user intentions. Representing intentions in tasks provides opportunities for focusing on some parts of a task. This is particularly useful when the task structure is large such as in BPMN, UML, etc.

References

1. Georgeff RMP (1995) BDI-agents: from theory to practice. In: Proceedings of the first international conference on multiagent systems (ICMAS'95)
2. A free online English dictionary http://www.thefreedictionary.com
3. Netlogo https://ccl.northwestern.edu/netlogo
4. Parameswaran N, Chakrapani PN (2014) User intentions in interactive tasks. In: Lecture notes in engineering and computer science: proceedings of the world congress on engineering and computer science 2014, WCECS 2014, San Francisco, USA, 22–24 Oct 2014, pp 216–220
5. Grosz BJ, Sidner CL (1986) Attention, intentions, and the discourse structure. Comput Linguis 12(3):175–204
6. Scheer R (2004) The mental state theory of intentions. Philosophy 79(307):121–131
7. Bratman M (1984) Two faces of Intention. Philos Rev XCIII(3)
8. Pollack ME (1990) Plans as complex mental attitudes. In: Cohen PR, Morgan J, Pollack ME (eds) Intentions in communication. MIT Press, Cambridge
9. Hunsberger L, Ortiz CL Jr (2008) Dynamic intention structures I: a theory of intention representation. Auton Agent Multi-Agent Syst 16(3):298–326
10. Stone M (2004) Intention, interpretation and the computational structure of language. Center for Cognitive Science, Rutgers, The State University of New Jersey, Newark
11. White SA, Miers D (2008) BPMN modeling and reference guide: understanding and using BPMN. Future Strategies Inc.

Nonlinguistic Disaster Information Sharing System Using Visual Marks

Kakeru Kusano, Tomoko Izumi and Yoshio Nakatani

Abstract This work aims to support the actions of disaster victims, including tourists and persons who do not understand local language, via information sharing when disasters occur. Existing disaster support systems are often targeted at local residents, and understanding the local language is a prerequisite for use. In consideration of this situation, our system converts linguistic information to a nonlinguistic approach including maps, pictograms. We propose a system that can express disaster information visually, and operate information collection and provision intuitively without using languages. In this paper, we describe the prototype system, and show evaluation results to verify the effectiveness of our system. Moreover, we propose a solution for the problems which are indicated in the evaluation results. Specifically, an interface to collect ant provide multidimensional information is proposed in this paper.

Keywords Disaster information · Disaster prevention · Information sharing system · Mobile devices · Nonlinguistic · Pictograms

1 Introduction

Disaster prevention is one of the critical issues worldwide. In this section, we describe the present situation in Japan, and present the overview of our contributions.

K. Kusano (✉)
Graduate School of Information Science and Engineering, Ritsumeikan University,
1-1-1, Nojihigashi, Kusatsu, Shiga 525-8577, Japan
e-mail: is0013vi@ed.ritsumei.ac.jp

T. Izumi · Y. Nakatani
College of Information Science and Engineering, Ritsumeikan University,
1-1-1, Nojihigashi, Kusatsu, Shiga 525-8577, Japan
e-mail: izumi-t@fc.ritsumei.ac.jp

Y. Nakatani
e-mail: nakatani@is.ritsumei.ac.jp

H.K. Kim et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-94-017-7236-5_19

1.1 Background

Japan is affected by many disasters; in fact, the country is an area that suffers a concentration of disasters. Although this area is only 0.25 % of global landmass, it is the site of 20.5 % of all earthquakes of magnitude over 6.0, 7.1 % of all active volcanoes, and 16.0 % of disaster-related damage costs worldwide [1]. At the present time, various measures are in place for disaster mitigation and disaster reduction in, due to the influence of estimations of damage by a Nankai Trough Quake [1]. Moreover, during the Great East Japan Earthquake of 2011, social media played an important role in sharing disaster information among victims. Such utilization of social media during disasters, exemplified by Twitter, is gaining attention. A "Wisdom of Crowds" was constructed from multiple users' contributions (Tweets), and the effectiveness of this as an information infrastructure to ascertain the damage situation was recognized [2]. Interest in disaster mitigation and disaster reduction that utilizes such information infrastructure is increasing, due to the influence of anticipated major earthquakes and the disasters that have caused serious damage in recent years.

Meanwhile, Japan is aiming to make the country a global travel destination, which is being treated as important policy in the 21st century [3]. As an example, Kyoto city is visited by one million foreign tourists every year, and foreign visitors across Japan are expected to increase in the future. Paris is a famous model of an internationally competitive sightseeing city, visited by 45 million tourists every year, 60 % of whom are from overseas [4]. When a disaster strikes such sightseeing cities, many tourists will suffer from heavy damage. They are unfamiliar with the area, local language, culture and disasters. The safety of foreign tourists is critically important, and actions to protect them are the responsibility of the region and the nation.

However, the present information systems to support victims when disasters occur target only local residents, who are geographically familiar with the local area. That is, they require understanding of the local language. Tourists, who are likely to be geographically unfamiliar with the area, and non-local language speakers, tend to have difficulty in sharing and understanding information about the disaster, and to have greater damage than residents [5]. Therefore, a framework is required that can support all disaster victims, including persons predisposed to difficulty such as those described above.

1.2 Importance of Universal Disaster Mitigation

Methods to support disaster victims, including persons who cannot understand the local language and persons geographically unfamiliar with the area, by transmitting information visually and intuitively have been considered. Although multilingual system is one of the solutions to support foreigner disaster victims, there is no

standard of the languages which should be handled in the system, and information provision sides have a high load if they provide a large amount of information in the various languages with expedition during the occurrence of disasters.

These problems have been noticed by the Japanese FDMA (Fire and Disaster Management Agency). They are creating a universal design for disaster mitigation pictograms using such a method. Pictograms are diagrams that express meaning using color and shape, which can transmit the meaning of information without using language [6]. These pictograms are created in accordance with the principles of graphic symbols stipulated by JIS (Japanese Industrial Standards), and are registered by the ISO (International Organization for Standardization). In this way, universal methods to support disaster victims are starting to be considered in Japan.

The biggest advantage of this approach is to enable communication of information without the restraints of language. For example, signs such as 'emergency exit' and 'disabled access' are well known. In Japan, pictograms are adopted in places that are heavily used by the general public. Kunoki et al. of Ritsumeikan University considered the provision of unified design of disaster prevention pictograms at tourist sites [7]. They conducted an investigation into signs and guide plates at Kiyomizu Temple, a famous tourist attraction, as a target and suggested a method for improvement. The summary of their results is as follows:

1. Guide plates that display foreign languages or use pictograms are only about one quarter of the total.
2. Methods to express color, shape and content are different depending on the place. There is no sense of unity and this hinders understanding.

From the summary of results, Kunoki et al. advocated the necessity of offering information utilizing universal design to provide efficient evacuation guidance.

1.3 Our Contribution

In this research, we propose a disaster information sharing system that is able to collect and provide information visually and intuitively using GIS (Geographic Information System) and pictograms. We aim to construct a disaster information sharing system that supports disaster victims' decision-making, as well as their grasp of the situation in the surrounding area, by implementing an interface that enables disaster victims to post information easily when disasters occur. First, we constructed a system that makes use of zero-dimensional information that expresses information dealing with points. This paper shows the evaluation results to verify the effectiveness of our system. In the results, one problem is indicated, that is, we need to treat with multidimensional information in order to express target zones and area. Then, we propose an extended system which makes also use single dimensional information and two-dimensional information using lines and plane figures. There are currently no existing systems in which the user not only collects

multidimensional information about disasters, but can also post such information themselves.

In the next section, related work is described. We present outline of our first system and its evaluation results in Sect. 3, and in Sect. 4, we show our second system which makes use multidimensional information. Then evaluation experiment that we examining is presented in Sect. 5. Finally, a conclusion and future work are described.

2 Related Works

2.1 Analysis During the 2011 Tohoku Earthquake

In the 2011 Tohoku Earthquake, the Tokyo metropolitan area suffered from disaster for the first time after being developed as a modern city, and there were many people who had difficulty returning home. This is relevant to this study, as it is useful to design our system through analysis related to information media.

From the summary of a Cabinet Office survey conducted via the Internet, there were 5.15 million persons who had difficulty returning home in the metropolitan area, which includes Tokyo, Kanagawa, Chiba, Saitama and Ibaraki [5]. From the results of the survey into how people returned home, it was shown that the most common method of returning home was on foot, accounting for 37.0 % of the total victims, and the second most common method was by car. The reason for these being the most common is that trains ceased operation immediately after the earthquake. Disaster victims felt that certain information was necessary while returning home. In descending order, information on their families' safety, investigation of the damage, and time until the trains and subway would resume operation, were the most cited. If we exclude family safety information, we can see that victims need information about items required to return home.

The results of a survey conducted about future methods of obtaining information hoped for by disaster victims are also shown. Results are grouped for three conditions related to whether victims were capable of returning home. In particular, the provision of information by TV is strongly hoped for. This reason is considered to be that TV is capable of providing information visually, and that people watching TV can know an overview of the disaster in various places. The number of people who wish to be provided information by cell phone is high, because it is easy to use while on the move. In particular, there is a tendency for this to be desired by people who tried to return home but were prevented. This may be estimated as due to the necessity for information to decide a new course of action after being prevented from returning home.

2.2 Disaster Information Sharing Support System

Aoyama et al. proposed an information sharing system using WebGIS. Their system is a communication tool that is targeted at sightseeing spots [8]. Outside of disasters occurrence, businesses engaged in the tourism industry can use the system to provide tourism promotional information; during disasters, each of these businesses becomes a disaster shelter for tourists and local residents and transmits information. Their system assumes a wide range of users, including staff of local governments, residents, business personnel and tourists. A key feature of the system is the capability for users to mutually transmit information, whether during a disaster or in normal periods.

By dealing with information that is appropriate to the situation, Aoyama et al.'s system is grasped to play a role in ascertaining the local damage situation during the occurrence of a disaster. On the other hand, since it is necessary for users to input the location and details of the registered information in order to send such information, we can assume that the operation of the registration may be complicated for the victims because they must operate the information provision in times of emergency, i.e. the occurrence of a disaster, even if a wide range of users is assumed. Furthermore, understanding of the local language is a prerequisite of using Aoyama et al.'s system, so people who cannot understand the local language will have difficulty not only when providing information, but also when understanding the provided information. Accordingly, in our research we aim to resolve such problems by simplifying the system and providing information that have high visibility by using an intuitive interface.

2.3 Information Collection Support System for Disaster-Affected Areas Using Mini-Blogs

Yokobe and Nakatani proposed an information collection system using social media and mini-blogs to collect information when disasters occur [9]. Social media and mini-blogs have high real-time applicability, and are capable of responding to the fluid situations that arise when disasters occur. In this system, unnecessary and unreliable information is deleted in accordance with the following two assumptions.

1. Useful information for victims is sent from disaster-affected areas.
2. The reliability of information is evaluated by people in the area where the information is sent.

Based on these assumptions, Yokobe and Nakatani's system deletes information which is sent from areas geographically separated from the disaster affected area. Specifically, they assume that information with exact location data has high reliability, and the reliability of each Tweet and Twitter user is evaluated based on the location data and responses from people living around the location where the Tweet

is sent. From the results of their evaluation, it was demonstrated that Tweets that included geographical names had a strong relationship with the disaster-affected area. This fact implies that geographical names are important in deciding the reliability of Tweets.

However, most Tweets do not contain location data, and there is no such service in Twitter itself. Thus, the ascertainment of reliability based on location data is not available in many cases. Moreover, a problem inherent in Twitter is that when Tweets are sent by a huge number of users in a short time, all these Tweets are displayed on the Time Line, making it difficult for the user to comprehend all the information. As a vast amount of the latest information is displayed during the time taken by a user to read a single Tweet, there is a risk that information valuable to the user may be missed.

In October 2012, disaster prevention training utilizing social media was conducted in Japan [10]. In this training, disaster victims identified evacuation shelters that were posted on social media, and traveled to those evacuation shelters in reality. During the training, there was a major information gap in proportion to the level of comprehension when using an information system, due to problems such as users who were not familiar with dealing with social media being unable to understand where the evacuation shelter information was posted on social media, and users who could not understand how to use social media itself. In this way, the difficulty of using social media effectively during disaster situations was confirmed. It is thus necessary to make innovations such as providing an easily understandable interface to users in which large amounts of information are condensed before displaying. To resolve such problems, we attempt to utilize pictograms and GIS for input and output of information in this research.

3 Outline of Our Proposed First System

In this research, we aim to construct a system that performs both information collection and information provision when disasters occur, with all disaster victims as a target. This system will enable sharing of disaster information among disaster victims via implementation of an interface that allows information to be collected and provided visually and intuitively, even if users do not know how to operate the system.

3.1 System Functions

We constructed a system for sharing disaster information that converts linguistic information to GIS and pictograms, as a nonlinguistic approach [11]. By allocating pictograms on GIS, the system expresses information visually about the disaster situation at specific locations. Additionally, this system enables the user to execute

information provision intuitively by incorporating pictograms and gesture operation using a touchscreen display when users operate the system.

This system uses a client-sever model in which the client is a smart phone application and the server is a web application utilizing a database and PHP programs. The database stores the pictogram type, location (latitude and longitude), provided time, provider and pictogram ID inputted by the user.

The system initial screen is showed in Fig. 1. The user can collect information on their current surroundings via map information and icons on this screen. In the screen, the current position of disaster victim getting from network and GPS (Global Positioning System) is the center. For the surrounding area up to a 3 km radius, the pictograms are placed on the screen. By tapping an icon, the users can find out the provider and posted time, which can aid in judging the situation.

During disasters, the time and operations required to provide information should be minimal. In this system, users can provide with three steps, as shown in Fig. 1.

1. Change to "Information Provision" screen
2. Select the pictograms corresponding to the information the user wants to provide from the icons at the bottom of screen
3. Tap "Provide" Button

Via this operation, the client-system sends the server information that includes the current position, pictogram name, time of provision and pictogram ID. Disaster information to be provided by users is collected in the server.

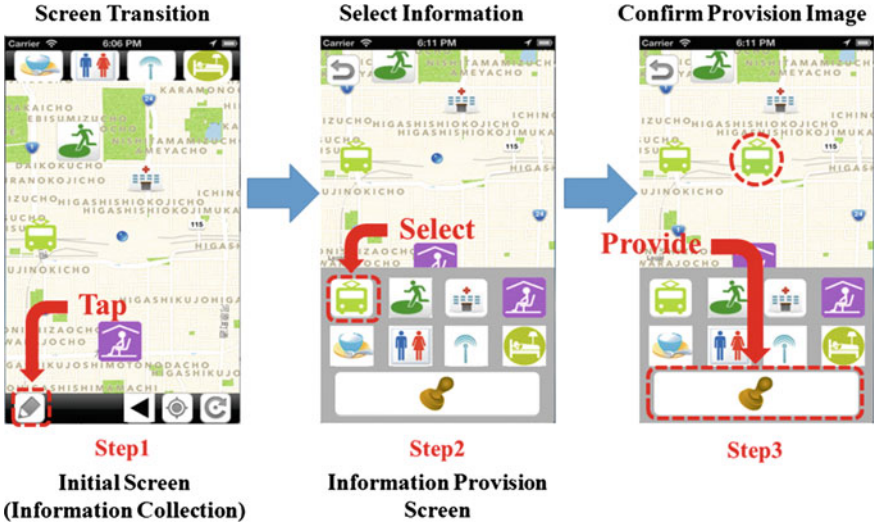


Fig. 1 The flow for providing information

4 System Evaluation Experiment

In the evaluation experiment, we conducted two separate trials: for disaster mitigation experts and for normal users. The expert group consisted of three persons from the Disaster Management Office of Kyoto City and twelve persons from Konan Regional Fire Administrative Organization, Shiga Prefecture as experts, performed simulated operation of the system and answered a questionnaire survey after we explained about the system. Two students from China and six students from Japan participated in the experiment as normal users, and answered the same survey without hearing an explanation of how to use this system. Three of the students had not used social media or smartphones before. The list of questions in the questionnaire is shown in Table 1.

The following is a summary of the evaluation by the experts.

1. Response to Question 1
   - There are big advantages for evacuating to shelters and obtaining information for returning home.
   - Information is available quickly, as it is expressed using easy-to-understand pictograms.
   - Users can be psychologically reassured, as they can gather information on their surrounding area.
   - Comprehension of the situation can be increased, as information about the destination is easy to understand.
2. Response to Question 2
   - Pictograms were appropriate (13 out of 15 experts).
   - Users should be able to change the size of the pictograms (2 out of 15 experts).
3. Response to Question 3
   - If users already have smart phones, the system could be operated without problems.
   - Limiting the amount of provided information leads to ease of use.
   - Mental barriers to using this system are higher for elderly people.

Table 1 The list of the questionnaire survey

Question number	Question contents
1	Do you think comprehension of the situation during a disaster would increase by using this system?
2	Were the sizes and designs of pictogram appropriate?
3	Was the system sufficiently easy to use?
4	Are the types of information provided sufficient?
5	Do you have other points of feedback or improvement?

4. Response to Question 4
    - Information about unavailable facilities, not only available facilities, is necessary.
    - Information about means of transport (operation status and area) is necessary.
5. Response to Question 5
    - Need accurate and safe user-provided information.
    - Setting display periods for pictograms would improve management of information.
    - Wish to utilize Wi-Fi spots during disaster situations, for example bus stops.

The following is a summary of the evaluation by non-Japanese users (Chinese).

- The provided types of information and size of pictogram are appropriate.
- It is easy to ascertain the situation in the surrounding area via the combination of GIS and pictograms.
- If we have experience of this system as social media during non-disaster times, we can use it when a disaster occurs.
- We could not understand the difference between "Evacuation Site" and "Place for Rest" only from the pictogram.

The following is a summary of the evaluation by general users.

- The provided information types, size of pictograms and denotation are suitable on the whole.
- It is easy to use this system because it only uses pictograms.
- If we have operated this system once, we can use when a disaster occurs.
- Because the display is based on maps, we could ascertain both the geography and situation of our surroundings simultaneously.
- We could understand the meaning of the information quickly, as there was no need to read text.
- The pictograms for "Evacuation Site", "Place for Rest" and "Available Accommodation" are too similar to distinguish. They should be revised or guidelines should be made.

Although some deficiencies in the provided information and comprehension problems for some of the pictograms were pointed out, we were able to confirm that this system can be used as an easy-to-operate disaster information sharing system through evaluation from experts, non-Japanese and general Japanese users. It is highly beneficial for disaster victims to obtain information during their evacuation or going back their home. By using our system, victims are able to obtain information immediately, because information is simply expressed by using pictograms and GIS. So, victims are able to know the situation the surrounding area.

On the other hand, subjects gave two suggestions for improving disaster information using the system: One is that the system should provide not only facilities which are not available, but also available ones. The other is that in the case of providing transportation information, the traffic section needs for disaster victims, not limited to traveling condition.

5 Multidimensional Information Sharing

In the first system, information is shared as zero-dimensional information on GIS. This zero-dimensional information showed single-point information that corresponded only to one specific coordinate. However, among the information required by disaster victims during a disaster, there exists some information that is difficult to express as zero-dimensional information. One typical case is the zones of available public transportation operation. Although this information is important for disaster victims in order to evacuate or return home, it is desirable to handle this information as single-dimensional information, as a line connecting the transport route from the starting station to the terminal station. Similarly, when describing areas of damage such as flooding on GIS, this information is expressed as a plane, in other words, two-dimensional information. Specifically, this can be expressed by colorizing the corresponding area, as shown in Fig. 2.

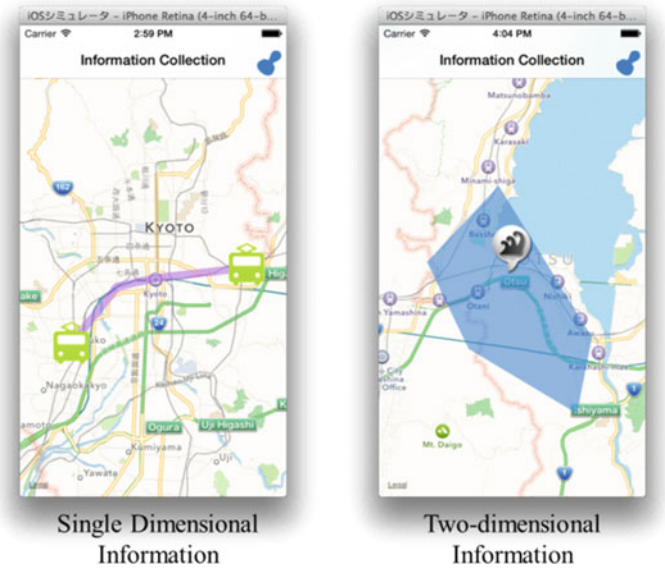


Fig. 2 Input/output of multidimensional information

In this section, we propose a multidimensional disaster information sharing system that not only provides single-dimensional information and two-dimensional information to show route and area, but also collects multidimensional information from disaster victims [12].

5.1 System Architecture

The system is consists of a client device, by which the user provides/collects information, and a server that responds to the client's requests. The database of the server application contains two tables. The first is used for saving zero-dimensional information, and the second for saving single-dimensional information and two-dimensional information. The zero-dimensional information represents single-point information, that is, it corresponds only to one specific coordinate. The single dimensional and two-dimensional information represent a route and a range respectively, that is, they must have multiple coordinates to represent the route or range. Thus these multidimensional information are saved in the another table of database.

The access to the database is executed by that PHP programs receiving HTTP requests of clients. The client@device transmits server requests, and thereafter the server processes data in accordance with the client's request, information is registered in the database, and users are provided with information from a server side system throughout client application's interface, as shown in Fig. 3.

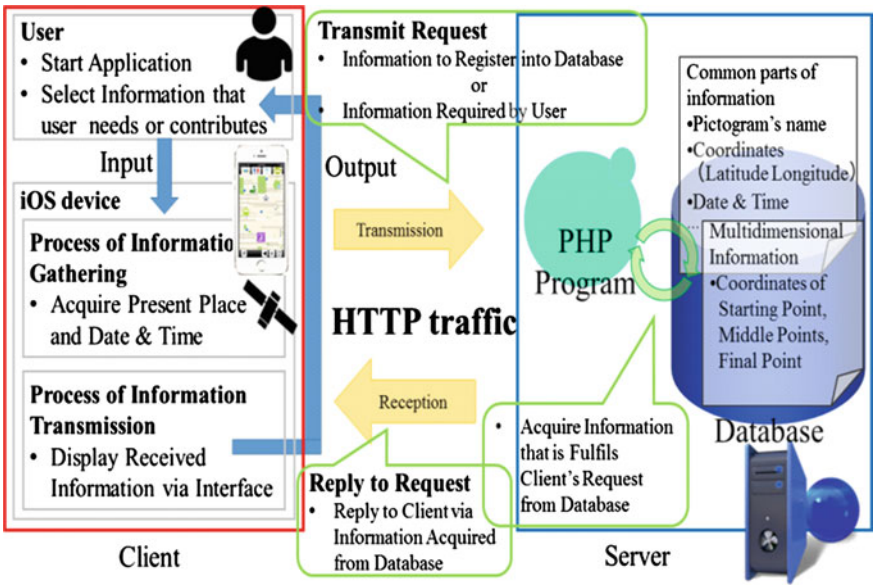


Fig. 3 System structure

5.2 System Functions

The functions of the prototype system are to collect information and to provide information. These functions are explained below.

The following describes the flow of the client application's processing when disaster victims collect information, as shown in Fig. 4.

1. The client application acquires the current location including latitude and longitude from GPS.
2. The client application acquires surrounding area information provided by other users from the server.
3. The information acquired from the server is displayed on GIS.

The following describes the flow of the client application's processing when disaster victims provide information, as shown in Fig. 5.

1. The system transitions from the information collection screen to the information provision screen.
2. The user selects a pictogram that corresponds to the information to be sent.
3. In the case of single dimensional information and two-dimensional information, the user input the routes or ranges of transmission information (Fig. 6).

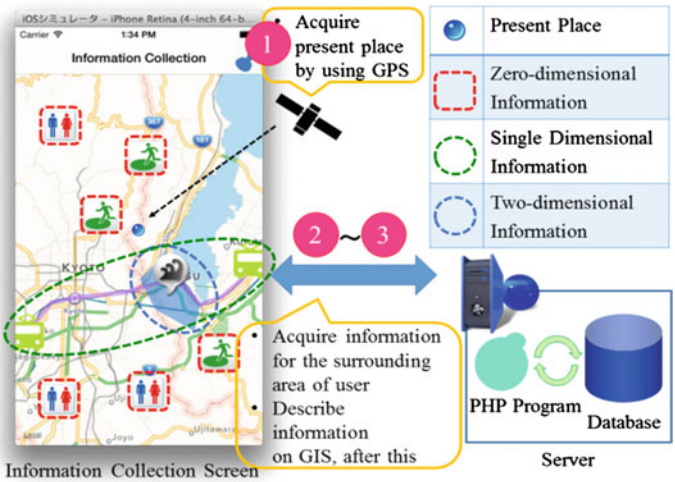


Fig. 4 System processing flow for the information collection

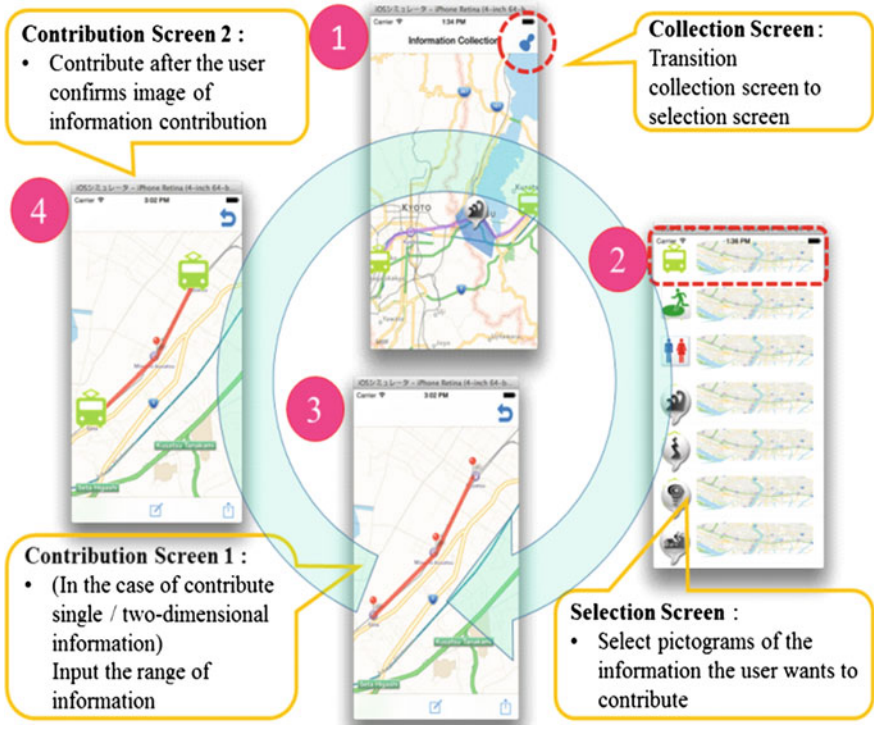


Fig. 5 System processing flow for the information provision

4. The user checks the output image, and confirms the information transmission.

When a user input single dimensional information, the user taps the positions of the start, middle and end points on the display, as shown in Fig. 6. When a user input two-dimensional information, the user taps the points of the boundary of the area in the same way, and then, taps the button to color the corresponding area. In these ways, the user are able to input multidimensional information visually.

6 System Evaluation

To verify the effectiveness of this system, we will conduct two types of evaluation experiment to examine the operability of the system and the visibility of information, respectively. Thus, we will evaluate the effectiveness of converting disaster information into nonlinguistic information from both of these aspects. In this section, we describe the evaluation experiments currently under consideration.

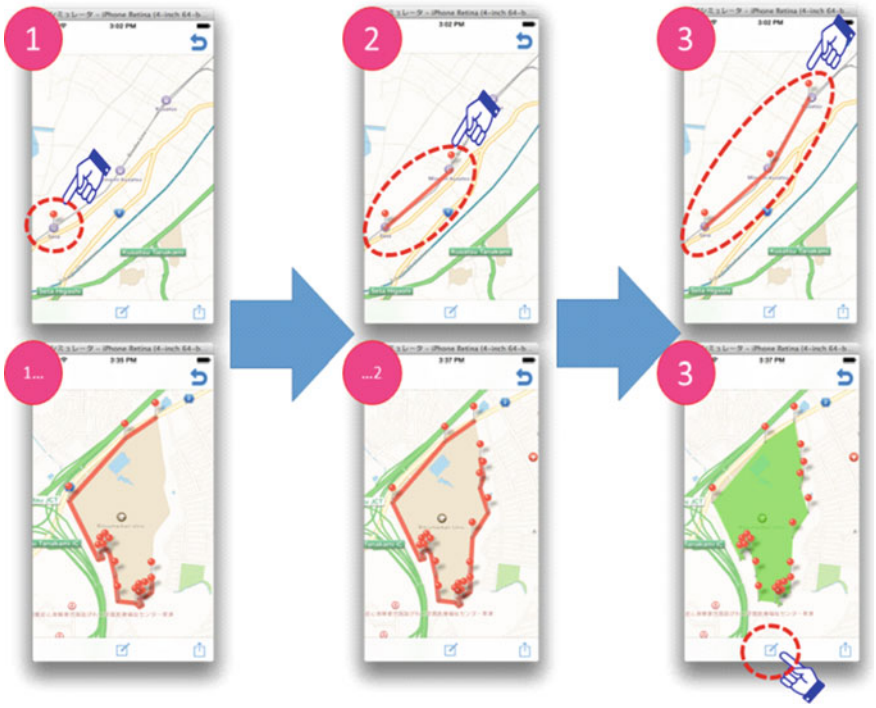


Fig. 6 System processing flow for the information collection

6.1 System Operability Test

The operability of this system will be verified using only the prototype system. We will measure the time of operating information provision, evaluate the accuracy of information quantitatively, and conduct a questionnaire survey on subjects. In this experiment, subjects will perform operation tasks (for example, providing information stating that "the trains are running from Station A to Station B"), and we will measure the time taken to transmit information, and also measure the ratio of correct answers.

6.2 Information Visibility Test

Visibility of information will be verified via a comparison between the proposed system and a version of the system that has been augmented with language-based information. Experiment subjects will be shown information and asked to fill in disaster information obtained from each system within a time limit. By establishing a time limit for system operation and recording disaster information, the

effectiveness of our nonlinguistic approach will be verified by measuring the information collected from these two systems within the time limit, both in terms of accuracy and volume of information.

As a result of the comparison in this evaluation experiment, if the difference is minimal in the amount of collected information or the visibility of the systems, or the nonlinguistic system is superior, it will be possible to verify the effectiveness of the nonlinguistic approach. We consider that higher results for information visibility may be measured for the system with added language-based information. However, we also anticipate that the system with added language-based information will obtain lower evaluation results for the volume of collected information, as the attention of the subjects will be concentrated on the language-based information.

7 Conclusion

In this research, we proposed a system that can be utilized by a wide range of disaster victims during the occurrence of a disaster, including tourists and persons who do not understand the local language, by constructing a system for disaster information collection and provision using a nonlinguistic approach. Moreover, in order to share information that contains zone and area information, which is vital when disasters occur, we proposed a system that enables disaster victims to collect and provide multidimensional information.

In this research, collection of information from a large number of disaster victims is a means to achieve the objective of creating a "Wisdom of Crowds" during disasters. Accordingly, a future task will be to consider the safety and reliability of information that is shared when disasters occur. In future, we are considering methods such as peer-evaluation of user-provided information, or to verify the reliability of information based on the location of the information provider, as in Yokobe and Nakatani's research [9]. In addition, as stated in Sect. 5, we will verify the effectiveness of this nonlinguistic system by conducting evaluation experiments with subjects including persons who are geographically unfamiliar with the area, persons who cannot communicate in a local language, and experts on disaster mitigation such as local government officials and fire fighting personnel.

Acknowledgment This work was supported in part by the Nakajima Foundation.

References

1. Cabinet Office (2010) Government of Japan: white paper on disaster management 2010. Tokyo (in Japanese)
2. Koide T (2011) Crisis responses to the great east Japan earthquake 1: fixed telephone network and communication service primary action for east Japan great disaster. Inf Process Mag 52(9):1062–1063 (in Japanese)

3. Ministry of land, infrastructure, transport and tourism, Japan tourism agency: tourism nation promotion basic law. http://www.mlit.go.jp/kankocho/kankorikkoku/kihonhou.html (in Japanese)
4. Sakai H (2011) General investigative report: sightseeing volunteers in Paris 'Parisien d'un Jour' (in Japanese)
5. Cabinet Office Council for disaster victims unable to return home during a Tokyo metropolitan earthquake: results of survey on status of measures for disaster victims unable to return home: response to 3/11 and subsequent initiatives, Japan. http://www.bousai.metro.tokyo.jp/_res/projects/default_project/_page_/001/000/439/4-1122kaigi.pdf (in Japanese)
6. Fire and Disaster Management FDMA (2005) The report of committee related graphic symbols for disaster mitigation, pp 16–19 (in Japanese)
7. Kunoki S, Izuno K, Yagi Y (2012) Universal design for disaster mitigation in tourist resorts. In: Disaster mitigation of cultural heritage and historic cities, vol 6 (in Japanese)
8. Aoyama T, Ichii T, Murakami M, Hisada Y (2006) WebGIS used in normal and disaster situations for sightseeing spot: application for the ITO in IZU peninsula. In: Summaries of technical papers of annual meeting architectural institute of Japan 2006, pp 491–492 (in Japanese)
9. Yokobe W, Nakatani Y (2011) Information collection, distribution, and provision assistance in disasters through mini-blogs. In: Innovations in information and communication science and technology IICST 2011, pp 32–41
10. Saito E (2012) The active role of social media in rapid diffusion of information: demonstration of the "wisdom of crowds" and exposure of the "idiocy of crowds". Nikkei BPnet ITpro. http://itpro.nikkeibp.co.jp/article/COLUMN/20110407/359234/ (in Japanese)
11. Kusano K, Izumi T, Nakatani Y (2013) Disaster information sharing system using pictograms only. In: Proceedings of the sixth international conference on advances in human-oriented and personalized mechanisms technologies and services (CENTRIC 2013), pp 67–72
12. Kusano K, Izumi T, Nakatani Y (2014) Disaster information sharing system using pictograms: representation of multidimensional information, lecture notes in engineering and computer science. In: Proceedings of the world congress on engineering and computer science WCECS 2014. San Francisco, pp 171–176

Dynamic Proximity Clouds on the GPU

Ryan Thomas and Sudhanshu Kumar Semwal

Abstract Ray tracing is a widely used technique in rendering realistic scenes in Computer Graphics. Its main drawback has been that it is time consuming, requiring the rendering to finish from hours to sometimes days. For decades, the goal has been to speed up the processing of these scenes. Two popular grid traversal techniques have emerged: (a) Three Dimensional Digital Differential Analyzer (3DDDA) and (b) the Proximity Cloud (PC), which is a variation of 3DDDA. Both of these techniques try to limit the number of collision tests, which can be the most time consuming part of the algorithm. While both techniques allow impressive speedups, large dynamical varying scenes topology still challenge the real time rendering process. These techniques are optimal on static scenes, but object movement forces recalculation of the scene. This is a problem when using CPUs because parallelization is not easily available. Running these on GPUs, however, allows for parallelization. Apart from briefly summarizing some of our previous results from Ryan and Semwal (Proceedings of the world congress on engineering and computer science 2014, San Francisco, pp. 376–381 [1]), we also look to answer some of the more relevant questions about ray tracing, and what future holds for this area.

Keywords Acceleration · CUDA · GPU · Graphics · Grids · Parallelization · Raytracing

R. Thomas (✉) · S.K. Semwal
University of Colorado at Colorado Springs, 1420 Austin Bluffs Pkwy Colorado Springs, Colorado Springs, CO 80918, USA
e-mail: rythomas@gmail.com

S.K. Semwal
e-mail: ssemwal@uccs.edu

H.K. Kim et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-94-017-7236-5_20

1 Grid Acceleration Techniques

The 3DDDA algorithm is a method that works with grod of voxels. These voxels are populated with the objects in the scene and allow the ray to test only objects in the voxel the ray is currently in. The 3DDDA method removes many unnecessary collision detection tests because we test only those objects which are along the path of the ray. The PC method builds from 3DDDA. PCs allow for a ray to skip a larger portion of the grid by computing how far the ray can safely jump before it might have a collision. Both methods are summarized below and are also explained in this paper [1].

2 3DDDA

Fujimoto, Tanaka, and Iwata proposed the 3DDDA algorithm in 1986 [2]. The algorithm defines a grid of voxels and determines which objects inside the voxel need to be tested against. This method would divide the scene into a reasonable sized grid. This grid is used to speed up the ray tracing process. When a ray is fired, it detects whether it hit the grid. If the grid is missed, the ray reflects the background color back showing that no object can be seen from the eye at that position. If the ray hits the grid, the cell (voxel) the ray hits is determined. The traditional ray tracing algorithm is then run on objects that are contained in the current voxel. If no objects are hit in the voxel, the ray then traverses the next voxel on the grid and continues to test each voxel in the path of the ray. This allows for a ray to only perform collision detection on objects that are in its path rather than every object in the scene, as also explained in [1].

The size of the grid does matter, but it is usually not clear what the optimal size of the grid should be in the sense of reducing the image generation time. There is a tradeoff between memory and speedup. The smaller the voxels the fewer the objects a ray will have to test against. If a scene is represented by four voxels then it is still possible that lot of unnecessary collision calculations would have to be computed, but the memory footprint of voxels would be small. If the grid represents a single $1 \times 1 \times 1$ unit in space, then the scene can truly optimize the number of collision detections but would have a larger memory footprint. The size of the grid needs to be determined by the user because the hardware may not be able to support the required memory needed to hold it. This 3DDDA algorithm also has the possibility of being optimized by testing fewer voxels if they are empty, which allows for skipping in the algorithm. There is the potential a large number of grid cells in a row can be empty. This observation is the basis for the Proximity Cloud method, see also [1].

2.1 Proximity Clouds

In 1994 Cohen and Sheffer [3] proposed another grid traversal technique to speed up the ray tracing algorithm. Similar to the 3DDDA method, the scene is divided into many voxels. Once divided the grid cells determine a safe distance a ray may skip between voxels. This results in faster traversal of the voxel grid. When a ray hits a voxel, it first determines if any objects need to be tested. If the voxel is empty or the ray did not intersect with any object, the ray skips ahead by the value the voxel says is safe to skip. This allows for fewer calculations when traversing the grid. There are a few ways this can be accomplished. Using distance transforms, these safe values are calculated during preprocessing.

A method that determines a safe distance to skip is the minimum Euclidian distance between a voxel and all of the non-empty voxels. This would give the most accurate reading for a safe distance. A more optimized method is the city block distance method. The city block distance is calculated by finding the distance in x, y, and z direction of the current voxel to a non-empty voxel and calculating it. This is performed for all of the voxels in the grid. This is not as accurate as the Euclidian distance, but it lets the program avoid calling the square root function on each of the cells. In order to compensate for the possibility of overshooting the target, the ray is normally brought back a grid cell to ensure that no objects are missed. It then continues after checking for possible intersection with that voxel. In the worst case Proximity Clouds perform the same as the 3DDDA method. This technique works great when a scene is divided across large areas; but for very close groups, the cost of building the cloud may not justify the traditional 3DDDA method. This is also explained in [1].

2.2 Directed Safe Zones

The Directed Safe Zones method was developed in 1997 and was an addition to Proximity Clouds [4]. A Proximity Cloud, as described above, finds the minimum safe distance that can be skipped by the ray as it leaves the voxel. This does not take into account the direction that the ray is traveling in. The Directed Safe Zone method takes this into account by leveraging the knowledge it has about the rays direction. When computing the clouds, the method finds the safest distance for each face of the voxel, allowing the rays to skip a larger distance depending upon which face it emerges from. This requires six times the memory, one for every face, to store the variables for the skip in each direction. When a ray leaves a voxel, the face it emerges from is determined. A lookup is performed to find the minimum safe distance the ray can travel based on the face that the ray emerges. This allows the ray to skip an even greater distance because there could be more empty space on one side of a voxel than the other. The performance of this method at its worst is again equal to the 3DDDA method because the minimum safe distance in all

directions is the proximity cloud distance. Similar to Proximity Clouds, the goal of Directed Safe Zones is to travel through the scene at a faster rate. Also see [1] for more explanation.

2.3 Slicing Extent Technique

A different approach to the grid is the Slicing Extent Technique (SET) developed in 1987 [5]. The SET method projects three-dimensional space into two so the grid method is done on a 2D plane. The slicing is done by taking perpendicular slices to the x-axis, perpendicular to the y, and perpendicular to z. Each two dimensional slice is then divided into cells. The ray traversal occurs when the ray moves from one two dimensional cell to another along the path of the ray.

2.4 Modified Slicing Extent Technique—MSET

The MSET method is an extension of the SET method for dividing up rays [5]. The SET method has a couple of shortcomings that the MSET improved upon. The first is slices are proportional to the number of objects. This would result in an unmanageable amount of slices, which made traversal time consuming. The second was the usage of floating point operations to traverse the slices. MSET evenly spaces the slices along the axis similar to the 3DDDA method. This allows for faster 3DDDA grid traversal. MSET takes into account the direction of the ray as well. When the object list inside of a cell is built, it breaks them up based on what direction the ray can hit it. The example is rays traveling up and down through the cell can only intersect with objects above and below the cell, not left and right. This method's ability to predict the size of the data structure will make it ideal for porting to the GPU in the future. The Directed Safe Zones, Slicing Extent Technique, Dual Extent, and MSET methods were not tested in this paper but warrant further research in the future, because in our experiments, as described in [4], DSZ methods outperformed the Proximity Clouds method, based on both experiments and theoretical analysis.

2.5 GPU Computing

Graphical Processing Units have become common in most computers today. They have been around since the 1980s with the goal of speeding up a rendering process for the system. This allows for better processing by freeing up the CPU from rendering by having dedicated hardware to draw to the screens. GPUs are built with many processing cores that run in parallel. The recent trend has been that the GPU

manufacturers are providing APIs for traditional processing on the GPU rather than just graphics rendering. Object hierarchy methods such as k-d tress [6] and BVHs [7] are used for these implementations. But the above mentioned approaches are based on partitioning the object space and may not be suitable for ray tracing a 3D volume data-set, such as that available from CT/MRI slices or even the visible human project. This is because volume data does not have any objects to build the object hierarchies on.

GPU computing has been growing over the previous few years. Supercomputers like performance can be provided on our desktop with multiple GPUs. In our implementation, the 3DDDA algorithm remains the same when ran on the GPU or the CPU. Proximity Clouds were allowed a different approach on the GPU. The principles of the algorithm remain the same but are rewritten using a parallel processing implementation. The benefit is the simplified loop, which allows the distance calculations to be run simultaneously while building the Proximity Clouds. These adjustments allow Dynamic Proximity Clouds implementation.

Modern day CPUs generally have two to four cores on the standard desktops and up to sixteen cores on server CPUs. A device running on the GPU has the potential to have thirty-two threads on a single core. Threading is also made easier in languages such as CUDA. This provides a massive amount of computing power in the average consumer computer. There are three main libraries that allow processing on the GPU: Microsoft's DirectCompute (which is bundled with DirectX 11), OpenCL, and Nvidia's CUDA library. We have used CUDA for our implementation.

The disadvantage which implementations face on the GPU is the memory constraints. For a GPU implementation, usually we are limited to the available memory on the device. This is somehow copied from the host (CPU) memory space, but this operation can be time consuming. This could be tricky when handling dynamics objects and topologies, such as for game applications. In this case, one set of data needs to be exchanged with another modified set. This research does not investigate acceleration techniques to buffer memory, but that would be interesting question for the future [1].

3 Ray Tracing on the GPU

3.1 Implementation

The first step to setup ray tracing implementation on the GPU is to set up and run the traditional algorithms. On the GPU the rays that emanate from the eye are divided into grid segments and run on concurrent threads as described by threading in CUDA. This results in a tremendous speedup from the version on the CPU. The algorithm remains the same for the kernel drawing the spheres [1].

3.2 3DDDA

The 3DDDA method on the GPU is similar to the version on the CPU. A grid is still built in a similar manner. The GPU implementation has the same algorithm when traversing the grid. It does fire many concurrent rays similar to the traditional one. The threads then traverse the global grid. The pre-processing is done in one kernel, and the traversing is done in another. Although explained in more details in [1], here we briefly describe this approach.

The first step is to determine which voxels are filled and which are not. The grid is a set size on the GPU since dynamic allocating and releasing of memory prevents real time rendering. Each voxel is made up of a structure. Each voxel contains a boolean to determine if it contains any spheres, an array of indices for the spheres located in the voxel, and three integers describing the x, y, and z position. This helps in the parallelization since the voxel array is declared as a single dimension array.

Building the grid is done with two separate kernels. The first kernel is designed to empty the grid. This is needed to register spheres with the correct grid cell. This is threaded in the x and y directions. Each thread loops over the z direction setting each cell to empty and resetting the array value to −1, which represents the end of the list of spheres contained in the voxel. This array is static because the GPU prefers static memory to dynamic. This can become a limitation of the system; however, when looking at enough spheres to fill a single voxel other memory issues may be introduced. In that case a new grid of a different size can be built or that array can be increased. On certain cards the memory limitations may be met building the grid and should be taken into account.

Once the grid is reset, it needs to be populated with the spheres contained in the spheres array defined using a float3 data type. The float3 data type is defined in CUDA and contains three floats. The min and max of the sphere is determined in order to create a bounding box used to fill the voxels. Spheres also define a material, which determines their reflectiveness and color. An array on the GPU defines the spheres that are to be rendered. This global array is used to populate the voxels in a second kernel. This kernel is a single dimensional kernel that threads on the number of spheres. Each sphere fills the cells that are contained within its bounding box. Once this kernel is finished the grid traversal can begin.

The final kernel in the process draws the scene. It is broken up in the same way as traditional ray tracing, but it does not compute the collision for each sphere, rather traversing the grid in a device function as explained in [1].

3.3 GPU Proximity Clouds

Proximity Clouds was built using a variety of distance computations. All of them require at least an n squared algorithm to compute, usually two pass algorithm as

described in [3] is implemented. These algorithms can be parallelized to run on the GPU. The algorithm presented in this paper is as follow:

- The index represents the voxel that needs to determine its distance.
- The thread in the y direction represents a voxel that is not empty.
- If a thread receives a voxel that is empty, the thread returns and asks for the next voxel.
- Once all threads in a block are completed a new block is loaded on the GPU.

This builds the clouds in a way that can be scaled to multiple GPUs and can run on any GPU. It is possible to run this in a single warp when sufficient threads are available. This algorithm can take any distance computation as input to determine a safe distance to jump. The speed of this algorithm comes from the parallelization of the system. The first list is divided up into many different threads so they can run in sync. The Proximity Cloud generated for the experiment consisted of a $40 \times 40 \times 40$ grid. The list is 64,000 in length and the second list would be of equal size for the worst case, as also explained in [1].

4 Results

The results were run on a Windows 7, 64 bit PC with 6 GB of RAM, an Intel i7 2.66 GHz processor, and an NVidia GTX 570 graphics card. The GTX 570 has 480 CUDA cores with a graphics clock of 732 MHz and a processor clock of 1464 MHz [1]. Figure 1 shows a variety of techniques we have developed and their relationships with the other existing techniques.

4.1 Clustered Spheres Performance

We started with a rendering of up to 15,000 spheres and a $20 \times 20 \times 20$ grid. This grid size is important because sometimes larger grid side could lead to time out when filling the grid, as will be discussed later. This scene has the quality that the scene had no real gaps. So we expected that the cloud and the grid would perform at the same speed. The proximity cloud only produced a 0.3 % increase in speed, but it did speed up traditional ray tracing by 91.8 %. For more discussion, please refers to [1].

4.2 Clustered Spheres with Large Grid

To better demonstrate the performance of a larger grid, the same scene as above was generated using a $40 \times 40 \times 40$ grid. Since larger grid size was used, we needed to reduce the number of spheres. Far fewer spheres were used due to the timeout that

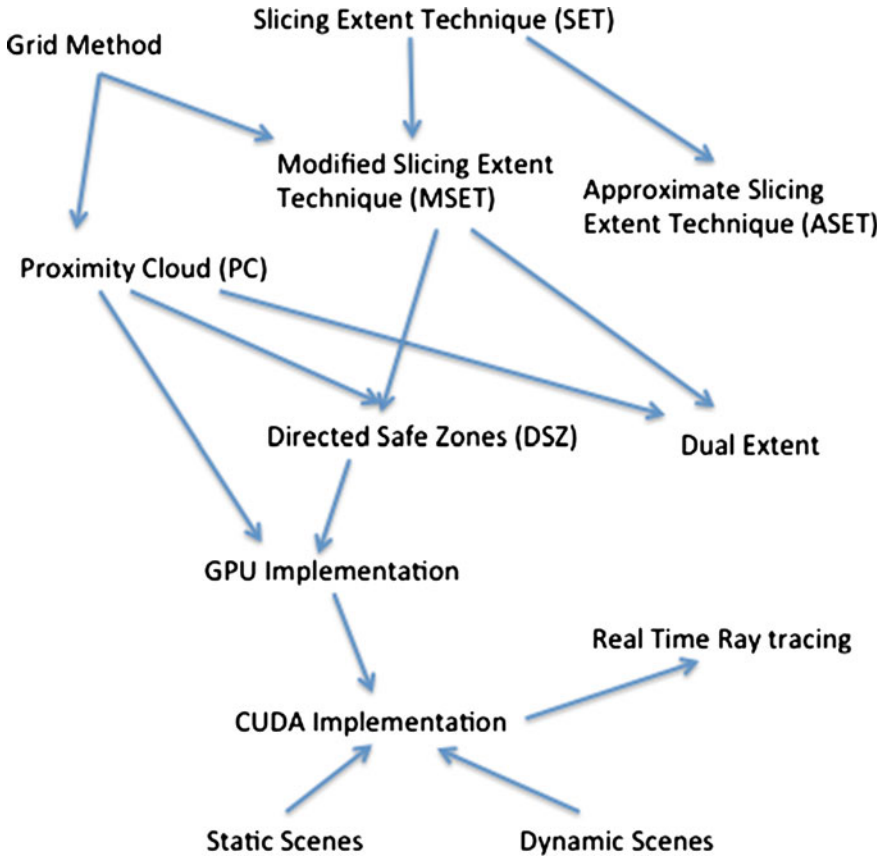


Fig. 1 GPU implementation towards possible real time rendering

can occur when building the Proximity Clouds with the addition of rendering. [1] Once again there is not a huge distinction between the grid and the Proximity Cloud method. This is because there are no large empty areas. There is a 6 % increase in speed when looking at proximity clouds over the grid method, and a 79 % increase in speed over traditional ray tracing.

4.3 Rendering Planes of Spheres

In order to showcase the advantage that has been claimed for Proximity Clouds, a new scene was generated where two planes of spheres are created. One plane is toward the front of the scene, while the second plane is far at the back. This allows for a large gap between them demonstrating how Proximity Clouds increases

rendering time. The grid and Proximity Cloud methods show much faster performance over non optimized ray tracing. The difference between proximity clouds and the 3DDDA method is not as large as expected. This is partially due to the size of the grid. A 40 × 40 × 40 grid can only skip the length of the grid in the best case. This does not allow for huge performance increases when each thread is traversing the grid simultaneously. The average increase in speed was only 6.4 % when looking at all points on the graph, but a 94.2 % increase from the non-optimized ray tracer was achieved. As the scene grows and becomes sparser this should only increase as also indicated in our paper [1].

4.4 Proximity Cloud Generation Speed

The next set of data that was looked at is the speed at which Proximity Clouds can be generated. This is based on the size of the Proximity Cloud as well as how populated the scene is. Each test below generates spheres in random locations and tests the speed at which the Proximity Clouds can be generated. When building a 10 × 10 × 10 grid, it can be put in a single warp. Building it only takes 1 ms, which is a single loop through the scene. A 20 × 20 × 20 cloud is a little bigger, but it experiences a similar behavior to the 10 × 10 × 10 cloud. Once at 350 ms it begins to level out no matter the number of spheres added because the scene has a sphere in every voxel in our implementation for this case.

While slower the 30 × 30 × 30 grid is still manageable. Similar to the 20 × 20 × 20 grid, the 30 × 30 × 30 builds in the same time scale. This is due to a large number of CUDA cores working in parallel. The 30 × 30 × 30 grid requires the same number of warps as the 20 × 20 × 20, thus completing in the same amount of time.

The 40 × 40 × 40 grid requires multiple loops in order to build. Building the grid takes over 1.4 s, and when built with other kernels, it has the potential of running into the kernel timeout. These results do show that running across multiple cards can increase the speed of Proximity Clouds so that it's closer to the speed it takes to build the 10 × 10 × 10 grid. The speed to build the Proximity Clouds, plus the time to render, is still faster than the traditional ray tracer and a slight improvement over the 3DDA method [1].

5 Future Work

GPUs have permeated the graphics industry and we expect their usage to only grow more. One area to extend this work is to use some form of Cellular Automata for implementing dynamic changes in the scene because the changes usually are localized. Cellular Automata [8, 9] could implement multi-level interactions and emergence of diseases [10, 11]. Complex Systems science [12] has been applied to

model events occurring in nature. Works by Prigogine [13], in thermodynamics, and earlier work by Poincare's on sensitivity of dynamical systems to initial conditions provide the basis for complex systems research for Cellular Automata research. Limitations of simulating organic life by using computational models have been discussed before, these include (i) brittleness [14] of the computational medium, and (ii) the limitations of reductionist approaches to model organic life, which is well documented in [15]. Because Cellular Automata uses local interactions, not the reductionist approaches, it could provide a suitable platform to model organic behavior such as cancerous growth patterns. Local interactions, usually implemented for every cell, could create subtle interactions mimicking organic behavior. Many examples, such as flocking, and 3D games have shown remarkable variety of emergence when a cell's next state is based on consulting nearby voxels. For example twenty-seven cells could be consulted for ($3 \times 3 \times 3 = 27$; 26 immediate vicinity, and 1 itself) to decide the next state. Different non-linear and dynamics pattern could emerge using different local interactions strategies [16].

Volume Data provides one-to-one correspondence for use by a Cellular Automata. The Visible Human Project supported (1989–2000) by US National Library of Medicine (NLM) provides a detailed volume data of human body. The process created a very detailed database of volume data 1 mm apart for the male cadaver with 1871 slices, which when stacked create a 3D grid of volume. This created 40 GB of static 3D grid data which might have to be ray traced, or variation of ray tracing called ray-casting could be used. 3D Morphing techniques [17], and for medical applications [18, 19] have been implemented using cellular automata on volume data. However, real-time manipulation of such large data is not possible with the computer systems of today, yet GPU computing provides a promising research direction. The rise of GPU computing has been growing over the previous few years. GPU computing allows for parallelization of algorithms when the algorithm allows for it. The 3DDDA traversal algorithm remains the same when run on the GPU and the CPU. Proximity Clouds are allowed a different approach on the GPU. The traditional algorithms proposed can be mapped on the GPU. The principles of the algorithm are the same but now are rewritten using a parallel processing implementation. The benefit is the loop is simplified allowing the distance calculations to be run simultaneously while building the Proximity Clouds.

6 Conclusion

Dynamic Proximity Clouds were achieved by dividing the problem into several kernels and allowing the GPU to compute each section. Each generated scene consisted of building a blank voxel grid, filling the grid with the spheres, computing the Proximity Clouds, rendering, and updating the positions of the objects. This allowed for a nice animation when the number of spheres was manageable, but can quickly become choppy, i.e. non real-time, due to the size of the grid.

On average it took 1.44 s to generate the clouds, which sometimes had the potential to reach the timeout of the GPU. The 3DDDA provided the largest performance gain in terms of overall speed but was slower when it came to rendering using Proximity Clouds. This algorithm can be run across multiple GPUs providing real time rendering, but new ways should be addressed, perhaps increasing the speed across a single GPU. By running it in parallel, it is possible to build the Proximity Clouds in a single cycle. The speed increase between the Proximity Clouds and 3DDDA on the GPU demonstrates that more investigation is needed to come up with better traversal methods. Continued parallelization of the algorithm will only result in a better speedup. Finally, cloud computing could provide better and novel solutions to this interesting problem.

Acknowledgments This paper is an invited Chapter based on our earlier publication [1] at the WCECS 2014 conference. Although we have added several new sections, some remnants of the old paper still might be present as we started with our original submission [1]. Both authors want to thank the WCECS 2014 conference organizers for inviting us to submit this book Chapter.

References

1. Ryan T, Semwal SK (2014) Ray tracing using 3D grid simulations, lecture notes in engineering and computer science. In: Proceedings of the world congress on engineering and computer science WCECS 2014. San Francisco, pp 376–381, 22–24 Oct 2014
2. Fujimoto A, Takayuki T, Kansei I (1986) ARTS: accelerated ray-tracing system. IEEE Comput Graph Appl 6(4):16–26 (print)
3. Cohen D, Sheffer Z (1994) Proximity clouds an acceleration technique for 3D grid traversal. Vis Comput 11(1):27–38
4. Semwal SK, Hakan K (1997) Directed safe zones and the dual extent algorithms for efficient grid traversal during ray tracing. In: Graphics interface 1997, pp 76–87 (print)
5. Semwal SK, Kearney CK, Moshell JM (1993) The slicing extent technique for ray tracing: isolating sparse and dense areas. In: Graphics, design and visualization 1993, pp 115–122 (print)
6. NVIDIA CUDA (2009) C programming best practices guide. In: CUDA programming guide. NVIDIA, Web. 10 Mar 2011
7. Manocha D, Lauterbach C (2006) Ray tracing dynamic science using BVHs. In: SigGraph 2006 presentation, pp 1–47
8. Sarkar P (2000) A brief history of cellular automata. ACM Comput Surv (CSUR) 32(1):80–107
9. Wolfram S, A new kind of science, book on cellular automata. Wolfram Media Company, London, pp 1–849
10. Bezzi M, Modeling evolution and immune system by cellular automata. http://citeseer.nj.nec.com/429312.html
11. Sosic R, Johnson RR (1995) Computational properties of self-reproducing growing automata. BioSystems 36:7–17
12. Melaine M (2009) Complexity: a guided tour. Oxford University Press, Oxford, pp 1–337
13. Prigogine I, From being to becoming, freeman (ISBN 0-7167-1107-9)
14. Ray T (2000) An evolutionary approach to synthetic biology: zen and the art of creating life, Chap. 2. In: Book on best papers from VW98 Paris conference

15. Stephen R, Lessons from the living cell: the limits of reductionism. McGraw Hill, New York, pp 1–300
16. Rabinovich MI, Ezesky AB, Weidman PD (2000) The dynamics of patterns, World Scientific, Singapore, pp 1–324
17. Semwal SK, Chandrashekher K (2005) 3D morphing for volume data. In: The 18th conference in central Europe, on computer graphics, visualization, and computer vision, WSCG 2005 conference, pp 1–7
18. Fang S, Raghavan R, Richtsmeier J (1996) Volume morphing methods for landmark based 3D image deformation. In: SPIE international symposium on medical imaging
19. Forsyth T (2002) Cellular automata for physical modeling. Game Program Gems 3:200–214

Sectional NoC Mapping Scheme Optimized for Testing Time

Zhang Ying, Wu Ning and Ge Fen

Abstract NoC architecture has been increasingly applied to complex SoC chips and how to efficiently map the specific application to NoC infrastructure is an important topic urgently needed to study for NoC. At the same time, there are many challenges for NoC embedded IP cores testing. This paper proposes a sectional NoC mapping algorithm optimized for NoC IP cores testing. Associated with the pre-designed test structure, sectional NoC mapping firstly adapts the Partition Algorithm to arrange IP cores into parallel testing groups to minimize testing time. Then, it applies genetic algorithm for NoC mapping based on the traffic information between IP cores. The experiment results on ITC'02 benchmark circuits showed that the mapping costs decreased by 24.5 % on average compared with the random mapping and the testing time can be reduced by 12.67 % on average as well, which illustrated the effectiveness of the sectional NoC mapping scheme.

Keywords Genetic algorithm · NoC mapping · NoC testing · Partition algorithm · Testing optimization · Testing schedule

Z. Ying (✉) · W. Ning · G. Fen
The College of Information Science and Technology, Nanjing University of Aeronautics and Astronautics, 29# Yudao Street, College No.4, NUAA, Nanjing 210016, Jiangsu Province, China
e-mail: tracy403@nuaa.edu.cn

W. Ning
e-mail: wunee@nuaa.edu.cn

G. Fen
e-mail: gefen@nuaa.edu.cn

H.K. Kim et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-94-017-7236-5_21

1 Introduction

With the increasing development of the semiconductor technology and chip integration, NoC (Network on Chip) becomes an important solution to complex SoC (System on Chip) architecture [1]. NoC mapping is meant to assign a specific application on the NoC platform in accordance with certain rules, which will greatly influence the system performance. Mapping is one of the hot issues in NoC research [2]. The existing mapping schemes are mainly based on various optimization algorithms to minimize the traffic or energy consumption [3, 4], which only consider functional performance. However, the difficulty of complex system chip testing increases dramatically [5], which brings the requirement of considering test optimization while mapping.

NoC-based SoCs often reuse NoC communication architecture as TAM (Test Access Machine) for embedded IP cores testing. Parallel testing can be achieved by multiple ATE (Automatic Test Equipment) simultaneously access the NUT (NoC Under Test) through multiple ports. There have been some research [6, 7] discussing on the multiple access ATE schemes. Authors of Amory et al. [6] proposed a DFT scheme reusing the NoC as TAM and described the associated testing structure. Although it focuses on the design of ATE interface and test wrapper, the multi-ATE input/output and test scheduling are involved. The diagram of NoC partition and three accessible ATEs is provided and the test partition is aimed to minimize the testing time. However, it is worth noting that the partition is only applied to manufactured chips. Agrawal et al. [7] proposed the test data delivery optimization for NoC. It co-optimizes quantity and position of the access points and the assignment of cores to access points. Testing time minimization is modeled as a NoC partitioning problem and solved with some optimized algorithms. And yet the scheme is also only concerned with the manufactured chips. Our proposed solution is to take into account the needs of parallel testing during NoC mapping process. Based on the optimized test access structure, partition IP cores are partitioned to testing groups in advance, which effectively improves the efficiency of NoC embedded IP cores testing.

We propose a sectional NoC mapping scheme which is optimized for testability. Firstly, uses Partition Algorithm to assign IP cores into groups to minimize testing time; then maps the specific application to NoC structure; finally optimizes the NoC mapping based on genetic algorithm in order to minimize mapping overhead.

Hereafter, Sect. 2 gives the analysis of embedded IP cores testing time and introduces the testing optimization NoC mapping problem. Section 3 presents the sectional NoC mapping scheme with the optimization of embedded IP cores testing. Section 4 explains the results of testing time and mapping performance evaluation and compares with other schemes. Section 5 draws the conclusions.

2 Ip Cores Testing Time Analysis and the Problem Description

NoCs can be defined as a set of structured routers and point-to-point channels interconnecting IP cores (Resources). The topology of NoC can be represented as an undirected connected graph $G(N, L)$, where $N = \{n_1, n_2, \ldots n_i\}$ is the set of nodes and $L = \{l_1, l_2, \ldots l_j\}$ is the set of links in the corresponding network. For regularity of its structure, Mesh network are easy to implement and have good scalability. We proposed testing optimized NoC mapping is based on 2D Mesh NoC.

The optimization objective of IP cores parallel testing is testing time (the number of test cycles), so the independent testing time of each core needs to be firstly determined. Our testing optimized NoC mapping is applied to the ITC'02 benchmark circuits [8], which contain 12 typical SoC circuits provided by the research institutions and corporations.

Circuits in benchmark are made up with cores (modules). For each module, number of test patterns is p_m, number of input signals is i_m, number of output signals is o_m, number of bidirectional signals is b_m, number of scan chains is s_m, length of scan chain k is $l_{m,k}$, total number of scan FFs is f_m and $f_m = \sum_{k=1}^{s_m} l_{m,k}$. The number of test stimulus and response for each test vector is s_{im} and s_{om}, so Eqs. (1) and (2) can be obtained.

$$s_{im} = i_m + b_m + f_m \tag{1}$$

$$s_{om} = o_m + b_m + f_m \tag{2}$$

The input data and output data of each test vector can be transmitted simultaneously in the scan test except the last one. Moreover, there is an extra cycle needed for function. Therefore, the independent testing time for an IP core T_c without considering the TAM width is as Eq. (3).

$$T_c = \{1 + \max(s_{im}, s_{om})\} \times p_m + \min(s_{im}, s_{om}) \tag{3}$$

Assume the TAM width is w, T_c will be as Eq. (4).

$$T_c = \left\lceil \frac{\max(s_{im}, s_{om}) \times p_m + \min(s_{im}, s_{om})}{w} \right\rceil + p_m \tag{4}$$

For NoC-based SoCs, TAM is usually the NoC communication channel and the width is typically 16 or 32 bits. Test data are transmitted in the form of packets, which are composed of flits in wormhole routing. One flit is transferred in one clock cycle. The testing time will be different when there is scan chain or not in IP core and is various when the number of the scan chain less than, equal to or greater than

the transmission bandwidth. Therefore, we analyze the testing time of embedded IP cores under various circumstances and the results are summarized as follows.

1. IP core has no scan chain

Flits number of each packet is $N_f = \frac{\max(s_i, s_o)}{w}$, $s_i = i_m + b_m$, $s_o = o_m + b_m$, testing time is as Eq. (5).

$$T_c = \max\left\{\left\lceil\frac{s_i}{w}\right\rceil, \left\lceil\frac{s_o}{w}\right\rceil\right\} \times p_m + \min\left\{\left\lceil\frac{s_i}{w}\right\rceil, \left\lceil\frac{s_o}{w}\right\rceil\right\} + p_m \tag{5}$$

2. IP core has scan chains

A. the number of scan chain is much smaller than current bandwidth

Since the input and output can be transmitted together in the scan chain, so that $N_f = \max(l_{m,k}) \quad k = 1, \ldots, s_m$, testing time is as Eq. (6).

$$T_c = N_f \times p_m + p_m + N_f \tag{6}$$

B. the number of scan chain is close or equal to the bandwidth

The input and output will not be transmitted together in the scan chain, so $N_f = N_{f1} + N_{f2}$, $N_{f1} = avg(l_{m,k}) \quad k = 1, \ldots, s_m$, $N_{f2} = \max\left\{\left\lceil\frac{s_i}{w}\right\rceil, \left\lceil\frac{s_o}{w}\right\rceil\right\}$, and $T_c = T_{c1} + T_{c2}$

$$T_{c1} = N_{f1} \times p_m + N_{f1} + p_m \tag{7}$$

$$T_{c2} = N_{f2} \times p_m + \min\left\{\left\lceil\frac{s_i}{w}\right\rceil, \left\lceil\frac{s_o}{w}\right\rceil\right\} \tag{8}$$

C. the number of scan chain is greater than the bandwidth

so $N_f = \max\left\{\left\lceil\frac{f_m+s_i}{w}\right\rceil, \left\lceil\frac{f_m+s_o}{w}\right\rceil\right\}$, testing time is as Eq. (9).

$$T_c = N_f \times p_m + p_m + \min\left\{\left\lceil\frac{f_m + s_i}{w}\right\rceil, \left\lceil\frac{f_m + s_o}{w}\right\rceil\right\} \tag{9}$$

According to the above analysis, the testing time of each IP core for ITC'02 circuits will be obtained. For example, IP core testing times of d695 circuit are shown in Table 1.

When applying parallel testing on NoC -based SoC, ATE can access the NUT by various ports. Ref. [9] provided performance evaluation of several ATE access modes and proposed the best solution as shown in Fig. 1. The ATEs access from the East, South, West, and North corners of Mesh. So that four sub-blocks testing can

Table 1 Testing time of d695 circuit

Core	W = 16		W = 32	
	Flits/packet	T_c	Flits/packet	T_c
1	2	38	1	25
2	13	1029	7	588
3	32	2507	32	2507
4	54	5829	54	5829
5	109	12,192	55	6206
6	50	11,978	41	9869
7	43	4219	34	3359
8	46	4605	46	4605
9	128	1659	64	836
10	109	7586	55	3863

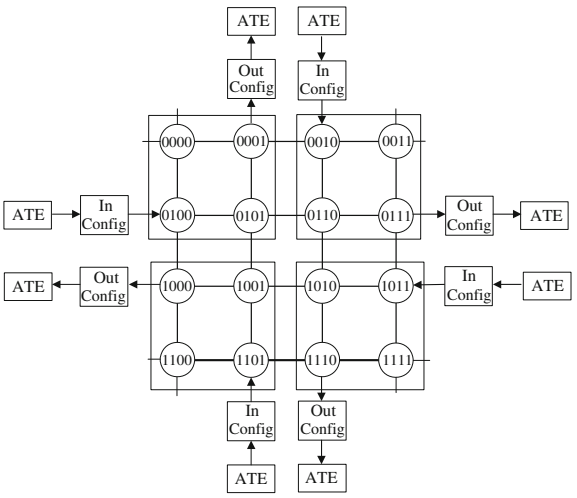


Fig. 1 Optimized ATE access mode

be executed simultaneously. The test architecture will significantly optimize the testing time and has relatively low testing cost.

According to the test access mode, IP cores of NoC need to be pre-grouped and then mapped to a 2D Mesh architecture. By that means, the efficiency of NoC embedded IP core testing will be significantly improved. At the same time, NoC IP cores testing optimization has converted into the IP cores partition problems [10]. It can be summarized as follows.

Given the set of IP cores $C = \{c_1, c_2, \ldots, c_N\}$, assume the width of NoC transfer channel is W, then the individual testing time of each IP core T_i can be induced. Partition C into four groups P_i $(i = 0, 1, 2, 3)$, the core number and testing time of P_i is separately Num_i and TP_i, so that the grouping optimization objective is as Eq. (10).

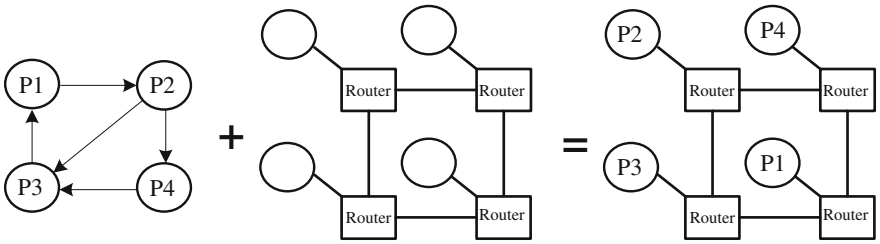


Fig. 2 2D mesh NoC mapping

$$\text{Minimize}(\text{Max}(TP_i))\ i = 0, 1, 2, 3 \tag{10}$$

And the following constraints should be satisfied in the meantime.

$$\sum_{i=0}^{3} Num_i = N \tag{11}$$

$$Num_i \leq \left\lceil \frac{N}{4} \right\rceil \tag{12}$$

On the other hand, assuming the core structure and IP cores traffic information are provided, each IP core is allocated to corresponding NoC resources node. This process is called NoC architecture mapping for specific applications. The 2D NoC mapping process is shown in Fig. 2. The mapping performance can be evaluated by the objective function.

The description of 2D Mesh NoC mapping problem needs to provide three definitions as follows firstly.

Definition 1 Given directed acyclic weighted graph $G(V, E)$ as application characteristics chart, each vertex $v_i \in V$ shows the associated IP core, directed arc $e_{i,j} \in E$ shows the communication relationship between v_i and v_j, coefficient $\omega_{i,j}$ shows the traffic between v_i and v_j.

Definition 2 Given directed graph $P(R, P)$ as NoC structure feature graph, each vertex $r_i \in R$ shows resource node, directed arc $p_{i,j} \in P$ shows the router between v_i and v_j, $E(p_{i,j})$ shows the power consumption of one bit transfer between v_i and v_j.

Definition 3 power consumption functions are defined as Eq. (13) and (14).

$$Energy = \sum_{\forall e_{i,j}} \omega_{i,j} \times E(p_{i,j}) \tag{13}$$

$$E(p_{i,j}) = n_{router} \times E_{Sbit} + (n_{router} - 1) \times E_{Lbit} \tag{14}$$

n_{router} is number of routers, E_{Sbit} and E_{Lbit} are separately power consumption of routers and interconnections. Considering the 2D Mesh architecture and XY routing algorithm, $E(p_{i,j})$ is actually determined by the hamming distance between v_i and v_j, so that it can be induced as Eq. (15).

$$Cost = \sum_{i=0}^{N-1} \sum_{j=0}^{M-1} \omega_{i,j} * distance_{i,j} \tag{15}$$

Among them, *Cost* is the mapping overhead, $\omega_{i,j}$ is the traffic between v_i and v_j, $distance_{i,j}$ is the hamming distance between v_i and v_j, and the NoC is N × M 2D Mesh architecture.

NoC mapping problem: given G and P, search for mapping function $map()$, for minimizing *Cost* and satisfying the following constraints.

$$\forall v_i \in V \Rightarrow map(v_i) \in R \tag{16}$$

$$\forall v_i \neq v_j \Leftrightarrow map(v_i) \neq map(v_j) \tag{17}$$

$$size(G) \leq size(P) \tag{18}$$

For identified optimal test structure, the original problem of parallel testing has transformed into IP cores partition. Considering the traffic and power constraints of NoC application, group selection of NoC embedded IP cores can be firstly implemented. After that, optimization of the NoC mapping is applied based on the traffic information of groups, which will realize the sectional test optimization of NoC mapping.

3 Sectional Testing Optimized NoC Mapping

When testing time of IP cores are determined, sectional testing optimized NoC mapping firstly apply P_A algorithm for grouping IP cores so that the parallel testing time of NoC is minimized. The detail of the P_A algorithm process is shown in Fig. 3.

Firstly, sorts all IP cores in descending order according to their independent testing time. Each IP core is then successively assigned to the group, whose length of testing time after this assignment is closest to, but not exceeding the current maximum test time. That means, each IP core is assigned to the group in which it achieves the best fit. If there is no such a group available, the IP core will be assigned to the current minimum testing time group.

Fig. 3 Pseudo-code of P_A algorithm

Procedure P_A algorithm

1. Define N_p, which is the upper limit of cores number in each partition based on the total cores number;
2. Sort cores in descending order of test cycles;
3. Add the first core (whose test cycles is maximum) to P_0;
4. Add the successive three cores to P_1, P_2, P_3 respectively;
5. For each of other core c_i
   i. Find partition $P_{\max}$ with current maximum test cycles;
   ii. Find partition $P_{\min}$ with current minimum test time and cores number is less than N_p;
   iii. Assign c_i to partition p, such that {test time($P_{\max}$) - (test time(p)+test time(c_i))} is minimum and cores number is less than N_p;
6. If there is no such partition p, then assign c_i to $P_{\min}$.

According to IP cores grouping results and traffic information, the initial layout of IP cores in NoC may be determined. The specific mapping algorithm needs to be applied for further optimizing mapping performance.

Genetic Algorithm is one of evolutionary algorithms, which generate solutions to optimization problems using techniques inspired by natural evolution, such as inheritance, mutation, selection, and crossover. When using genetic algorithms to solve practical problems, optimization parameters or variables need to be firstly translated into code string (usually in binary), these are so-called chromosomes and the process is encoding. After encoding is finished, construct the initial possible solution set, which provides the primary population size and randomly generates the chromosome set. In each generation, selects individual from problem domain based on the size of fitness, and implements the crossover and mutation by the genetic operator, resulting in a population representative of new solution set. This process will obtain the best individual in the offsprings, which is the same as the natural evolution. Finally, the last individual will be decoded and can be approximated as the optimal solution.

As far as the NoC mapping problem is concerned, we adopt the Genetic Algorithm as the optimization mapping algorithm. The mapping scheme based on GA algorithm is shown in Fig. 4.

The communication power consumptions are calculated as the fitness. CCG (C, A) is the NoC application characteristics chart and NAG (R, P) is the IP cores partition result. The appropriate genetic operations (selection, crossover and mutation) execute successively to produce a new generation of NoC mapping. This process will be iterated until the optimal mapping solution will be produced. GA algorithm steps are described in detail as follows.

Fig. 4 Mapping scheme based on GA algorithm

```
Procedure GA algorithm
input : CCG(C,A), NAG(R,P)
output : IP cores mapping generation
1. Initialize the population with segmented sequences;
2. Read the communication data;
3. Set the parameters;
4. for i=1 to N_max
    {
      for each solution
     {
      for each ai,j in CCG in current mapping
        {
         count IP cores allocated in different ports;
         compute the coordinates of the mapped nodes map(ci)
         and map(cj);
         generate placement;
         calculate Cost;
         if (Cost is minimum) save mapping as the best one.
         }
      select();
      crossover();
      mutate();
      upgrade_ population();
     i=i+1;
    }
5. output the optimal solution.
```

A. Initialization
Based on the grouping results, IP cores are assigned to four ports. So the initialization of the Genetic Algorithm is divided into four sections. Among them, the range of chromosome segment is determined by grouping results. An example of chromosome coding is shown in Fig. 5.

B. Calculating fitness
Fitness is the communication power consumption, whose value is related to the communication distance and the traffic between cores and can be obtained by Eq. (3).

C. Selection
Roulette selection method is applied to select individuals in the populations, in which the probability of an individual being selected is inversely proportional to its fitness. It reflects the proportion of the individual fitness to the sum of the entire population fitness.

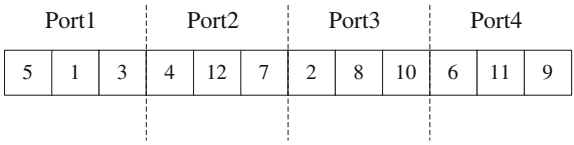


Fig. 5 Example of chromosome encoding

D. Crossover
Crossover operation is to reconstruct parts of two parent individuals to generate new individuals. We adopt single-point crossover, a random number is generated within the length of the chromosome as a scheduled crossing point. Then two individual chromosomes exchange the order on this crossover point. Figure 6 shows the improved sequence exchange example.

E. Mutation
Mutation operation is to change the gene value of the individual, in order to ensure that the algorithm has the ability of random search and maintain the local population diversity. Usually selects one or more genes by random in the individuals and alters them at the preset probability.

4 Performance Evaluation of Mapping Schemes

In order to evaluate the proposed testing optimization mapping scheme, some experiments are applied to four circuits of ITC'02 benchmark and they are d695, g1023, p22810, p93791. P_A algorithm is written in C++, and its compiler and simulation environment is Visual C++ 9.0. Genetic algorithm is programmed under the environment of MATLAB R2009a. Experiments are executed on CPU for the Intel (R) Core (TM) i5-2400 3.1 GHz and 4 GB memory. Firstly, P_A algorithm is implemented and the comparison of testing time with the Ref. [11] is shown in Table 2.

Fig. 6 Example of sequence exchange

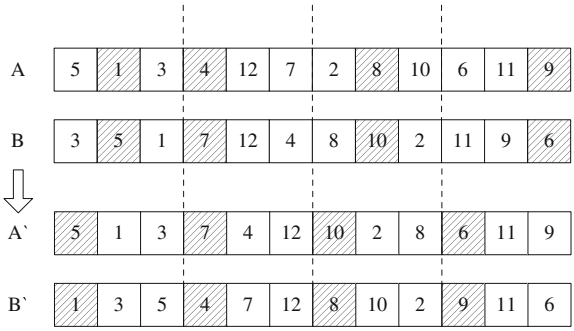


Table 2 Comparison of testing time

Circuit	W = 16		W = 32	
	Base case [11]	P_A case	Base case [11]	P_A case
d695	16,197	13,462 (−16.9 %)	10,705	9869 (−7.8 %)
g1023	17,925	14,953 (−16.6 %)	16,489	14,953 (−9.3 %)
p22810	166,800	154,310 (−7.5 %)	150,921	135,909 (−9.9 %)
p93791	502,876	453,923 (−9.7 %)	333,091	228,287 (−31.5 %)

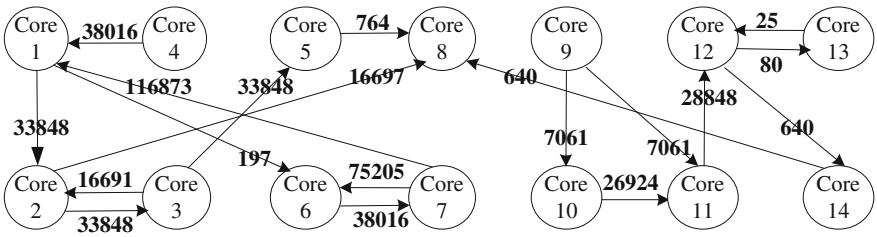


Fig. 7 Traffic diagram of g1023 circuit

The results in Table 2 showed that the P_A algorithm significantly optimized testing time. When data transfer width W is 16, the testing time is reduced by 12.67 % on average; while W is 32, the average reduction is 14.63 %.

NoC mapping needs the determined traffic information between various IP cores. Since g1023 has the same number of IP cores as H.263, it adopts traffic diagram of H.263 [12], while other three ITC'02 circuits adopt TGFF randomly generated traffic [13]. Figure 7 shows the traffic diagram of g1023.

According to the known traffic information and partition results, the layout of the NoC may be initially determined; then the actual communication power consumption(fitness) will be calculated based on the cores distance; and the appropriate genetic operations will execute(such as selection, crossover and mutation) to produce a new generation of IP core layout. The process will be repeated until achieving the best optimal layout.

For example, testing optimized grouping of g1023 is (4, 11), (5, 7, 12, 14), (2, 6, 8, 13), (1, 3, 9, 10). Applying the random mapping and the mapping result is showed as Fig. 8. The mapping cost of Fig. 8 is 4970 according to Eq. (3).

After adopting genetic algorithms, the mapping result of g1023 is shown in Fig. 9 and its mapping cost is 3214.

Applying the genetic algorithm to d695, g1023, p22810, p93791 circuits and the transfer width is 16 and 32. The mapping cost comparisons with random mapping are shown in Fig. 10.

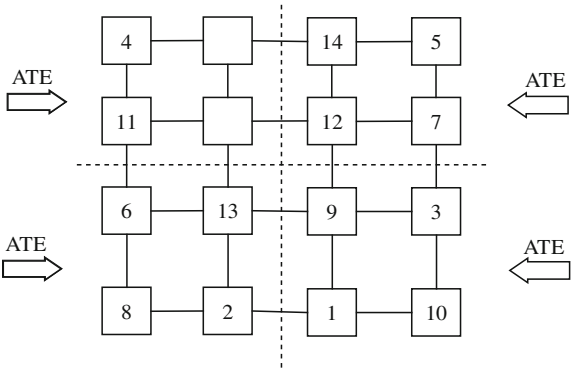


Fig. 8 Random mapping of g1023 circuit

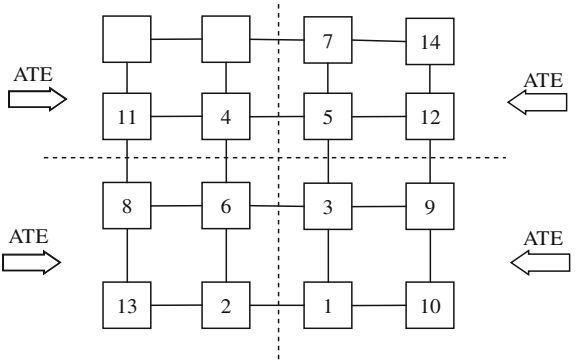


Fig. 9 Optimized mapping of g1023 circuit

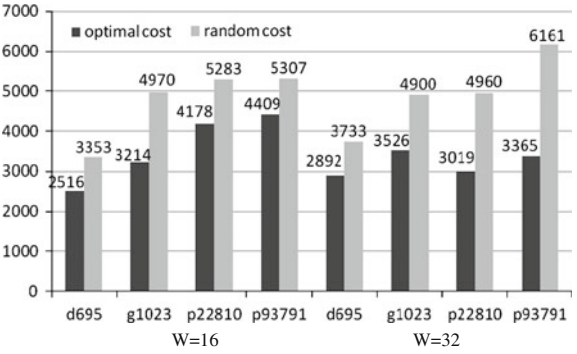


Fig. 10 Mapping cost comparisons

It can be concluded that the proposed mapping scheme is superior to the random mapping. When transfer width W is 16, the average reduction of mapping cost is 24.5 %; while W is 32, the average reduction is 33.75 %. It is obvious that the proposed mapping scheme optimizes testing efficiency while greatly reducing the mapping overhead.

5 Conclusion and Future Work

This paper described the testing time of each IP core on the condition that NoC scan test structure and transmission bandwidth are provided. Then it introduced a sectional NoC mapping scheme optimized for NoC IP cores testing. The main part of this scheme is a partition algorithm cooperating with optimized test structure and a NoC mapping scheme based on genetic algorithm. The former is applied to minimize testing time and the latter is aimed to acquire the minimum mapping overhead. The experiment results on ITC'02 benchmark circuits demonstrated the effectiveness of the testing optimization mapping scheme. Moreover, a collaborative optimization scheme for NoC testing and mapping is currently on the research.

Acknowledgment This work was supported by the Natural Science Foundation of China under Grant 61076019, 61106018 and 61376025, the Aeronautical Science Foundation of China under Grant 20140652008, Prospective joint research project of on the Integration of Industry, Education and Research of Jiangsu Province 2014003-05.

References

1. Bjerregaard T, Mahadevan SA (2006) Survey of research and practices of network-on-chip. ACM Comput Surv 38(1):1–51
2. Yang S, Li L, Gao M et al (2008) An energy- and delay-aware mapping method of NoC. Acta Electronica Sinica 36(5):937–942
3. Hu JC, Marculescu R (2005) Energy-and performance-aware mapping for regular NoC architectures. IEEE Trans Comput Aided Des Integr Circuits Syst 24(4):551–562
4. Wang L, Ling X (2010) Energy-and latency-aware NoC mapping based on chaos discrete particle swarm optimization. In: Proceedings of the communications and mobile computing conference CMC 2010, Shenzhen, pp 263–268. 12–14 Apr 2010
5. Chakrabarty K (2001) Optimal test access architectures for system-on-a-chip. ACM Trans Des Autom Electron Syst 6(1):26–49
6. Amory AM, Ferlini F, Lubaszewski M et al (2007) DfT for the reuse of networks-on-chip as test access mechanism. In: Proceedings of the 25th IEEE VLSI test symposium VTS 2007, Berkeley, pp 435–440. 6–10 May 2007
7. Agrawal M, Richter M, Chakrabarty K (2012) A dynamic programming solution for optimizing test delivery in multicore SOCs. In: Proceedings of the ITC'12 conference ITC 2012, Anaheim, pp 1–10. 5–8 Nov 2012
8. Marinissen EJ, Iyengar V, Chakrabarty K (2002) A set of benchmarks for modular testing of SOCs. In: Proceedings of the ITC'02 conference ITC 2002, Baltimore, pp 519–528. Oct 2002
9. Zhang Y, Wu N, Ge F (2011, 2013) The co-design of test structure and test data transfer mode for 2D-mesh NoC. In: IAENG transactions on engineering technologies—special edition of the world congress on engineering and computer science 2011, 2013, pp 171–184
10. Zhang Y, Wu N, Ge F (2014) Novel NoC mapping scheme optimized for testing time, lecture notes in engineering and computer science. In: Proceedings of the world congress on engineering and computer science WCECS 2014, San Francisco, pp 15–19. 22–24 Oct 2014
11. Liu C, Shi J, Cota E et al (2005) Power-aware test scheduling in network-on-chip using variable-rate on-chip clocking. In: Proceedings of the 23th IEEE VLSI test symposium VTS 2005, Palm Springs, California, pp 349–354. 1–5 May 2005

12. Hu J, Marculescu R (2003) Energy-aware mapping for tile-based NoC architectures under performance constraints. In: Proceedings of the VLSI test Asia and south Pacific design automat conference ASP-DAC 2003, Kitakyushu, pp 233–239. 21–24 Jan 2003
13. Dick PR, Rhodes LD, Wayne W (1998) TGFF: task graphs for free. In: Proceedings of the 6th international hardware/software co-design workshop CODES 98, Seattle, pp 97–101. 15–18 Mar 1998

An Extension of Hard Switching Memristor Model

Wanlong Chen, Xiao Yang and Frank Z. Wang

Abstract Memristor was initially introduced by Professor Leon Chua in 1971 as the fourth passive fundamental circuit element. In this chapter, we revise and extend the hard switching memristor model, that is developed based on the HP's memristor model. This model matches most of the memristor characteristics such as the pinched hysteresis loops (the fingerprint of memristive devices). Different materials of memristive devices could require different memristor models, which are sufficiently accurate and easily to do their simulations. The hard switching memristor model could be fit to specific materials of memristive devices and particular memristive systems. In some cases, this model is more reliable and flexible than the existing models of memristive devices.

Keywords Memristive system · Memristor · Memristor model · Modelling · Nano devices · Simulation · Window function

1 Introduction

The memristor was first conceptualized in 1971 by Professor Leon Chua. It is a passive electrical component which provides a functional relationship between magnetic flux and charge [1]. After the concept of memristive devices was proposed, some researchers proclaimed a few devices which could own similar behaviours as the memristors, however, none of the physical model had been created. It took 37 years for our engineering abilities to catch up with the idea of

W. Chen (✉) · X. Yang · F.Z. Wang
Future Computing Group, School of Computing, University of Kent, Kent, UK
e-mail: W.CHEN@kent.ac.uk

X. Yang
e-mail: X.YANG@kent.ac.uk

F.Z. Wang
e-mail: F.Z.Wang@kent.ac.uk

H.K. Kim et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-94-017-7236-5_22

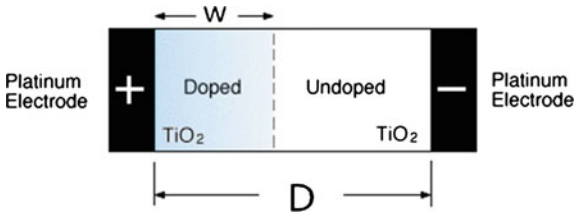


Fig. 1 Illustration of the two terminal memristor based on HP's model

Chua, the first physical memristor model was successfully created by a team of HP's labs in 2008 [2]. A thin film of titanium dioxide (TiO_2) is used in HP's memristor (shown in Fig. 1), that contains a stoichiometric layer of titanium dioxide (low resistance) and an oxygen deficient layer (high resistance), sandwiched between two platinum electrodes [2, 3]. The memristance of the HP's memristor is a sum of the resistances of the undoped regions (high resistance) and the doped regions (low resistance).

The research of memristive devices and systems [4] are revived by the realisation of HP's memristor and the promising features of memristors such as nano-scale size and memory effect. These features lead memristors to different applications, for examples, memories, neuromorphic engineering, and hybrid electronic circuits [5–7]. In [8–11], memristors are applied to content addressable memory. In [12–14], memristor-based neural networks were investigated under spiking-time-dependent plasticity (STDP).

In 2010, Chandra suggested that it was an interesting phenomenon as before the HP actual devices announced, not much attention been paid to the device upon the electrical engineering and nanotechnology [15, 16]. In fact, a nano-scale device has the phenomenon with its small voltages could produce enormous electric fields, which could produce significant non-linearity in the ionic transport [2]. Particularly, the boundary between the undoped and doped regions (TiO_2 and TiO_{2-x}) moves with a speed in a memristor, however this speed is limited when the boundary reaches either side [17]. In order to model the boundary condition effect and the non-linear drift of the memristor, a window function is often introduced.

In order to simplify and conquer the challenges of the memristive devices' design, we propose a new memristor model based on a window function. Unlike the most of existing memristor models, that are with symmetry functions; our model provides a flexible window function, which can provide not only symmetry functions but also asymmetry functions. There is no evidence that proves that the window function must be a symmetry function [18].

In this paper, by extending our preliminary work in [19], we systematically develop a rather complete set of properties and window functions for modelling the memristors of different materials. Section 2 reviews existing models of memristors. In Sect. 3, we present our hard switching memristor model with the explicit relationship between the memristance and the charge. The extension of the hard switching memristor model could be discussed in Sect. 4. Finally, the conclusion will be drawn in Sect. 5.

2 Background

2.1 Window Function

HP group was firstly introduced the window function on the memristor modelling with the boundary condition effect, manifestly it brought a significance to the memristor modelling in some cases [2, 3]. In the literature review, we will be discussing some current window function models.

Window function is defined as a function that owns a zero-value when the permissible value of the state variable is outside of some chosen intervals. A significant approach may be a rectangular window, which is the function shown the value of 0 at either boundary and the value of 1 at any other state between boundaries. Meanwhile, it is possibly to add a non-linear ion drift phenomenon to a different window to decrease the ion drift speed when it is approaching either end [17].

2.2 HP's Model

A thin film of titanium dioxide (TiO_2) is used in HP Memristor Model (shown in Fig. 1), D is the length of the device and w is the length of the doped region. The functions of the HP memristor model are given by

$$v(t) = ((R_{ON}\frac{w(t)}{D}) + R_{OFF}(1 - \frac{w(t)}{D}))i(t) \tag{1}$$

$$\frac{dw(t)}{dt} = \mu_V \frac{R_{ON}}{D} i(t) \tag{2}$$

where μ_V is the ion mobility, $v(t)$ is the voltage that applied to the memristor, $i(t)$ is the corresponding current of the device and $w(t)$ is limited to the range between 0 and D.

However, these state equations can not deal with the boundary non-linear dopant drift [20]. To overcome this, a window function $f(w)$ is multiplied to the right-hand side of Eq. (2), therefore,

$$\frac{dw(t)}{dt} = \mu_V \frac{R_{ON}}{D} i(t) \cdot f(w) \tag{3}$$

where $f(w) = w(D - w)/D^2$. However, the window function of HP is too simple to adapt different kinds of memristors since it lacks the flexibility.

2.3 Joglekar's Model

Joglekar proposed a window function, with regards to p as a positive exponent parameter [17]. They proposed this window function model to ensure zero drift at the boundaries.

$$f(x)_p = 1 - (2x - 1)^{2p} \tag{4}$$

where p is a positive exponent parameter.

Figure 2 shows the behaviour of Eq. (4) for some different values of p. It is clear that:

- for $0 \leq x < 0.5$, the greater x value the higher $f(x)$;
- for $0.5 < x \leq 1$, the greater x value the lower $f(x)$;
- when $x = 1$ and $x = 0$, $f(x)_{min}$ reaches, this could offer the zero drift at both boundaries of the devices;
- when $x = 0.5$, $f(x)_{max}$ arrives and this window function is symmetric regard to $x = 0.5$.

2.4 Prodromakis' Model

Prodromakis's model [21] is likely a kind of upgrade vision of Joglekar's model. This window function allows to change the $f(x)_{max}$ by setting the scaling parameter j and is of the form:

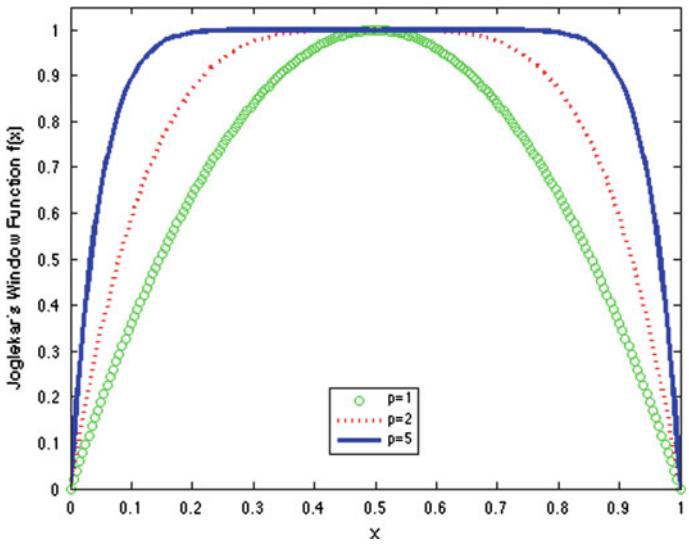


Fig. 2 The form of Joglekar's window function Eq. (4) with different values of its controllable parameter p

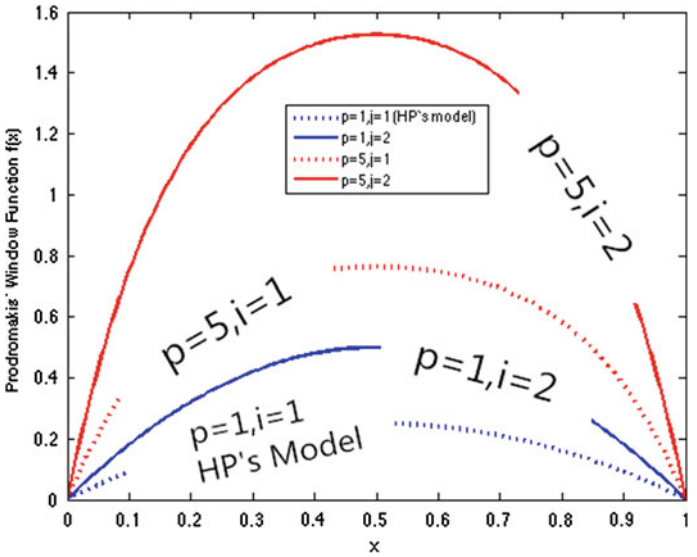


Fig. 3 The form of Prodromakis' window function Eq. (5) with different values of its controllable parameters p and j

$$f(x)_{p,j} = j(1 - [(x - 0.5)^2 + 0.75]^p) \tag{5}$$

where p is a controllable parameter like the 'p' in Joglekar's model, and j is another controllable parameter to scale $f(x)$ upward or downward. It improves the scalability and flexibility. This window function is illustrated in Fig. 3 for different values of p and j. For $p = 1$ and $j = 1$, it becomes identical to HP's window function $f(x) = x(1 - x)$. Also this window function is symmetric about $x = 0.5$, it lacks the necessary flexibility.

2.5 *Biolek's Model*

$$f(x)_{p,i} = \begin{cases} 1 - (x - 1)^{2p}; & i \leq 0 \\ 1 - (x)^{2p}; & i > 0 \end{cases} \tag{6}$$

There is also a controllable parameter p, a positive integer, which is used to control the flatness of the window around the maximum in a similar way to Joglekar's model. In Prodromakis' paper [21], it was suggested for 'terminal state problem' when the device cannot be further adjusted, there could applies an alternative window function, Biolek's Model where suggested an additional

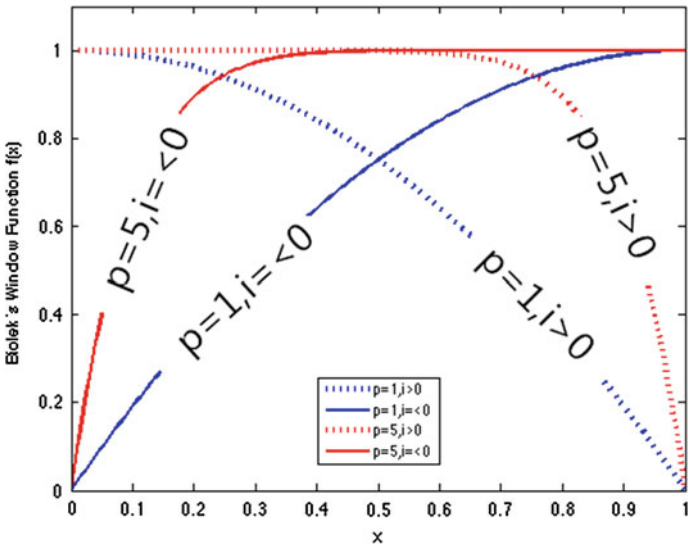


Fig. 4 The form of Biolek's window function Eq. (6) with different values of its controllable parameter p

parameter i, which is the current of the device, otherwise it is about the polarity of the input. Figure 4 shows that a positive current provides an increment on width of the doped region, in a way that '$0 \mapsto 1$', and a negative current offers a decrement on width of the doped region. However, this window function could not solve the boundary condition effect properly.

3 HSMM: Hard Switching Memristor Model

A new window function is proposed in this paper to model the memristor in the following form:

$$f(w) = \begin{cases} \frac{w^3}{a^3} & a \geq w \geq 0 \\ \frac{(1-w)^3}{(1-a)^3} & 1 \geq w > a \end{cases} \tag{7}$$

where $0 < a < 1$.

The window function has two sub-domains [0, a] and (a, 1] which constitute the whole domain range [0, 1]. In this work, we assume that $D = 1, R_{on} = 1, \mu_V = 1$.

So we get

$$\frac{dw(t)}{dt} = i(t) \cdot f(w) \tag{8}$$

Since

$$i(t) = \frac{dq(t)}{dt} \tag{9}$$

which yields the following formula:

$$\frac{dw(t)}{dt} = \frac{dq(t)}{dt} \cdot f(w) \tag{10}$$

For $a \geq w \geq 0, f(w) = \frac{w^3}{a^3}$, which yields:

$$\frac{dw(t)}{dt} = \frac{dq(t)}{dt} \cdot \frac{w^3}{a^3} \tag{11}$$

Then,

$$\frac{dw(t)}{w^3} = \frac{dq(t)}{a^3} \tag{12}$$

Let $t_0 = 0, w_0 = w(t_0)$ and $q_0 = q(t_0) = 0$. And when $t_a > \hat{t} > 0$, where t_a is the time when the boundary arrives the value a, let $\hat{w} = w(\hat{t})$ and $\hat{q} = q(\hat{t})$.

$$\int_{w_0}^{\hat{w}} w^{-3} dw(t) = \int_{0}^{\hat{q}} a^{-3} dq(t) \tag{13}$$

Then,

$$-\frac{1}{2w(t)^2}\Bigg|_{w_0}^{\hat{w}} = \frac{q(t)}{a^3}\Bigg|_{0}^{\hat{q}} \tag{14}$$

Then,

$$-\frac{1}{2 \cdot (\hat{w})^2} + \frac{1}{2 \cdot (w_0)^2} = \frac{\hat{q}}{a^3} \tag{15}$$

Then,

$$\hat{w} = \frac{1}{\sqrt{\frac{-2\hat{q}}{a^3} + \frac{1}{w_0^2}}} \tag{16}$$

When the boundary arrives the value a, where $\hat{w}$ of Eq. (16) is a, so that

$$q(\hat{w} = a) = \frac{a^3}{2}\left(\frac{1}{w_o^2} - \frac{1}{a^2}\right) \tag{17}$$

Noting that a is the boundary of the subregions $a \geq w \geq 0$ and $1 \geq w > a$, and let

$$Q = q(\hat{w} = a) \tag{18}$$

For $1 \geq w > a, f(w) = \frac{(1-w)^3}{(1-a)^3}$, which yields:

$$\frac{dw(t)}{dt} = \frac{dq(t)}{dt} \cdot \frac{(1-w)^3}{(1-a)^3} \tag{19}$$

Then,

$$\frac{dw(t)}{(1-w)^3} = \frac{dq(t)}{(1-a)^3} \tag{20}$$

Noting that t_a is the time when the boundary arrives the value a. From Eq. (18), let $a = w(t_a)$ and $Q = q(t_a)$. And when $\hat{t} > t_a$, let $\hat{w} = w(\hat{t})$ and $\hat{q} = q(\hat{t})$.

$$\int_a^{\hat{w}} (1-w)^{-3} dw(t) = \int_Q^{\hat{q}} (1-a)^{-3} dq(t) \tag{21}$$

Then,

$$\frac{1}{2(1-w(t))^2}\Bigg|_a^{\hat{w}} = \frac{q(t)}{(1-a)^3}\Bigg|_Q^{\hat{q}} \tag{22}$$

Then,

$$\frac{1}{2 \cdot (1-\hat{w})^2} - \frac{1}{2 \cdot (1-a)^2} = \frac{\hat{q} - Q}{(1-a)^3} \tag{23}$$

Then,

$$\hat{w} = 1 - \frac{1}{\sqrt{\frac{2}{(1-a)^3} \cdot (\hat{q} - Q) + \frac{1}{(1-a)^2}}} \tag{24}$$

Thus, combining the solutions of the two subregions $a \geq w \geq 0$ and $1 \geq w > a$, we can have the complete solutions:

$$w(q) = \begin{cases} \frac{1}{\sqrt{\frac{-2q}{a^3} + \frac{1}{w_0^2}}}; & Q \geq q \\ 1 - \frac{1}{\sqrt{\frac{2}{(1-a)^3} \cdot (q-Q) + \frac{1}{(1-a)^2}}}; & q > Q \end{cases} \tag{25}$$

where $Q = \frac{a^3}{2}(\frac{1}{w_o^2} - \frac{1}{a^2})$. Here '$Q$' is the boundary point of the hybrid function.

Since the memristance of the HP's memristor is a sum of the resistances of the doped regions (low resistance) and the undoped regions (high resistance), let $R_{off}/R_{on} = k$, so that,

$$M(w(q)) = R_{ON} \times w(q) + R_{OFF} \times (1 - w(q)) \tag{26}$$

Therefore, combing Eqs. (25) and (26), the explicit relation of the memristance and the charge could be solved as the following:

$$M(q) = \begin{cases} \frac{1-k}{\sqrt{\frac{-2q}{a^3} + \frac{1}{w_0^2}}} + k; & Q \geq q \\ \left(1 - \frac{1}{\sqrt{\frac{2}{(1-a)^3} \cdot (q-Q) + \frac{1}{(1-a)^2}}}\right)(1-k) + k; & q > Q \end{cases} \tag{27}$$

where $Q = \frac{a^3}{2}(\frac{1}{w_o^2} - \frac{1}{a^2})$, which is the boundary point of the hybrid function.

4 Discussion

The most important advantage of this new memristor model is that the formula of the memristance $M(q)$ could be obtained through Eq. (27), as well as the numerical value of the memristance $M(q)$ can be calculated by setting all other parameters.

As another advantage of this model, it can provide a more functional and flexible window function. The Heaviside step function Eq. (28) is added to our new proposed window function, that is flexiable to set and model the memristor behavior.

$$H(x) = \begin{cases} 0; & x < 0 \\ 1; & x \geq 0 \end{cases} \tag{28}$$

When the acceleration of the boundary between the doped regions and the undoped regions is increasing, we let $x \geq 0$; otherwise, we let $x < 0$.

The following new window function of the extension of the hard switching memristor model is proposed:

$$f(w) = b \cdot \left(H(i) \cdot \left(\frac{w}{a}\right)^c + H(-i) \cdot \left(1 - \left(\frac{w-a}{a}\right)^c\right) + H(j) \cdot \left(\frac{1-w}{1-a}\right)^c + H(-j) \cdot \left(1 - \left(\frac{w-a}{1-a}\right)^c\right)\right) \tag{29}$$

where $w \in [0, 1]$ is the state variable, $a \in (0, 1), b \in \mathbb{R}^+, c > 1$ and $i, j \in \mathbb{R}$. This new extension of the hard switching memristor model satisfies the boundary conditions enforcing zero drift at $w = 0$ and $w = 1$, since $f(0) = f(1) = 0$. This window function has a single maximum in the range $w \in [0, 1]$, when $w = a$. This may be extended in the future, since we are going to develop a more comprehensive and functional memristor model with piecewise window functions, so that multiple maximums should be achieved. i is the acceleration of the boundary of the memristor which accelerates from rest to a certain speed, j is the acceleration of the boundary of the memristor which decelerates to rest. The parameter a is the controllable parameter for changing the skewness of the window. For $a = 0.5$, the window function is symmetric regard to $a = 0.5$, which is the same with most of current reported models such as HP's model [2] and Joglekar's model [17]. For $a \in (0, 0.5)$, it is left skewed and for $a \in (0.5, 1)$, it is right skewed. The skewed window function is not mentioned by previously memristor models, so our model could capture some special features of the memristors and some memristive systems. Compared with Prodromakis' model [21], in some cases where the dopant's drift is such that $f(x)_{max}$ is greater or smaller than one, other controllable parameter b is provided in our proposed window function in order to adjust this particular issue. The parameter c is used to adjust the flatness of the window function. By introducing the parameters a, b and c, the plots of our proposed window function could be more flexible to adjust different memristor models.

5 Conclusion

The extension of the hard switching memristor model based on a window function with multiple controllable parameters, made up on the basis of HP's memristor state equations. The shape of the controllable window function could be flexibly set up to model different kinds of memristive devices and systems.

During past years, the scientific community is witnessing an enormous development of new computing technologies and more specifically in the field of nanotechnology. With developing nano-level memristors and discovering more memristor physical materials, memristor models would become more significant. Our model aims to identify and model the memristor in a simple and convenient way, which could make the design and analysis of the memristors and the memristive systems as easy and convenient as possible.

Theoretically, memristive devices are significant as they have the capability to extend the functionalities via incorporating the analog computation and the self-programming neural network. It helps approaching information processing as the way human brains process.

References

1. Chua L (1971) Memristor-the missing circuit element. IEEE Trans Circuit Theory 18 (5):507–519
2. Strukov DB, Snider GS, Stewart DR, Williams RS (2008) The missing memristor found. Nature 453(7191):80–83
3. Williams R (2008) How we found the missing memristor. Spectr IEEE 45(12):28–35
4. Chua LO, Kang SM (1976) Memristive devices and systems. Proc IEEE 64(2):209–223
5. Wang F, Chua L, Yang X, Helian N, Tetzlaff R, Schmidt T, Li C, Carrasco J, Chen W, Chu D (2013) Adaptive neuromorphic architecture (ana). Neural Networks Official J Int Neural Network Soc 45:111
6. Yang X, Chen W, Wang FZ (2014) The staircase memristor and its applications. In: 21st IEEE international conference on electronics, circuits and systems (ICECS). IEEE, pp 259–262
7. Gandhi G, Aggarwal V, Chua LO (2014) The detectors used in the first radios were memristors. In: Adamatzky, A, Chua L (eds) Memristor networks. Springer, Berlin, pp 53–66
8. Yang X, Chen W, Wang FZ (2013) A memristor-cam (content addressable memory) cell: new design and evaluation. In: International conference on computer science and information technology, pp 1045–1048
9. Chen W, Yang X, Wang FZ (2013) Delayed switching applied to memristor content addressable memory cell. In: Proceedings of the world congress on engineering (WCE 2013), London, UK. Lecture notes in engineering and computer science, vol 1, pp 354–357, 3–5 July 2013
10. Eshraghian K, Cho K-R, Kavehei O, Kang S-K, Abbott D, Kang S-MS (2011) Memristor mos content addressable memory (mcam): hybrid architecture for future high performance search engines. IEEE Trans Very Large Scale Integr (VLSI) Syst 19(8):1407–1417
11. Chen W, Yang X, Wang FZ (2014) Memristor content addressable memory. In: IEEE/ACM international symposium on nanoscale architectures (NANOARCH). IEEE, pp 83–87
12. Yang X, Chen W, Wang FZ (2013) A supervised spiking time dependant plasticity network based on memristors. In: IEEE 14th international symposium on computational intelligence and informatics (CINTI). IEEE, pp 447–451
13. Wang FZ, Helian N, Wu S, Yang X, Guo Y, Lim G, Rashid MM (2012) Delayed switching applied to memristor neural networks. J Appl Phys 111(7):07E317
14. Yang X, Chen W, Wang FZ (2014) The memristor-based associative learning network with retention loss. In: IEEE 15th international symposium on computational intelligence and informatics (CINTI). IEEE, pp 249–253
15. Chandra S (2010) On the discovery of a polarity-dependent memory switch and/or memristor (memory resistor). IETE Tech Rev 27(2):179–180
16. Williams RS (2010) Reply to–on the discovery of a polarity-dependent memory switch and/or memristor (memory resistor). IETE Tech Rev 27(2):181–182
17. Joglekar YN, Wolf SJ (2009) The elusive memristor: properties of basic electrical circuits. Eur J Phys 30(4):661
18. Mu X, Yu J, Wang S (2015) Modeling the memristor with piecewise linear function. Int J Num Model. Electron Networks Devices Fields 28(1):96–106

19. Chen W, Yang X, Wang F (2014) Hsmm: hard switching memristor model. In: Proceedings of the world congress on engineering and computer science notes in engineering and computer science (WCECS 2014), San Francisco, USA, vol 1, pp 11–14, 22–24 Oct 2014
20. Juntang Y, Xiaomu M, Xiangming X, Shuning W (2013) A memristor model with piecewise window function. Radioengineering 22(4):969–974
21. Prodromakis T, Peh BP, Papavassiliou C, Toumazou C (2011) A versatile memristor model with nonlinear dopant kinetics. IEEE Trans Electron Devices 58(9):3099–3105

On Circulant Graphs with the Maximum Leaf Number Property and Its One-to-Many Communication Scheme

Felix P. Muga II

Abstract This paper determines the maximum leaf number and the connected domination number of some undirected and connected circulant networks. We shall tackle this problem by working on the largest possible number of vertices between two consecutive jump sizes. This paper also determines the communication steps of its one-to-many communication scheme.

Keywords Broadcast scheme · Circulant graph · Height of a tree · Maximum leaf number · Minimum connected domination number · Spanning tree

1 Introduction

Let G be a simple connected graph. The spanning tree of G is its subgraph that contains all its vertices and has no cycles. A maximum leaf spanning tree (MLST) of G has the most possible number of leaf vertices among all the spanning trees of G.

The number of leaf vertices of a MLST of G denoted by $\ell(G)$ is the *maximum leaf number* of G.

The problem of finding a maximum leaf spanning tree of G is the *MLST* problem.

A connected dominating set of G is a subset D of the vertex set of G that induces a connected subgraph of G such that every vertex in G is either in D or adjacent to a vertex in D. The minimum connected dominating set (MCDS) has the smallest possible number of vertices among all connected dominating sets of G.

The number of elements of a *MCDS* denoted by $d(G)$ is the *connected domination number* of G (Sampathkumar and Walikar [3]).

Douglas [1] showed that $d(G) + \ell(G)$ is the order of G.

Hence, the problem of finding the *MCDS* of G is equivalent to the *MLST* problem.

F.P. Muga II (✉)
School of Science and Engineering, Ateneo de Manila University, Quezon City, Philippines
e-mail: fmuga@ateneo.edu

H.K. Kim et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-94-017-7236-5_23

2 Leaf Number of a Family of Circulant Networks

Let N and k be integers such that $N \geq 3$ and $k \geq 1$.

Consider a connected and undirected circulant network $G = C(n; \pm(s_1, s_2, \ldots, s_k))$ of order N and degree $2k$ such that $N \geq 2k + 1$.

The vertices s_i and $N - s_i$, for all $i = 1, 2, \ldots, k$ are called the *jump sizes* of G. The jump sizes under consideration are ordered such that $1 = s_1 < s_2 < \cdots < s_k < \frac{N}{2}$.

Let the vertices of G be labelled as $0, 1, \ldots, N - 1$.

Suppose S is the ordered list that contains all the jump sizes of G. Then we can also write G as $C(N; S)$.

We shall find $d(G_i)$ and $\ell(G_i)$ for $i = 1, 2$ where

1. $G_1 = C(N; \pm(iq + 1)), \forall i = 0, 1, \ldots, k - 1, q \geq 1$, and $N = (2k - 1)q + 2$.
2. $G_2 = C(n; \pm(s_1, s_2))$ where $N = (2k - 1)q + r + 2$, $s_1 = iq + 1$, and $s_2 = (k - r_1 + j)q + j + 2$, $\forall i = 0, 1, \ldots, k - r_1 - 1$, $\forall j = 0, 1, \ldots, r_1 - 1$, $q \geq 1$, and $r = 1, 2, \ldots, 2k - 2$ with $r_1 = \left\lfloor \frac{r}{2} \right\rfloor$.

Note that if $r = 1$, then $r_1 = 0$ and the jump of sizes of G_2 are similar to those of G_2.

We shall show that

$$\mathscr{M}(n, 2k) = \begin{cases} d(G_1) & \text{if} \quad \text{r = 0} \\ d(G_2) & \text{if} \quad \text{r } > 0 \end{cases}$$

$$\mathscr{L}(n, 2k) = \begin{cases} \ell(G_1) & \text{if} \quad \text{r = 0} \\ \ell(G_2) & \text{if} \quad \text{r } > 0 \end{cases}$$

where

$$\mathscr{M}(n,\ 2k) \stackrel{\text{def}}{=} \min\{d(G) | \forall G = C(N; S), |S| = 2k\}$$

$$\mathscr{L}(n, 2k) \stackrel{\text{def}}{=} \max\{\ell(G) | \forall G = C(N; S), |S| = 2k\}$$

over all connected and undirected circulant networks G of order $N = (2k - 1)q + r + 2$ and degree $2k$ where $q \geq 1, k \geq 1$, and $r = 0, 1, \ldots, 2k - 2$.

In the rest of the paper we shall always assume that G is a circulant network of order n and degree $2k$ such that its jump sizes are in the ordered list S of length $2k$ whose first term is 1.

Theorem 1(Muga[2])

$$\left\lceil \frac{N-2}{2k-1} \right\rceil \leq \mathscr{M}(N,2k) \leq n-1$$
$$1 \leq \mathscr{L}(N,2k) \leq \left\lfloor \frac{(2k-2)N+2}{2k-1} \right\rfloor$$

Proof Consider a spanning tree $\mathscr{T}$ of G rooted at u.

Let $A(v) = [v, p[v], L[v]]$ be the adjacency list of vertex v where $p[v]$ is the parent vertex of v and $L[v]$ is the list of child vertices of v in $\mathscr{T}$.

Since u is the root of $\mathscr{T}$, it follows that $A(u) = [u,$ none, $L[u]]$ where $L[u] \subseteq S$.

The adjacency lists of the other internal vertices are $[v, p[v], L[v]]$, where $|L[v]| > 0$, and those of the leaf vertices are $A(z) = [z, p[z], L[z]]$, where $L[z]$ is empty.

The number of internal vertices of $\mathscr{T}$ is the maximum if each internal vertex has one child vertex only.

Thus, $d(G) \leq N-1$ and $1 \leq \ell(G)$. Hence,

$$\mathscr{M}(N,2k) \leq d(G) \leq N-1$$
$$1 \leq \ell(G) \leq \mathscr{L}(N,2k)$$

Since G is $2k$-regular and since the root u has no parent vertex, it follows that $|L[u]]| \leq 2k$, and since each of the other internal vertices of $\mathscr{T}$ has a parent vertex, we have $|L[v]| \leq 2k-1$ for all the other internal vertices v of $\mathscr{T}$.

Hence, the number of internal vertices of $\mathscr{T}$ can be minimized if the list of child vertices are filled up to its maximum capacity.

The total number of child vertices is $N-1$ since the root has no parent vertex.

If a vertex which is adjacent to the root is first chosen to be its child vertex, then we have $N-2$ child vertices to be distributed to all the $L[v]$'s of the internal vertices, each of which can accommodate up to $2k-1$ child vertices.

Hence, $\left\lceil \frac{N-2}{2k-1} \right\rceil \leq d(G)$. This inequality is true for all connected and undirected circulant networks of order n and degree $2k$.

Thus, $\left\lceil \frac{N-2}{2k-1} \right\rceil \leq \mathscr{M}(N,2k)$. Consequently, the number of leaf vertices of each of the spanning trees of G is as large as

$$n - \left\lceil \frac{N-2}{2k-1} \right\rceil = n + \left\lfloor -\frac{N-2}{2k-1} \right\rfloor = \left\lfloor \frac{(2k-2)N+2}{2k-1} \right\rfloor.$$

Hence, $\ell(G) \leq \left\lfloor \frac{(2k-2)N+2}{2k-1} \right\rfloor$. This inequality is true also for all connected and undirected circulant networks of order N and degree $2k$.

Thus, $\mathscr{L}(N, 2k) \leq \left\lfloor \frac{(2k-2)N+2}{2k-1} \right\rfloor$.

Therefore,

$$\left\lceil \frac{n-2}{2k-1} \right\rceil \leq \mathscr{M}(N, 2k) \leq N-1$$
$$1 \leq \mathscr{L}(N, 2k) \leq \left\lfloor \frac{(2k-2)n+2}{2k-1} \right\rfloor.$$

□

Let $G = C(N; S)$ be a connected and undirected circulant network of order N and degree $2k$ and let $\delta(G)$ be the largest number of vertices between two consecutive jump sizes in G.

Theorem 2(Muga[2]) *Suppose that* $N = (2k-1)q + r + 2$ *where* $k, q \in \mathbb{Z}^+$, *and* $r = 0, 1, \ldots, 2k-2$. *Then*

1. $\delta(G) = q - 1$, *if* $r = 0$ *and* $S = \{\pm(iq+1)\}$, $\forall i = 0, 1, \ldots, k-1$, *or.*
2. $\delta(G) = q$, if $r > 0$ *and* $S = \{\pm s_{1,i}, \pm s_{2,j}\}$ *where* $s_{1,i} = \pm(iq+1)$, $s_{2,j} = \pm((k - r_1 + j)q + j + 2)$, $\forall i = 0, 1, \ldots, k - r_1 - 1$, $\forall j = 0, 1, \ldots, r_1 - 1$, *with* $r_1 = \left\lfloor \frac{r}{2} \right\rfloor$.

Note that the elements of S are computed as residues under modulo N,

Proof

1. If $r = 0$, then $N = (2k-1)q + 2$ and if $s_i = iq + 1$, for $i = 0, 1, \ldots, k-2$, then

$$s_{i+1} - s_i - 1 = q - 1,$$
$$(N - s_i) - (N - s_{i+1}) - 1 = q - 1,$$
$$(N - s_{k-1}) - s_{k-1} - 1 = q - 1$$

   Thus, $\delta(C(N; \pm(iq+1))) = q - 1$.
2. For $i = 0, 1, \ldots, k - r_1 - 1$, and for $j = 0, 1, \ldots, r_1 - 1$, with $r_1 = \left\lfloor \frac{r}{2} \right\rfloor$, let $s_{1,i} = iq + 1$ and $s_{2,j} = (k - r_1 + j)q + j + 2$.
   We find the number of vertices between two consecutive jump sizes of $G = C(n; \pm(s_{1,i}, s_{2,j}))$

$$\begin{aligned} s_{1,i+1} - s_{1,i} - 1 &= q - 1 \\ (N - s_{1,i}) - (N - s_{1,i+1}) - 1 &= q - 1 \\ s_{2,j+1} - s_{2,j} - 1 &= q \\ (N - s_{2,j}) - (N - s_{2,j+1}) - 1 &= q \\ s_{2,0} - s_{1,k-r_1-1} - 1 &= q \\ (N - s_{1,k-r_1-1}) - (N - s_{2,0}) - 1 &= q \end{aligned}$$

and in the two consecutive jump sizes at the center of S, we have $(N - s_{2,r_1-1}) - s_{2,r_1-1} - 1 = q + r - 2r_1 - 1$.
If r is odd, then $2r_1 = r - 1$. If r is even, then $2r_1 = r$.
Thus,

$$(N - s_{2,r_1-1}) - s_{2,r_1-1} - 1 = \begin{cases} q & \text{if } r \text{ is odd, or} \\ q - 1 & \text{if } r \text{ is even} \end{cases}$$

Hence, $\delta(G) = q$. □

3 An Algorithm to Construct S, p and L of a Spanning Tree of G_1 or G_2

Let $\mathscr{A} = \langle A(0), A(1), \ldots, A(n-1) \rangle$ be a sequence of adjacency lists of a tree $\mathscr{T}$ rooted at 0 such that the adjacency list of the root 0 is $[0, \text{None}, S]$ and for $v \neq 0$, we have $A(v) = [v, p[v], L[v]]$ where $p[v]$ is the parent vertex of v in $\mathscr{T}$ and $L[v]$ is the list of child vertices of v in $\mathscr{T}$.

Suppose that the respective data structure of p and L are dictionaries where

1. $p = \{0 : \text{None}, 1 : p_1, n-1 : p_{n-1}\}$ such that $p[0] = \text{None}$, and $p[v] = p_v$, the parent vertex of v, for each $v = 1, 2, \ldots, N-1$.
2. $L = \{0 : S, 1 : L_1, \ldots, , N-1 : L_{N-1}\}$ such that $L[0] = S$, and $L[v] = L_v$, the list of child vertices of v, for each $v = 1, 2, \ldots, N-1$.

ALGORITHM 1 Find S, p and L in G_1 or G_2.

Require: $N, k \in \mathbb{Z}$, $n \geq 3$ and $k \geq 1$
Ensure: S, p, and L
if $n - 2 * k - 1 < 0$ **then**
  return None
end if
$S \leftarrow [\,]$, $p \leftarrow \{\,\}$, **and** $L \leftarrow \{\,\}$
for $i = 0$ **to** $n - 1$ **do**
  $p[i] \leftarrow$ None **and** $L[i] \leftarrow [\,]$
end for
Find $(q, r) = \text{divmod}(N - 2,\ 2 * k - 1)$ {where q is the quotient and r is the remainder when $N - 2$ is divided by $2k - 1$.}
if $r = 0$ **or** $r = 1$ **then**
  for $i = 0$ **to** $k - 1$ **do**
    $u = i * q + 1$
    $S \leftarrow S + \left[u, N - u\right]$ {Jump sizes in G_1}
  end for
else $\{r \geq 2\}$
  compute $r_1 = \left\lfloor \frac{r}{2} \right\rfloor$
  for $i = 0$ **to** $k - r_1 - 1$ **do**
    $u_1 = i * q + 1$
    $S \leftarrow S + \left[u_1,\ N - u_1\right]$ {jump sizes in G_2}
  end for
  for $j = 0$ **to** $r_1 - 1$ **do**
    $u_2 = (k - r_1 + j) * q + j + 2$
    $S \leftarrow S + \left[u_2,\ N - u_2\right]$ {another jump sizes in G_2}
  end for
end if
Sort S in ascending order.
$L[0] \leftarrow S$ {All the jump sizes are child vertices of the root}
for all $v \in S$ **do**
  $p[v] \leftarrow 0$ {0 is the parent vertex of all the jump sizes}
end for
for $i = 0$ **to** $2k - 2$ **do**
  compute $d = S[i + 1] - S[i] - 1$
  if $d > 0$ **then**
    for $j = 1$ **to** d **do**
      $v \leftarrow S[i] + j$ $\{v \notin S\}$
      if $v \leq q$ **then**
        $p[v] \leftarrow v - 1$ **and** $L[v - 1] \leftarrow L[v - 1] + \left[v\right]$
      else $\{v > q\}$
        $p[v] \leftarrow j$ **and** $L[j] \leftarrow L[j] + \left[v\right]$.
      end if
    end for
  end if
end for
return S, p **and** L.

Theorem 3(Muga[2]) *Let $N, k \in \mathbb{Z}$ such that $N \geq 3$ and $k \geq 1$ and consider S, p and L generated in Algorithm 1.*

Suppose that $\mathscr{A} = \langle [v,\ p[v],\ L[v]] \mid \forall v \in V(G(n; S)) \rangle$ is the sequence of adjacency lists of the subgraph $\mathscr{T}$ of G. Then

1. *$\mathscr{T}$ is a spanning tree of G.*
2. *The number of internal vertices of $\mathscr{T}$ is q if $r = 0$, or $q + 1$ if $r \geq 1$, and*
3. *The number of leaf vertices of $\mathscr{T}$ is $N - q$ if $r = 0$, or $N - q - 1$ if $r \geq 1$.*

Proof $\mathscr{T}$ is a spanning subgraph of G since it has all the vertices of G where $G = G_1$ or $G = G_2$.

We shall show that $\mathscr{T}$ is a subtree of G.
Since 0 has no parent vertex in $\mathscr{T}$, it is the root of $\mathscr{T}$.
Let s and t be two consecutive jump sizes in G such that $q = t - s - 1$.
Let v be a nonzero vertex in G. Then v is in one of two closed intervals: $I_1 = [1,\ q]$, or $I_2 = [q + 1,\ N - 1]$.

1. Suppose $v_1 \in I_1$. Since $s_1 = 1$ and $s_2 = q + 1$, it follows that vertex 1 is the only vertex in I_1 that is a child vertex of the root 0.
   Since $p[v_1] = v_1 - 1$ for all $v_1 \in I_1$, it follows that every vertex in $[1,\ q)$ is a parent vertex in $\mathscr{T}$ and that every vertex in I_1 is connected to the root by a path that passes through vertex 1.
   Since a child vertex has a single parent vertex only, it follows that the path between $v \in I_1$ and 0 is unique for every vertex in I_1.
   (a) If $r = 0$, then $\delta(G_1) = q - 1$. Thus, vertex q has no child vertex in $\mathscr{T}$.
   (b) If $r \geq 1$, then $\delta(G_1) = q$ or $\delta(G_2) = q$.
   Thus, q has a child vertex in $\mathscr{T}$. Its child vertex is u where $s < u < t$ and $u = s + q$.
   Hence, the number of internal vertices of $\mathscr{T}$ is $q - 1$ if $r = 0$, or it is q, if $r \geq 1$.
2. Suppose $v_2 \in I_2$.
   If $v_2 \in S$, then $p[v_2] = 0$.
   If $v_2 \notin S$, then there exist two consecutive jump sizes s_1 and t_1 such that $s_1 < v_2 < t_1$ and $v_2 = s_1 + u$ for some u in $[1,\ t_1 - s_1 - 1]$.
   Thus, $p[v_2] = u$, since $u \leq t_1 - s_1 - 1 \leq q$.
   This implies that $p[v_2] \in I_1$.
   Hence, the parent vertex for every vertex in I_2 is in I_1.
   This means that $L(v_2)$, which was initialized as the empty list, is always empty.
   Consequently, every vertex in I_2 is a leaf vertex.
   Since each vertex $v_1 \in I_1$ is connected to the root 0 by a path that is unique to v_1 and 0, and since $p[v_2] \in I_1$ for every vertex $v_2 \in I_2$, it follows that v_2 is connected to the root 0 by a unique path in $\mathscr{T}$.
   Hence, the number of leaf vertices of $\mathscr{T}$ is $N - q - 1$ if $r = 0$, or it is $N - q$, if $r \geq 1$. □

Let $m(\mathscr{T})$ and $l(\mathscr{T})$ be the number of internal vertices and the number of leaf vertices of a tree $\mathscr{T}$.

Theorem 4 *Suppose that G is the circulant network G_1 or G_2 of order N and degree $2k$ where $N \geq 3$ and $k \geq 1$. Then $d(G) = \left\lceil \frac{N-2}{2k-1} \right\rceil$ and $\ell(G) = \left\lfloor \frac{(2k-2)N+2}{2k-1} \right\rfloor$.*

Therefore, $\mathscr{M}(N, 2k) = d(G)$, and $\mathscr{L}(N, 2k) = \ell(G)$.

Proof By Theorem 3, if $\mathscr{T}$ is the spanning tree of G generated by $\mathscr{A}$ where $G = G_1$ or $G = G_2$ is of order $N = (2k-1)q + r + 2$ and degree $2k$ with $r = 0, 1, \ldots, 2k-2$, $k \geq 1$ and $q \geq 1$, then

$$m(\mathscr{T}) = \begin{cases} q & \text{if } r = 0, \text{ or} \\ q+1 & \text{if } r \geq 1. \end{cases},$$

$$l(\mathscr{T}) = \begin{cases} N-q & \text{if } r = 0, \text{ or} \\ N-q-1 & \text{if } r \geq 1. \end{cases}$$

Since $\mathscr{M}(N, 2k) \geq \left\lceil \frac{N-2}{2k-1} \right\rceil$ and

$$\left\lceil \frac{N-2}{2k-1} \right\rceil = \left\lceil \frac{(2k-1)q + r}{2k-1} \right\rceil$$

$$\left\lceil \frac{N-2}{2k-1} \right\rceil = q + \left\lceil \frac{r}{2k-1} \right\rceil = m(\mathscr{T})$$

Thus, $m(\mathscr{T}) \leq \mathscr{M}(N, 2k)$.

However, by the minimality of $\mathscr{M}(N, 2k)$ and $d(G)$, we have $\mathscr{M}(N, 2k) \leq d(G) \leq m(\mathscr{T})$,

Therefore, $\mathscr{M}(N, 2k) = d(G) = \left\lceil \frac{N-2}{2k-1} \right\rceil$.

Also, $\mathscr{L}(N, 2k) \leq \left\lfloor \frac{(2k-2)N+2}{2k-1} \right\rfloor$ and

$$\begin{aligned} \left\lfloor \frac{(2k-2)N+2}{2k-1} \right\rfloor &= \left\lfloor \frac{(2k-1)N - N + 2}{2k-1} \right\rfloor \\ &= \left\lfloor \frac{(2k-1)N - (2k-1)q - r}{2k-1} \right\rfloor \\ &= N - q + \left\lfloor \frac{-r}{2k-1} \right\rfloor \end{aligned}$$

$$\left\lfloor \frac{(2k-2)N+2}{2k-1} \right\rfloor = N - q - \left\lceil \frac{r}{2k-1} \right\rceil = l(\mathscr{T})$$

Thus, $\mathscr{L}(N, 2k) \leq l(\mathscr{T})$.

However, by the maximality of $\mathscr{L}(N, 2k)$ and $\ell(G)$, we have $\mathscr{L}(N, 2k) \geq \ell(G) \geq l(\mathscr{T})$.

Therefore, $\mathscr{L}(N, 2k) = \ell(G) = \left\lfloor \frac{(2k-2)N+2}{2k-1} \right\rfloor$. □

4 A Spanning Tree of G with Optimal $d(G)$ and $\ell(G)$ but with Lesser Height

Let $G = C(N, S)$ be a circulant graph of order N and degree $2k$ with the maximum leaf number property and consider the maximum leaf spanning tree of G constructed by Algorithm 1.

The nonzero internal vertices of the spanning $\mathscr{T}$ form a path which starts at the root. Vertex 1 is the only internal vertex that is a child vertex of the root. The rest of the nonzero internal vertices are between 1 and the next child vertex of the root (in terms of value) or the next higher jump size.

The height of a vertex in a tree is the distance of the vertex from the root.

Hence, the height of the root u in a rooted tree $\mathscr{T}$ is $h(u, \mathscr{T}) = 0$ and the height of a child vertex of u is 1.

The height of a tree $\mathscr{T}$ is defined to be

$$h(\mathscr{T}) \stackrel{\text{def}}{=} \max\{h(v, \mathscr{T}) | v \in V(\mathscr{T})\}$$

where $V(\mathscr{T})$ is the vertex set of $\mathscr{T}$.

Hence, the height of the spanning tree $\mathscr{T}$ of G_1 or of G_2 generated by S, p, L of Algorithm 1 is equal to the number of internal vertices of $\mathscr{T}$ including the root. The height of the spanning is equal to the height of a child vertex of the farthest internal vertex from the root in the path of internal vertices.

Thus,

$$h(\mathscr{T}) = m(\mathscr{T}) = \left\lceil \frac{N-2}{2k-1} \right\rceil.$$

This tree can be shortened by half of its height by modifying the internal vertices obtained in Algorithm 1.

Modification of the Internal Vertices

Let $S[1] = s_1, S[2] = s_2, \ldots, S[2k] = s_{2k}$ be the jump sizes of $G = C(N, S)$ such that

$$s_1 = 1 < s_2 < \cdots < s_{2k} = N - 1.$$

For each $j = 1, 2, \ldots, 2k-1$, let $s_j = S[j]$ and $t_j = S[j+1]$.

Then we define the following variables d_j, $c_{j,1}$, and $c_{j,2}$ with

$$\bullet\ d_j = t_j - s_j - 1, \qquad \bullet\ c_{j,1} = \left\lceil \frac{d_j}{2} \right\rceil, \qquad \bullet\ c_{j,2} = \left\lfloor \frac{d_j}{2} \right\rfloor.$$

Note that $d_j = c_{j,1} + c_{j,2}$.

Thus

1. for $i_1 = 1, 2, \ldots, c_{j,1}$,

let $p[s_j + i_1] = i_1$ and append $s_j + i_1$ to $L[i_1]$.
Hence, $1, 2, \ldots, c_{j,1}$ are nonzero internal vertices.

2. for $i_2 = 1, 2, \ldots, c_{j,2}$,
let $p[t_j - i_2] = N - i_2$ and append $t_j - i_2$ to $L[N - i_2]$.
Hence, $N - 1, N - 2, \ldots, N - c_{j,2}$ are nonzero internal vertices.

Thus, the new nonzero internal vertices are partitioned into two paths where each path starts from the root 0.

These two paths of internal vertices are:

$$\langle 0, 1, 2, \ldots, c_{\delta,1} \rangle \text{ and } \langle 0, N-1, N-2, \ldots, N - c_{\delta,2} \rangle$$

where $\delta = \max\{d_1, d_2, \ldots, d_{2k-2}\}$

Hence, we have the following theorem.

Theorem 5 *The spanning tree $\mathscr{T}^*$ of G_1 or G_2 produced by the modification of the internal vertices of Algorithm 1 as stated above has the following properties:*

1. $m(\mathscr{T}^*) = \left\lceil \frac{N-2}{2k-1} \right\rceil$,
2. $l(\mathscr{T}^*) = \left\lfloor \frac{(2k-2)N+2}{2k-1} \right\rfloor$, *and*
3. $h(\mathscr{T}^*) = \left\lceil \frac{\delta(G)}{2} \right\rceil + 1 \leq \frac{m(\mathscr{T})}{2} + 1$ *where $\mathscr{T}$ is the MLST generated in Algorithm 1.*

Proof Let v be a nonzero vertex in $\mathscr{T}^*$.

If $v \in S$, then $p[v] = 0$, i.e., the parent vertex of v is the root 0. Suppose $v \notin S$. Then either

- $v = s_j + i_1$ where $i_1 = 1, 2, \ldots, c_{j,1}$ or
- $v = t_j - i_2$ where $i_2 = 1, 2, \ldots, c_{j,2}$.

1. Suppose $v = s_j + i_1$ where $i_1 = 1, 2, \ldots, c_{j,1}$. Then v and i_1 are adjacent vertices in G and $p[v] = i_1$ in $\mathscr{T}^*$.
If $v = 1 + i_1$, where $s_1 = 1$ and $v \in (1,\ 1 + c_{1,1}]$, then $i_1 = v - 1$.
Consider the vertex $c_{\delta,1} = \left\lceil \frac{\delta(G)}{2} \right\rceil$.
Since $d_j \leq \delta(G)$, it follows that $c_{j,1} \leq c_{\delta,1}$.

   (a) Suppose $\delta(G) = q - 1$.
   Then $G = G_1$ where $S = \{\pm(iq + 1)\}$, for all $i = 0, 1, \ldots, k - 1$. Thus, $d_1 = \delta(G)$.
   Hence, $c_{1,1} = c_{\delta,1}$.
   (b) Suppose $\delta(G) = q$.
   Then $G = G_2$. This means that $d_1 = \delta(G)$ or $d_1 = \delta(G) - 1$.
   Thus, either $c_{1,1} = c_{\delta,1}$ or $c_{1,1} = c_{\delta,1} - 1$.

Thus, $c_{\delta,1} \in [1,\ 1 + c_{1,1}]$. Consequently, $p[c_{\delta,1}] = c_{\delta,1} - 1$.
Hence, $p[v] = i_1$ in $\mathscr{T}^*$ where $v = s_j + i_1$ and $i_1 \in [1,\ c_{\delta,1}] \subseteq [1,\ 1 + c_{1,1}]$.
Note that vertex $c_{\delta,1} + 1$ has no child vertex since it is larger than $c_{\delta,1}$.
Hence, $1, 2, \ldots, c_{\delta,1}$ are the nonzero internal vertices in the closed interval $[1,\ 1 + c_{1,1}]$ of $\mathscr{T}^*$.
These internal vertices form a path between $c_{\delta,1}$ and 0 that passes through 1. This path is unique since a child vertex has a single parent vertex only.

2. Suppose $v = t_j - i_2$ where $i_2 = 1, 2, \ldots, c_{j,2}$. Thus, v and $N - i_2$ are adjacent vertices in G and $p[v] = N - i_2$ in $\mathscr{T}^*$.
If $v = t_{2k} - i_2 = N - 1 - i_2$ where $1 \le i_2 \le c_{2k-1,2}$, then $v + 1 = N - i$. Hence, $p[v] = v + 1$, $v \in [N - 1 - c_{2k-1,2},\ N - 1)$.
Consider the vertex $N - c_{\delta,2}$.
Since $\left\lfloor \frac{d_j}{2} \right\rfloor \le \left\lfloor \frac{\delta(G)}{2} \right\rfloor$, it follows that $c_{j,2} \le c_{\delta,2}$.
Thus, $N - c_{\delta,2} \le N - c_{j,2}$.

   (a) If $\delta(G) = q - 1$, then $d_{2k-1} = \delta(G)$. Thus, $c_{2k-q,2} = c_{\delta,2}$. Hence, $N - c_{\delta,2} = N - c_{2k-1,2}$.
   (b) If $\delta(G) = q$, then $d_{2k-1} = \delta(G) - 1$ or $d_{2k-1} = \delta(G)$. Thus, $c_{2k-1,2} = c_{\delta,2}$ or $c_{2k-1,2} = c_{\delta,2} - 1$.
   Hence, $N - c_{\delta,2} = N - c_{2k-1,2}$ or $N - c_{\delta,2} = N - 1 - c_{2k-1,2}$.
   This implies that $N - c_{\delta,2} \in [N - 1 - c_{2k-1,2},\ N - 1]$. Thus, $p[N - c_{\delta,2}] = N - c_{\delta,2} + 1$.
   Hence, $p[v] = N - i_2$ where $v = t_j - i_2$, $i_2 = 1, 2, \ldots, c_{j,2}$ such that $N - i_2 \in [N - 1 - c_{2k-1,2},\ N - 1]$.
   Note that vertex $N - c_{\delta,2} - 1$ has no child vertex in $\mathscr{T}^*$ since $c_{\delta,2} + 1$ is larger than $c_{\delta,2}$.
   Hence, the nonzero internal vertices of $\mathscr{T}^*$ in the closed interval $[N - 1 - c_{2k-1,2},\ N - 1]$ are $N - 1, N - 2, \ldots, N - c_{\delta,2}$.
   This forms a path that passes through $N - 1$ between the root and the internal vertices in this interval.

Thus, the number of nonzero internal vertices are $c_{\delta,1} + c_{\delta,2} = \delta$.
Hence, $m(\mathscr{T}^*) = \delta + 1 = \left\lceil \frac{n-2}{2k-1} \right\rceil$.
Consequently, $l(\mathscr{T}^*) = \left\lfloor \frac{(2k-2)n+2}{2k-1} \right\rfloor$.

Since $c_{\delta,1} \ge c_{\delta,2}$ and since the path $\langle 0, 1, \ldots, c_{\delta,1} \rangle$ is of length $c_{\delta,1}$, a child vertex of $c_{\delta,1}$ is of height $c_{\delta,1} + 1$.
Hence, $c_{\delta,1} + 1$ is the maximum of all heights in this new tree $\mathscr{T}^*$. Therefore, The height of the new tree $\mathscr{T}^*$ is

$$h(\mathscr{T}^*) = c_{\delta,1} + 1 \leq \left\lceil \frac{\delta(G)}{2} \right\rceil + 1 \leq \frac{m(\mathscr{T}^*)}{2} + 1 = \frac{m(\mathscr{T})}{2} + 1$$

Therefore, $h(\mathscr{T}^*) \leq \frac{h(\mathscr{T})}{2} + 1$. □

5 A One-to-Many Communication Scheme for the Circulant Graph with the Maximum Leaf Number Property

The topology of a communication network is a connected graph. A vertex of the graph is considered as the a processor of the network. An edge of the graph is a communication link between two processors in the network. A network port is a device in each processor where information or data packets can be sent and received from. A processor may use a single or multiple network ports to send or receive data. A network port that can send and receive data simultaneously is synchronous. Otherwise, it is asynchronous.

5.1 The Communication Network Model

Let us consider the undirected and connected circulant graph with the maximum leaf number property as the underlying topology of our communication network.

We shall assume a uniform number of ports per processor and that each network port can handle a synchronous exchange of data packets.

A *one-to-many communication* in the network is the sending and receiving of a data packet of messages or a data packet of personalized messages from one vertex to t other vertices in the network where $t = 1, 2, \ldots, N-1$.

A vertex which is not a receiver may help in forwarding a data packet of messages in the network.

A data packet sent by vertex u is a nonempty list of messages

$$\Pi[u] = [[\rho[v_1], \mu[v_1], [\rho[v_2], \mu[v_2], \ldots, [\rho[v_t], \mu[v_t]]$$

where a message to vertex v is pair $[\rho[v], \mu[v]]$

1. $-c_{\delta,2} \leq \rho[v] \leq c_{\delta,1}$ such that $c_{\delta,1} = \left\lceil \frac{\delta(G)}{2} \right\rceil$, $c_{\delta,2} = \delta(G) - c_{\delta,1}$, and
2. $\mu[v_i]$ is the body of the message to vertex v_i.

The message that reached its intended receiver v is valid if it is of the form $[0, \mu[v]]$.

ALGORITHM 2 One-to-Many Communication Scheme

Note that addition and subtraction of vertex labels are done under modulo N.

Require: A circulant graph G with the maximal leaf number property with
vertex labels : $0,1,\ldots,N-1$,
jump sizes : $s_1,s_2,\ldots,s_{2k}$,
where $1=s_1<s_2<\ldots<s_{2k}=N-1$,
number of nonzero internal vertices : δ,
source : u,
receivers : $v_1,v_2,\ldots,v_t$

Ensure: $[0,\mu[v_i]]$ is received by v_i for each $i=1,2,\ldots,t$.

1: **if** $\delta=0$ **then**
2: u sends the packets $\Pi[v_i]=[0,\mu[v_i]]$ to its child vertices v_i, for each $i=1,2,\ldots,t$.
3: **end if**
4: **if** $\delta>0$ **then**
5: find $c_{\delta,1}=\left\lceil\frac{\delta}{2}\right\rceil$ and $c_{\delta,2}=\delta-c_{\delta,1}$.
6: **end if**
7: For each receiver v, the message $\Pi[v]=[\rho[v],\mu[v]]$ is prepared such that
8: **if** v is a child vertex of u **then**
9: $\rho[v]=0$
10: **else** $\{v$ is not a child vertex of $u\}$
11: find two successive child vertices $u+s$ and $u+t$ such that $u+s<v<u+t$ where s and t are jump sizes of G
12: **if** $v\le u+\left\lceil\frac{t-s-1}{2}\right\rceil$ **then**
13: $\rho[v]=v-u-s$ where $1\le v-u-s\le c_{\delta,1}$, **and** append $[\rho[v],\mu[v]]$ to the data packet $\Pi[u+1]$
14: **else** $\{v>u+\left\lceil\frac{t-s-1}{2}\right\rceil\}$
15: $\rho[v]=v-u-t$, where $-c_{\delta,2}\le v-u-t<0$, **and** append $[\rho[v],\mu[v]]$ to the data packet $\Pi[u-1]$.
16: **end if**
17: **end if**
18: The transmission of the data packets from u to its child vertices of u are given as follows and this is order of the transmission in a 1-port network, otherwise, if we have a $2k$-port network this can be done simultaneously:
19: $\Pi[u+1]$, if it exists
20: $\Pi[u-1]$, if it exists
21: $\Pi[u+s]=[0,\mu[u+s]]$, if it exists, and s is a jump size of G with $s\neq 1$ and $s\neq N-1$
22: Suppose that a vertex z receives a data packet $\Pi[z]$.
(If we have a 1-port network, the following is is the order of the transmission of the messages in $\Pi[z]$.
Otherwise, if we have a $2k$-port, then Steps 23–24, Steps 25–26 and Steps 27–31 can be done simultaneously).
23: **if** $\Pi[z]$ has a message with $\rho[v]>1$ **then**
24: z creates data packet $\Pi[z+1]$ containing $[\rho[v]-1,\mu[v]]$ for all v with $\rho[v]>1$, **and** z sends it to $z+1$
25: **else** $\{\Pi[z]$ has a message with $\rho[v]<-1\}$
26: z creates data packet $\Pi[z-1]$ containing $[\rho[v]+1,\mu[v]]$ for all v with $\rho[v]<-1$, **and** z sends it to $z-1$.
27: **else** $\{\Pi[z]$ has a message with $\rho[v]=1\}$
28: z sends the message $[0,\mu[v]]$ to each of the v with $\rho[v]=1$.
29: **if** $\Pi[z]$ has a message with $\rho[v]=0$ **then**
30: the message belongs to z and accepts it.
31: **end if**
32: **end if**
33: Steps 22 to 32 are repeated until all the messages are accepted.

Theorem 6 *Consider $\mathcal{T}^*$, the MLST of a circulant graph with the maximum leaf number property, which was constructed in Theorem 5 and assume a 1-port communication network.*

Let $v_1, v_2, \ldots, v_t$ be the receivers of the messages sent by u in Algorithm 2.

If the network communication setup is a 1-port, then number of communication steps to transmit completely the messages to the receivers is equal to

$$\sigma = \max\{h(v_i, \mathscr{T}^*) + \kappa(v_i) + \varepsilon(v_i)\}$$

for each v_i where $h(v_i, \mathscr{T}^*)$ is the height of v_i in $\mathscr{T}^*$, $\kappa(v_i)$ is the number of siblings of v_i who are receivers also, and

$$\varepsilon(v_i) = \begin{cases} 1 & \text{if } [\rho[v_i], \mu(v_i)] \in \Pi[u-1] \text{ and } \Pi[u+1] \neq \emptyset \\ 0 & \text{if otherwise.} \end{cases}$$

If the network communication setup is a $2k$-port, then the number of communication steps to transmit completely the messages to the receivers is equal to $\sigma = \max\{h(v_i, \mathscr{T}^*)\}$.

Proof Suppose σ is the number of communication steps to complete the transmission of the messages from vertex u to $v_1, v_2, \ldots, v_t$.

Let us first consider the 1-port case.

1. Suppose $t = 1$.
   If v_1 is a child vertex of u, then $\sigma = 1$.
   If v_1 is not a child vertex of u, then there exists two successive jump sizes s and t such that $u + s < v < u + t$.
   If $u + s < v_1 \leq u + \left\lceil \frac{t-s-1}{2} \right\rceil$, then $[\rho[v_1], \mu[v_1] \in \Pi[u+1]$.
   Otherwise, $[\rho[v_1], \mu[v_1] \in \Pi[u-1]$.
   Then the $\rho[v_1], \mu[v_1]$ is transmitted along the nonzero internal vertices either through $u+1$ or through $u-1$.
   This is exactly the path for computing $h(v_1, \mathscr{T}^*)$.
   Hence, $\sigma = h(v_1, \mathscr{T}^*)$.
2. Suppose $1 < t \leq 2k$.
   The path in transmitting a message to a receiver v_i also follows that path for computing $h(v_i, \mathscr{T}^*)$.
   However, if the parent of v_i has other child vertices then the transmission steps is delayed by at most $\kappa(v_i)$, the number of siblings of v_i.
   If the message for v_i is transmitted from $u-1$, then there an additional delay of 1 step if $\Pi[u+1] \neq \emptyset$. Otherwise, there is no delay in the transmission.
   Hence, $\sigma = \max h(v_i, \mathscr{T}^* + \kappa(v_i) + \varepsilon(v_i))$.

If the communication network has $2k - port$ per vertex, then the number of siblings of each vertex and the transmission from $u-1$ when $\Pi[u+1] \neq \emptyset$ do not affect the communication of the message to its receiver.

Hence, $\sigma = \max h(v_i, \mathscr{T}^*)$. □

Acknowledgments This research is supported by Ateneo de Manila University and the Commission on Higher Education (CHED) of the Republic of the Philippines.

References

1. Douglas RJ (1992) NP-completeness and degree restricted spanning trees. Discrete Mathematics 105(1–3):41–47
2. Muga II FP (2014) On the maximal leaf number of a family of circulant graphs. 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, pp 46–49
3. Sampathkumar E, Walikar HB (1979) The connected domination number of a graph. J Math Phys Sci 13:607–613

A Model for Quality Assurance in Higher Education: A Case Study with Nigeria Higher Education

Moses Emadomi Igbape and Omame Philipa Idogho

Abstract Nigeria being the most populous nation in Africa with over 55 % of its population as children and youths of school age has not witnessed a corresponding development in infrastructure, educational facilities and resources to cater for the growing population of students. In quest for better standards, especially since the mid-1980s, there has been high mobility of students, academic staff and professionals. Also, new methods of delivery educational services and cross-border providers have emerged over the years, coupled with globalization issues in education, Quality Assurance (QA) in education has drawn concerns from governments, industry and other stakeholders in international education. A number of QA procedures have been presented by regulatory bodies of various nationalities. As a result of the observed disparity between the high population of school age youth in Nigeria and the inadequate educational resources, coupled with the low level integration of Information Technology (IT) tools in administration of education in the country, QA in education has not been fully achieved. It is of the view that except conscious and serious effort is made to develop environment specific technologies to deploy IT tools in educational administration, manual procedures for pursuing QA programmes may not yield desirable results. In this study, a model for building a system that can be used to evaluate the performance of academic programmes to determine level of compliance to Minimum Acceptable Standards (MAS) set by the appropriate regulatory bodies is provided with the architectural framework for building the performance evaluation system.

Keywords Accreditation · Data integration · Decision support · Higher education · Information technology · International education · Minimum acceptable standards · Performance evaluation · Quality assurance

M.E. Igbape (✉)
Office of the Rector, Auchi Polytechnic, P.M.B 13, Auchi, Edo State, Nigeria
e-mail: emadomi2@yahoo.com; emigbape@auchipoly.edu.ng

O.P. Idogho
Rector, Auchi Polytechnic, P.M.B 13, Auchi, Edo State, Nigeria
e-mail: philipaidogho@yahoo.com; rector@auchipoly.edu.ng

H.K. Kim et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-94-017-7236-5_24

1 Introduction

In the recent past, Information Technology (IT) and computer technologies have advanced tremendously giving rise to networks and internet technologies. With developments in Graphical User Interface (GUI), hand held devices and internet appliances technologies, computer applications have become extensively integrated into virtually all facets of business and human endeavours. Thus, computers now have great impact on the way we buy at supermarkets, the way postal services and health care services are rendered, travels, banking operations, telecommunications, educational administration activities etc. all depend on appropriate use of IT tools.

Nigeria has an estimated population of 167 million persons and is regarded as the sixth most populous nation in the world after China, India, USA, Indonesia and Brazil [1].

According to the United States Census Bureau, the population of Nigeria will reach 402 million by 2050. Nigeria will then be the 4th most populous country in the world. The age structure by year 2011 estimates is as follows; 0–14 years: 40.9 % (male 32,476,681/female 31,064,539), 15–64 years: 55.9 % (male 44,296,228/female 42,534,542) 65 years and over: 3.1 % (male 2,341,228/female 2,502,355) (2011 est.) An estimated figure of over 50 % of this population is made up of children and youth of school age [2].

Planning and developing adequate educational facilities and resources to cater for this ever increasing population pose serious challenges to education providers. The consequence of this population explosion on quality education is not far-fetched. The facilities and other resources are grossly inadequate. This greatly impacts on Quality Assurance (QA) in education at all levels.

In quest for better standards, especially since the 1980s, there has been high mobility of students, academic staff and professionals. In parallel, new delivery modes and cross-border providers have appeared, such as campuses abroad, electronic delivery of higher education and for-profit providers. These new forms of cross-border higher education offer increased opportunities for improving the skills and competencies of individual students and the quality of national higher education systems [3]. Coupled with globalization issues in education, QA in education has drawn concerns from governments, industry and other stakeholders in international education. Every institution strives to have a fair share of the global market and gain competitive advantage. This calls for serious consideration for QA issues in education.

Regulatory bodies of various nationalities have outlined procedures for QA. However, QA has not been fully achieved in the Nigerian education system; (i) there is low level integration of IT tools in educational administration and service delivery; (ii) the resources available for delivering educational services are inadequate due to the explosive population of school age persons in Nigeria [4].

It is of the view that except conscious effort is made to develop environment specific technologies to deploy IT tools in school administration, manual procedures for pursuing quality assurance programmes in Nigeria may not yield desirable results.

Consequently, the goal of this work is to present a model and architectural framework for building an Academic Information Management System that can be used to evaluate performance of academic programmes in order to determine compliance with Minimum Academic Standards (MAS) set by the appropriate regulatory agency.

To achieve the above objectives, a review of available literature on QA procedures was carried out. Thereafter, the accreditation guidelines provided by National Board for Technical Education (NBTE) and the Nigerian Universities Commission (NUC) have been examined. The model and architectural framework for developing a performance evaluation system provided here are based on the procedures specified by NBTE and NUC for evaluating the performance of academic programmes in Nigeria tertiary institutions.

It is hoped that developers will find this framework suitable for developing enterprise solutions for QA in Nigeria schools.

2 Higher Education in Society and Quality Assurance

2.1 Higher Education (HE)

In a society full of diversity, ideologies and opinions, HE means different things to different people [5]. There is more to Higher Education than the higher level of educational structure in a country. In terms of the level, HE includes college and university teaching-learning towards which students' progress is assessed to attain higher educational qualification. HE imparts indebt knowledge and understanding so as to advance the students to new frontiers of knowledge in different walks of life (subject domain). It is about knowing more and more about less and less. It develops the students' ability to question and seek truth and makes him/her competent to critique on contemporary issues. It broadens the intellectual powers of the individual within a narrow specialization but also gives him or her, wider perspective of the world around [5].

There are four predominant concepts of higher education [6];

(i) Higher Education as the production of qualified human resources. HE is viewed here as a process in which the students are counted as "products" absorbed in the labour market. Thus HE becomes input to the growth and development of business and industry;
(ii) Higher Education as training for a research career. HE is preparation for qualified scientists and researchers who would continuously develop the frontiers of knowledge. Quality in this sense is more about research publications and transmission of the academic rigor to do quality research;
(iii) Higher Education as the efficient management of teaching provision. Many believe that teaching is the core of educational institutions. Thus, HE institutions focus on efficient management of teaching-learning provisions by

improving the quality of teaching enabling a higher completion rate among the students; and

(iv) Higher Education as a matter of extending life chances. Higher Education is seen as an opportunity to participate in the development process of the individual through a flexible continuing education mode.

It is noted that all these four concepts of HE are not mutually exclusive, rather they are integrated and give an overall picture of what is higher in education.

2.2 *Role of Higher Education in Society*

Higher Education is generally understood to cover teaching, research and extension. It is the source of feeder system in all walks of life and therefore supplies the much needed human resources in management, planning, design, teaching and research. Scientific and technological advancement and economic growth of a country are as dependent on the HE system as they are on the working class. Development of indigenous technology and capabilities in agriculture, food security and other industrial areas are possible when there is quality HE system and infrastructure. Higher education also provides opportunity for life-long learning, allowing people to upgrade their knowledge and skills from time to time based on the societal needs.

The Mishra (2006) as quoted in [5], listed the roles of higher education institutions in modern society as follows;

(i) To seek and cultivate new knowledge, to engage vigorously and fearlessly in the pursuit of truth, and to interpret old knowledge and beliefs in the light of new needs and discoveries;

(ii) To provide the right kind of leadership in all walks of life, to identify gifted youth and help them develop their potential to the full by cultivating physical fitness, developing the powers of the mind and cultivating right interests, attitudes and moral and intellectual values;

(iii) To provide the society with competent men and women trained in agriculture, arts, medicine, science and technology and various other professions, who will also be cultivated individuals imbibed with a sense of social purpose;

(iv) To strive to promote quality and social justice, and to reduce social and cultural differences through diffusion of education; and

(v) To foster in the teacher and students, and through them in the society generally, the attitudes and values for developing the "good life" in individuals and society.

The Delors Commission on Education for the 21st Century [3] highlighted four pillars of education as learning to know, learning to do, learning to live together and learning to be. HE in addition to inculcating these core values in the individual and the society should strive to achieve the following functions;

(i) To prepare students for research and teaching;
(ii) To provide highly specialized training courses adapted to the needs of economic and social life;
(iii) To be open to all, so as to cater to the many aspects of lifelong education in the widest sense; and
(iv) To promote international cooperation through internationalization of research, technology networking, and free movement of persons and scientific ideas.

2.3 Quality Assurance in Higher Education

Quality Assurance (QA) in education and accreditation is a means of evaluating standards and quality in Higher Education (HE).

Quality Assurance is the systematic review of educational programmes to ensure that acceptable standards of education, scholarship and infrastructure are being maintained [3].

Quality means different things to different people. Quality is defined as the totality of features and characteristics of a product or service that bears on its ability to satisfy stated or implied needs [5]. Quality can be viewed from five different approaches [7], which are;

(i) In terms of exceptional (exceeding high standards and passing a required standard);
(ii) In terms of consistency (exhibited through "zero defects" and "getting right the first time", making quality a culture);
(iii) As fitness for purpose (meaning the product or service meets the stated purpose, customer specifications and satisfaction);
(iv) As value for money (through efficiency and effectiveness); and
(v) As transformative (in terms of qualitative change).

A lot of people consider quality as a relative term that has many dimensions that form a fuzzy entity referred to as quality through social consensus rather than defining it. The various available definitions have therefore been classified into five main groups [5];

(i) Transcendent definitions. These definitions are subjective and personal. They go beyond measurement and logical description. They are related to concepts such as beauty and love;
(ii) Product-based definitions. Quality is seen as a measurable variable. The basis for measurement is objective attributes of the product;
(iii) User-based definitions. Quality is a means for customer satisfaction. This makes these definitions individual and partly subjective;
(iv) Manufacturing-based definitions. Quality is seen as conformance to requirements and specifications; and
(v) Value-based definitions. These definitions define quality in relation to cost. Quality is seen as providing good value for costs.

When quality is considered as absolute, it is given and considered as the highest possible standard [5]. Examples are the picture of "Monalisa" by Da Vinci, the Egyptian Pyramids and the Taj Mahal are works of high standards and quality. In product terms, they are attached with "high brand" values, status and positional advantages. Educational institutions such as Oxford, Cambridge and Stanford in the west have the absolute quality standard, though in the case of education it might still be perceptual. Quality as relative suggests that the quality of a product or service can be described in relative terms. Quality here can be measured in terms of certain specifications. The adherence to "product specification" is actually the minimum conditions for quality, but not the sufficient condition. The sufficient condition is a customer satisfaction and beyond [5].

Quality as a process suggests that in order to achieve quality of a product or service, it must undergo certain processes and conform to the procedural requirements. Thus, quality is the outcome of systems and procedures laid down for the purpose. Quality as a culture recognizes the importance of organizational view of quality as a process of transformation where each entity is concerned and acknowledges the importance of quality. In educational institutions, particular concern is with the latter, though all other ideas of quality too have their respective places [5].

Barnet (1992) as quoted in [6], summarily provided a definition of quality as "a high evaluation accorded to an educative process, where it has been demonstrated that through the process, the students' educational development has been enhanced. Not only have they achieved the particular objectives set for the course but, in doing so, they have also fulfilled the general educational aims of autonomy of ability to participate in reasoned discourse of critical self-evaluation, and of coming to a proper awareness of the ultimate contingency of all thought and action".

From the above, it can be adduced that the meaning of quality is contextual, ranging from "standard" to "excellence". These two concepts are deeply rooted in their respective values and operationalized in individuals, institutions as well as national practice. Standards can be defined in terms of "minimum threshold" by which performance is judged. In Nigeria for instance, Minimum Academic Standards (MAS) is being referred to, which are set by the supervisory agency for the various categories of higher institutions.

In this context, quality is assessment in terms of a set of norm-referenced standards such as NAAC and Commonwealth of Learning Criteria [5], European Association for Quality Assurance in Higher Education [8], UNESCO Guidelines for Quality Provision in Cross-border Higher Education [3], Nigerian Universities Commission (NUC) Minimum Academic Standards (MAS) and NUC Benchmarks MAS [9]. These are built around what is considered as minimum standards. In higher education, the objective is to achieve the standards and move towards excellence.

2.4 Need for Quality Assurance

The UNESCO guidelines for quality provision in Cross-border Higher Education [3] states the purposes of the guidelines for quality provision in HE as protection of students and other stakeholders from low quality provision and disreputable providers as well as to encourage the development of quality cross-border higher education that meets human, social, economic and cultural needs.

In recent times, especially since the 1980s, there has been high mobility of students, academic staff and professionals. In parallel, new delivery modes and cross-border providers have appeared, such as campuses abroad, electronic delivery of higher education and for-profit providers. These new forms of cross-border higher education offer increased opportunities for improving the skills and competencies of individual students and the quality of national higher education systems [3].

To have a share of the global market and gain competitive advantage is a desire of every institution. Therefore, all stakeholders should be conscious of quality of teaching programmes in institutions. [5] set out some reasons for QA in higher education.

(i) Competitions; (ii) Customer satisfaction; (iii) Maintaining standards; (iv) Accountability; (v) Improve employee morale and motivation; (vi) Credibility, prestige and status; and (vii) Image and visibility.

In [3] on quality assurance in cross border education, it is noted that in some countries, the national frameworks for quality assurance, accreditation and recognition of qualifications take into account cross-border HE, in many countries they are still not geared to addressing the challenges of cross-border provisions. Furthermore, the lack of comprehensive frameworks for coordinating various initiatives at international level, together with diversities and unevenness of the quality assurance and accreditation systems at the national level, create gaps in the QA of cross-border HE leaving some cross-border HE and provision outside any framework of QA and accreditation.

The quality of a country's HE sector and its assessment and monitoring is not only key to the social and economic well-being; it is also a determining factor affecting the status of that HE system at the international level. The establishment of quality assurance system has become necessary, not only for monitoring quality in HE delivered within the country, but also for engaging in delivery of HE internationally.

2.5 Quality Assurance Methodologies in Higher Education

Quality Assurance is the responsibility of everyone in the educational institution, though the top management sets the policies and priorities. QA should be a continuous and on-going process. It should not be considered as one time activity for

accreditation alone. However, External Quality Monitoring (EQM) can be found in all types of higher education systems [10].

It is highly recommended that every high institution develops internal QA mechanism. This unit within the institution will prepare the institution for External Quality Monitoring (EQM) which is accreditation exercise. Therefore, understanding the criteria for QA and adhering to best practices will significantly facilitate success in this direction. Across the world, QA is carried out in the following ways [5];

(i) Self-evaluation;
(ii) Peer review by panel of experts, usually including at least some external panel members and one or more site visits;
(iii) Analysis of statistical information and/or use of performance indicators or the best practices benchmarking;
(iv) Surveys of students, graduates, employers, professional bodies; and
(v) Testing the knowledge, skills and competencies of students.

At NAAC, a four stage process of EQM/assessment is undertaken covering;

(i) Identifying pre-determined criteria for assessment;
(ii) Preparation and submission of the self-study report by the unit of assessment;
(iii) On-site visit of the peer team for validation of the report and recommendation of the assessment outcome to NAAC; and
(iv) Final decision by the executive committee of NAAC.

Real quality that is sustainable is one that is assessed by self [5]. This is how to know what our strengths and limitations are. Self-evaluation is like looking ourselves in a mirror. It is therefore advisable that institutions submitting a self-study report need be self-critical and reflective. Otherwise the external input into quality control will not yield any positive result.

Many of the well-known approaches besides self-evaluation and self-study include;

2.5.1 Best Practices Benchmarking

Benchmarking entails the process of recognizing 'best practices' in the industry and implement them. It is defined as "a continuous systematic process for evaluating the products, services and work processes of organizations that are recognized as representing the best practices for the purpose of organizational improvement" (Mishra 2006) as quoted in Mishra [5]. As a process, this has four main activities; (i) Comparing one thing with the other; (ii) Creating and using criteria to evaluate differences between two things and recognizing which is better;

(i) Use the experience to identify the direction for change; and
(ii) Implement the required change to improve

Although benchmarking is a relatively new concept in education, its use is expected to bring about huge benefits in terms of continuous improvement of quality. As the best is compared and follow the best institutions, it becomes a tool for motivation to change.

2.5.2 External Quality Monitoring

EQM has become mandatory in many countries. It reassures external stakeholders such as employers, professional bodies and the general public about the legitimate quality of higher education institutions. It also offers an impartial and objective mechanism for assessing the educational institution by a peer team not directly related to the institution. Visit by a peer team is a common activity in EQM, which critically analyses the self-study report and the quality provisions based on established criteria.

Other approaches include the Unit of Assessment and Market-driven approaches. The issue of quality is so important these days that ranking of educational institutions has become a dominant factor in business. In order to capitalize on the internal quality and to add value to the quality assessment, EQM is highly recommended and is preferred above other approaches. Thus many countries use EQM as a strategy to assess the quality of educational institutions.

2.6 Quality Assurance Models

A detailed survey, analysis and critic of the existing QA models are outside the scope of this work. For the purpose of proper linkage to the model adopted by NUC in Nigeria, the available and commonly used model is mentioned. Across the world, institutions follow different models of QA; particularly country specific and institution specific models. Some of the most commonly used approaches [5], include; (a) Baldrige Criteria, (b) ISO 900:2000, (c) Capability Maturity Model, (d) Six Sigma, (d) Total Quality Management, (e) Towards Total Quality Care

A most commonly used model however is;

Accreditation Board for Engineering and Technology (ABET) Model. Established in 1932 as Engineer's Council for Professional development (ECPD) follows the accreditation tradition in the USA which requires a voluntary participation by institutions. An internal self-study evaluation forms the basis of the beginning of the accreditation process. Based on the self-study report, the appropriate ABET Commission forms an evaluation team for the site visit. Based on the visit, the peer-team provides the institution with a written report to allow for correction of errors or misrepresentation of facts. The peer team examines the following in comprehensive manner and recommends accreditation and relevant action;

a. Organization and management of the institution;
b. Education programmes offered;
c. Maturity and stability of the institution;
d. Admission process and number students enrolled;
e. Teaching staff and teaching load;
f. Physical facilities, finances etc.;
g. Curricula contents;
h. Sample student work;
i. Record of employment of graduates;
j. Support services to the students; and
k. Clearly stated academic policies.

Accreditation is usually granted for a period ranging from 2–6 years. Depending on the strength or weakness of the programme, the peer team recommends specific action to be taken by the Commission such as; Next General Review 6 years; Interim Report and Interim Visit both 2 years; Report Extended and Visit Extended 2 or 4 years; Show Cause 2 yrs; Show Cause Extended 2 or 4 years and Not Accredited.

Other known models are NBA Model, NAAC Model, ICAR Model and the DEC Model [5].

3 Modeling Approach and Methodology

The use of manual procedures for pursuing quality assurance programmes in Nigeria higher institutions have not yielded desirable results. Consequently, in this work the model and architectural framework presented in [11] for building a decision support system for an Academic Information Management System that can be used to evaluate performance of academic programmes is adopted. This is considered adequate for realizing desirable QA results.

Following design science approach [12, 13] and systems engineering hierarchy approach [14], we relied on the knowledge of the existing practice in Nigeria education sector, derived from work experience, interest and observations, to formulate innovative concepts and research idea that showed potential to improve actual human and organizational capabilities in evaluating performance of academic programmes in Nigerian higher schools.

In an academic institution like the university, the information required for accreditation come from six main performance indicators;

(i) Academic matters; (ii) Staff Quality; (iii) Physical facilities; (iv) Funding; (v) Library; (vi) Graduate rating by employers

The desired information come from operational source systems like

(i) Establishment; (ii) Bursary; (iii) Academic departments; (iv) Library; (v) Estate and Works department.

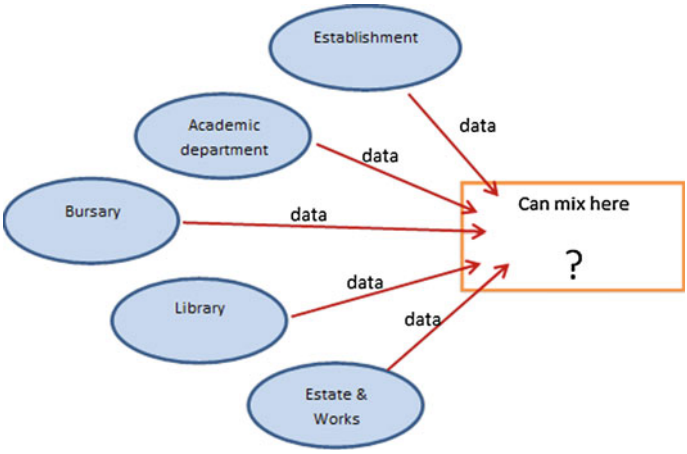


Fig. 1 Data integration concept for QA information system [4]

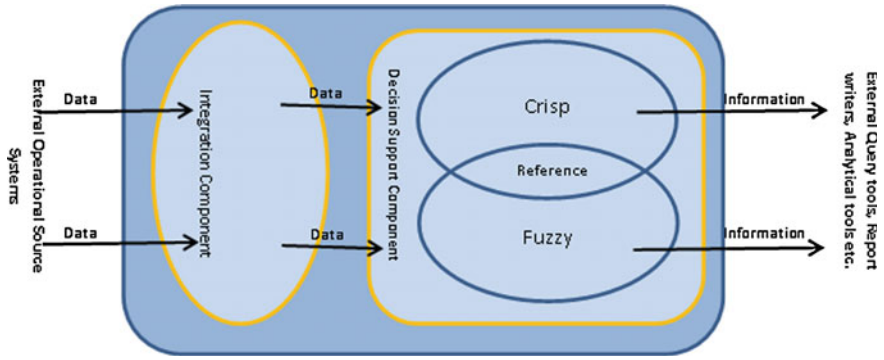


Fig. 2 Model of integrated academic information management system (IAIMS) [4]

Each of the operational source systems is a system in its own right, having the full components of a system. A data integration concept considered suitable for the QA information system is illustrated in Fig. 1. It is of the view that an integrated academic information management system concept shown in Fig. 2, with two basic components is desirable.

The Integration Component Layer

This layer is introduced to hold copies of data extracted from operational source systems. These may be text files or data from relational tables. Here the data from the diverse sources are scrubbed, conformed and standardized into dimension tables. There are no user queries at this layer.

The Decision Support Component Layer
Two processes take place here; computation and aggregation of assessment scores for classification of performance and optionally, the determination of membership degrees of the attributes in the resulting dimensional tables using trapezoidal function and classification of the attributes using fuzzy classification algorithm;

The content of the evaluation forms used by NUC were examined from which the data analyzed in this work are generated.

3.1 Grading

At the end of the accreditation visit, the executive committee at NUC reviews the reports submitted by the accreditation panel to take a decision on the accreditation status of the programme under consideration. An evaluated programme may earn any of 3 (three) possible accreditation status;

(i) FULL ACCREDITATION: a minimum of 70 % aggregate score and 70 % in each of the four core areas of academic matters, staffing, physical facilities and library. This accreditation is valid for 6 (six) years with mid-term review;
(ii) INTERIM ACCREDITATION: aggregate score of not less than 60 % or programme with a total score above 70 % but which scores less than 70 % in any of the indicated 4 (four) core areas. This accreditation is valid for 2 (two) years; and
(iii) DENIED ACCREDITATION: failed to satisfy MAS with less than 60 % aggregate score. Admission of new entrants into this programme ceases until deficiencies are remedied.

For the purpose of this study, four (4) programmes are used as follows; Computer Science, Law, Accountancy, Biotechnology. For each of the six (6) performance indicators, sample assessment was carried out following the evaluation indices contained in the NUC evaluation form. The result for Computer Science over the four year period is shown in Table 1.

By introducing time and programme dimensions, a data cube can be realized. The computed percentage (%) column from Table 1 can be used for analysis and reporting.

3.2 Data Analysis and Visualization

Data analysis applications look for trends and unusual patterns in data. They categorize data values and trends, extract statistical information and then contrast one category with another [15]. This is achieved through four steps;

Table 1 Assessment scores for computer science

Performance indicator	2008 Performance rating (%)	2009 Performance rating (%)	2010 Performance rating (%)	2011 Performance rating (%)	Total (%)
Academic matters (23)	82.607	73.913	65.217	69.565	72.826
Staffing matters (32)	59.375	65.625	71.875	84.375	70.313
Physical facilities (25)	56	68	80	84	72
Funding (5)	20	40	70	60	47.5
Library (12)	58.333	66.667	75	75	68.75
Employer rating (3)	66.667	66.667	66.667	83.33	70.833
Overall performance					67.037

(i) Formulating queries that extract relevant data from large database;
(ii) Extracting the aggregated data from the database into a file or table;
(iii) Visualizing the results in graphical way; and
(iv) Analyzing the results and formulating new queries.

Visualization and data analysis tools do 'dimensionality reduction' often by summarizing data along the dimensions of interest. For example, the performance of an academic programme (that is, to check the extent of compliance of a programme to set standards in MAS) is analyzed. It focuses on the role of programme of study, performance indicators and year. The overall performance of a programme across the performance indicators by year may be analyzed.

Queries are formulated to provide answers to enquiries such as;

(i) What is the overall rating of Computer Science programme across all performance indicators for 2008. This can be extended for subsequent years and visualized on the x-y plane of Fig. 3;
(ii) What is the overall performance of Computer Science programme in Academic Matters over the years (2008–2011)? This can be extended for other performance indicators and visualized on the x-z plane of Fig. 3;
(iii) What is the overall performance of Computer Science across all performance indicators over the 4 (four) years under review. This can be visualized on the z-y plane of Fig. 3; and
(iv) Assuming all programmes belong to a particular faculty or college, what is the overall performance of all the programmes over the 4-years under review?

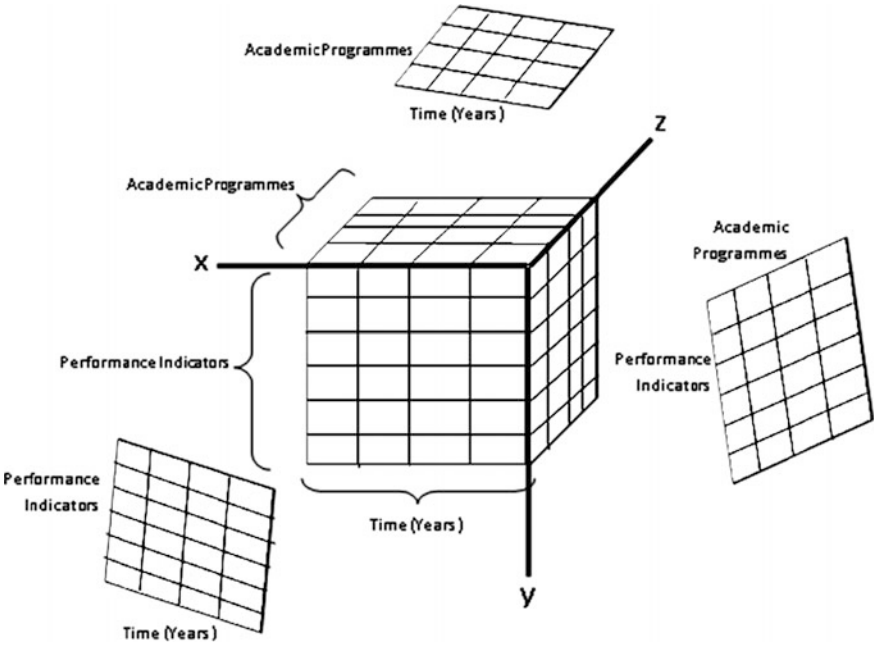


Fig. 3 XYZ planes visualization of QA data

3.3 Architectural Framework for Implementation

In this section, an architectural framework for the design of a decision support system for an integrated academic information management, which is considered suitable for QA is presented in Fig. 4. The prototype used to simulate examples in this work was achieved through interfaces created to enable access to the database system.

This requires knowledge of the structures of the database tables and meta-tables.

This architecture will work well for PostgreSQL 8 for Windows or MySQL database engine as well as Microsoft SQL Server database engine. It is to be noted that the database engine adopted will determine the specific syntax of the SQL commands. The three-tier architecture framework built with MySQL/PostgreSQL on three schema is adopted [16].

Developers can create frontend applications for the query processing layer using technologies provided by the.NET development environment. These include the visual studio.NET suite and the.NET framework Software.

Development Kit (SDK). PostgreSQL also provides development library for. NET programmers [17]. MySQL database engine and PHP scripts were also used to implement the prototype system.

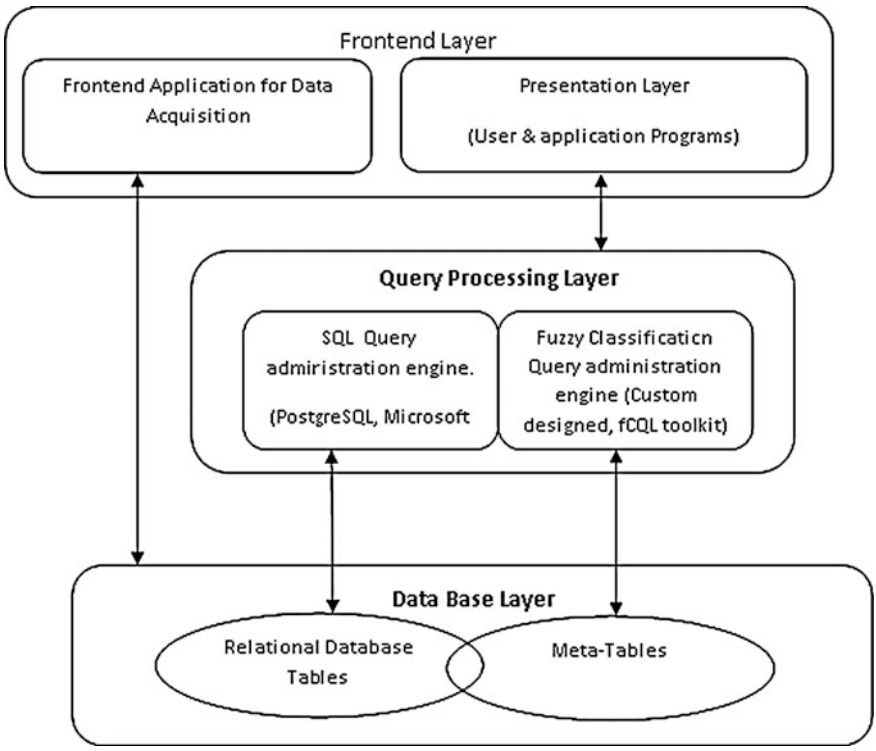


Fig. 4 Architectural framework for the system

3.4 Implementation of Model

The discussion of the features of standard SQL and some vendor-specific SQL-extensions is adequately covered in [15]. To achieve data analysis requirement of formulation, extraction, visualization and analysis, the concept of N-dimensional generalization of standard SQL aggregate functions and operators is adopted. In this case study, Academic Programmes are classified into segments such as Full Accreditation, Interim Accreditation and Denied Accreditation based on their performance in evaluation using indicators such as academic matters, staffing matters, physical facilities, funding, library and employer rating of graduates.

The Quality Assurance in higher Education with programmes accreditation case study example used in this work presents data with historical perspective, which creates a multi-dimensional or star schema data structure. The case study is therefore a valid example to demonstrate the viability of this decision support model of an Academic Information Management System.

Table 2 Result set for example 1 query

Programme	Year	Performance indicator	Performance rating (%)
Computer science	2008	ALL	62.02

Table 3 Computer Science rating

Programme	Year	Performance indicator	Performance rating (%)
Computer science	ALL	ALL	67.037

Example
The University administrators may want to know the overall rating of Computer Science programme across all performance indicators for 2008. This is a dice operation over a sub cube visualizing the overall assessment score grouped by programme but executed on the dimension of time (year = 2008). This visualization is a 1-dimension data cube. The result of the operation is shown in Table 2.

This query can be extended for all programmes under assessment. This is a 2-dimension data cube visualization as shown by the Y-Z plane of Fig. 3.

Example 2
The interest may be on the overall performance of Computer Science across all performance indicators over the 4 (four) years under review. This is a symmetric aggregation called cross-tabulation (cross-tab). It is a two-dimensional aggregation which is equivalent to the relational aggregation using the ALL values. Both generalize to N-dimensional cross-tab. This is a 2-dimension data cube visualization as shown by the X-Y plane of Fig. 3. The result set is shown in Table 3.

3.5 Conclusion

From the data analysis presented above, the department of Computer Science failed to achieve Full Accreditation status. With an overall score of 67.037, the accreditation status is INTERIM ACCREDITATION which is valid for 2 years.

The graphical model provides a framework for data integration which is considered adequate to resolve the problem of ad hoc incompatible information systems identified in higher institutions in Nigeria. Standards and best practices have been established in this respect [18].

To provide concrete instantiation through a real life example, University programmes accreditation scenario is used as a case study. The results of the analysis of data realized from the four (4) academic programmes and four (4) years used in this study show that this decision support system will be of tremendous value in administering scarce financial and other resources as well as effective budgeting.

A system built with this model can also be used to predict the performance behaviour of a programme in the future by varying the performance indices to

provide answers to what if questions. For example, information about staff retirement can be fed into the system so as to determine proactively, the behaviour of staffing matters and to visualize its impact on overall rating of the programme. The system will assist management to visualize performance of programmes at the end of each session to see where the weak indices lie so as to strengthen them and where the strong indices lie so as to sustain them. Management can also consider the effect of age of equipment or failure, budget cuts or low internally generated revenue on performance of programmes so as to plan proactively before the negative effects manifest.

It is therefore recommended that government and other agencies who are stakeholders in education should sponsor a development project that can use this model and the architectural framework provided to build the Integrated Academic Information Management System (IAIMS) that can be deployed for use in higher institutions for performance evaluation as a quality assurance means in Nigeria higher education.

3.6 Further Studies

Further work is being done on the implementation of this model as well as the introduction of fuzzy concepts to improve the decision support capability of the model.

Acknowledgment We acknowledge the numerous contributions of authors cited in this work. We thank Prof. E.A. Onibere of the department of Computer Science, University of Benin, Nigeria for his guidance in the process of this work.

References

1. NPC "Nigeria over 167 million population: implications and challenges" [Online]. Available: www.population.gov.ng
2. Otedo News "Demographics of Nigeria" [Online]. Available: www.ihuanedo.ning.com/profile/blogs/demorgraphics-of-nigeria
3. UNESCO "Guidelines for quality provision in cross-border higher education" [Online] Available: www.unesco.org/education/hed/guidelines
4. Igbape EM (2014) Fuzzy logic modeling of an integrated academic information management system. Ph.D. dissertation, Dept. Computer Science, University of Benin, Benin City
5. Mishra S (2006) Quality assurance in higher education: an introduction. National Assessment and Accreditation Council (NAAC), Bangalore
6. Barnett R (1992) Improving higher education: total quality care. SRHE and OU, Buckingham
7. Harvey L, Green D (1993) "Defining quality," assessment and evaluation in higher education. An Int J 18(1)
8. EAQAHA (2006) Quality assurance of higher education in Portugal: an assessment of existing system and recommendations for future system. European Association for Quality Assurance in Higher Education, Helsinki, [Online]. Available: www.enqa.eu/pubs.lasso

9. Suleiman RY (2008) Quality assurance and accreditation: the experience of the national universities commission of Nigeria. Department of Distance Education NUC, Abuja
10. Harvey L (1998) An assessment of past and current approaches to quality in higher education. Aust J Educ 42(3):237–255
11. Igbape EM, Idogho PO (2014) Performance evaluation model for quality assurance in Nigeria higher education. In: Proceedings of the world congress on engineering and computer science 2014, WCECS 2014. Lecture notes in engineering and computer science, 22–24 October, San Francisco, pp 334–343
12. Hevner AR, March ST, Park J, Ram S (2004) Design science in information systems research. MIS Quarterly 28(1)
13. March ST, Smith GF (1995) Design and natural science research on information technology. Decis Support Syst 15:251–266
14. Pressman RS (2005) Software engineering (A practitioner's approach), 5th edn. McGraw Hill Book Company, Boston
15. Gray J, Chaudhuri S, Bosworth A, Layman A, Reichrt D, Venkatrao M (1997) Data cube: a relational operator generalizing group by, cross tab, sub-totals. Data Min Knowl Discov 1:29–53 (Kluwer Academic Publishers)
16. Fasel D, Shahzad K (2010) A data warehouse model for integrating fuzzy concepts in meta-table structures. In: Proceedings of the 17th IEEE international conference and workshops on the engineering of computer based systems, pp 100–109
17. Blum R (2007) PostgreSQL 8 for Windows. McGraw Hill Book Company, New York
18. TrowBridge D, Roxburgh U, Hohpe G, Manolesscu D, Nadhan EG "Integration patterns," Available: http://www.msdn.microsoft.com/library/default.asp

Investigation of Radiation Dose and Image Quality in X-Ray Radiographic Imaging

Rafidah Zainon, Nor Syazreena Abu Talib, Siti Nur Amira Abu Bakar, Nur Hanisah Mohd Moner and Nurul Athirah Abdul Aziz

Abstract The main purpose of this study was to investigate the effect of tube potential on radiation dose and image quality; and to optimize the radiographic imaging by applying different type of parameters such as kilovoltage, types and thickness of filters. Studies were conducted to obtain the best quality radiograph of the RMI chest phantom using Toshiba Mobile Equipment. The first part of this study was to investigate the effect of kVp on radiation dose and image quality. The second part of this work was to optimize the image by evaluating parameters such as tube potential, filter types (Copper and Aluminum), and with and without filter thickness. The absorption dose increases as the kVp increases. The optimization of radiation dose and image quality can be obtained by using Copper filtration. Contrast to noise ratio is higher when using 0.2 mm copper thickness at 100 kVp, which denoted a better image quality. In conclusion, the optimization of x-ray radiographic imaging is needed to produce good quality image without exposing the patients with high radiation dose.

Keywords Filter · Image quality · Optimization · Radiation dose · Radiographic imaging · X-ray

R. Zainon (✉)
Advanced Medical and Dental Institute, Universiti Sains Malaysia, Bertam, 13200 Kepala Batas, Pulau Pinang, Malaysia
e-mail: rafidahzainon@amdi.usm.edu.my

N.S.A. Talib · S.N.A.A. Bakar · N.H.M. Moner · N.A.A. Aziz
School of Physics, Universiti Sains Malaysia, 11800 Kepala Batas, Pulau Pinang, Malaysia
e-mail: syazreenasyaz@yahoo.com

S.N.A.A. Bakar
e-mail: snuramira228@yahoo.com

N.H.M. Moner
e-mail: haniemoner@yahoo.com

N.A.A. Aziz
e-mail: nurul_athirah@ymail.com

H.K. Kim et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-94-017-7236-5_25

1 Introduction

Diagnostic imaging using film has been used since many years ago [1, 2]. During early days, the radiation dose given to the patient does not consider the side effect that will be suffered by the patient. After a number of examinations performed, and results on the risk of cancer increased due to ionizing radiation exposure, attention should be given to optimize the radiation dose given to the patient [2–4]. The quality of the image and the anatomic structure that appears within the image depends on the imaging parameters used and the amount of radiation exposed.

Basically, when the radiation dose is increased, the quality of the image is increased, but it also increased the radiation dose absorbs by the patient. Level of acceptance of image quality should be recognized before any optimization can be done [5–7]. This to ensure that any clinical diagnostic information needed is not missing out from the image produce. There are several ways to obtain optimal radiation dose and optimal image quality in x-ray radiographic imaging.

Optimization of radiation dose technique includes the use of optimum peak kilovoltage (kVp). When the kVp increases, the energy of x-ray produce will increase. The high energy of x-ray beams will penetrate denser material and produce better images. The high energy x-ray beam will produce more secondary x-rays [3, 4]. This secondary x-rays will be absorbed within tissue and increased the absorb dose of the patient. In order to produce optimal image quality and reduce the absorb dose, optimum kVp should be obtained.

When the x-ray beam penetrates into patient's body, the x-ray beam will undergo several reactions. Some of x-ray beam might penetrate through body and hit the film, some of x-ray beam will react with atom inside patient's body yield a Compton scattering reaction and photoelectric reaction and some of x-ray beam might simply being absorbed by the tissue inside the body [5–7].

Usually, soft x-ray is easily being absorbed by the tissue in the body and increase the radiation dose absorb by the patient. Soft x-ray does not contribute to production of image quality [7]. The presence of soft x-ray only increase the radiation dose absorb by the patient. Addition of filters will filter the beam coming out from the x-ray tube. All soft x-rays will be blocked by the filters and the x-ray beam with higher energy will pass through the patient. The amount of energy blocked by the filters depends on the thickness of the filters. The thicker the filter, the more energy is blocked by the filters and optimization of radiation dose can be achieved.

The main objective of optimization of radiation dose is to ensure that the diagnostic image produces achieve its acceptance level without giving any serious harm to the patient and also the radiographer. Therefore, this study focuses on the assessment and optimization of radiation dosimetry and image quality in radiographic imaging.

2 Materials and Methods

In this study, we imaged the RMI Radiographic Chest Phantom (model 170B) using Toshiba mobile x-ray machine. The image was captured on Radiographic film MG-SR Plus Konica Film. The radiation dose was measured using PTW UNIDOS ionization chamber. Two types of filters were used in this study: Copper and Aluminum filters with thickness of 0.2, 0.4 and 0.6 mm.

The assessment of image quality was performed at fixed tube current and exposure time of 1.0 mAs and varying the tube voltage. The RMI Radiographic Chest phantom was exposed with 56 up to 100 kVp with increment of 4 kVp. The film was processed. The images were analyzed using ImageJ software. The mean and standard deviation of background and object from the film were obtained by drawing region-of-interest on the image. The CNR are then calculated using Eq. (1).

$$\mathrm{CNR} = \left(\mu_{\mathrm{detail}} - \mu_{\mathrm{background}}\right)/\left(\sqrt{\left(\sigma^2_{\mathrm{detail}} + \sigma^2_{\mathrm{background}}\right)}\right) \quad (1)$$

μ_{detail} and $\mu_{\mathrm{background}}$ are mean of detail and background respectively. $\sigma^2_{\mathrm{detail}}$ and $\sigma^2_{\mathrm{background}}$ are the standard deviations for the detail and background respectively. A graph of CNR against tube voltage is plotted.

The measurement of radiation dose was done using the same parameters applied for assessment of image quality. In this study, the PTW UNIDOS ionization chamber was placed under the exposure window to get the measurement of absorbed dose. The measurements of radiation dose were recorded.

The effect of different types and thickness of filters on radiation dose and image quality was investigated. The tube current and exposure time was set at 1.0 mAs and the tube voltage was increased from 80 to 120 kVp. The thickness of Copper filter was added from 0.2 up to 0.6 mm. This procedure was repeated for Aluminum filter.

3 Results

The production of image of the body depends on several factors. It is important to understand how the image is produced, anatomical structure that we want to examined, the factors that affect the image quality and radiation dose received by the patient. The CNR are calculated to analyze the image quality. Four graphs of CNR against tube potential were plotted for different parts anatomical structure found in chest phantom: fat, vertebra, tumor and bone step wedge. Figure 1 shows the CNR of different diameters of fat versus tube potential. The highest CNR of fat can be seen in the range of 72–84 kVp for all diameters of fat.

Figure 2 shows the CNR of vertebra versus tube potential. Results show that the highest CNR of vertebra is achieved at 84 kVp. The CNR of vertebra decreases above 85 kVp.

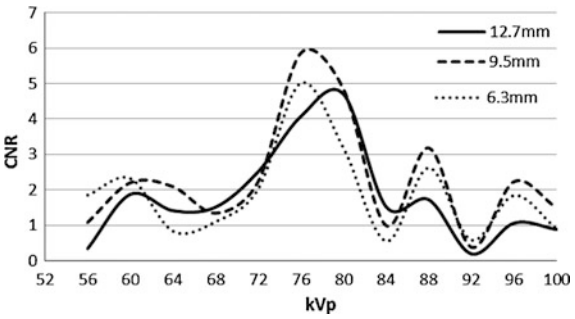


Fig. 1 CNR of fat with different diameters against tube potential

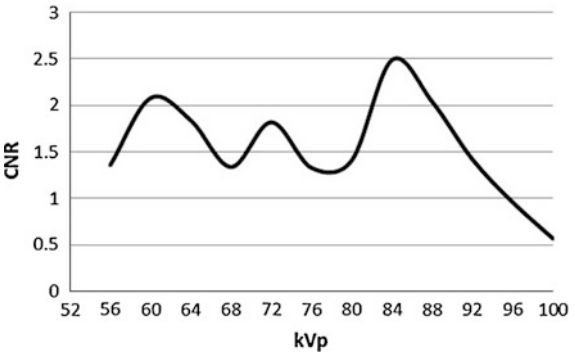


Fig. 2 CNR of fat with different diameters against tube potential

Figure 3 shows the CNR of all bone stepwedges versus tube potential. The CNR increases slowly as the tube potential increases.

Figure 4 shows the CNR of tumors of different diameters versus tube potential. The CNR of tumors is maximum at 76 kVp. The maximum CNR of all bone step wedge is achieved at 100 kVp.

Figure 5 shows the absorbed dose against tube potential. An increment of 25 % can be seen from 56 to 60 kVp. The increment of radiation dose is linear with the increment of tube potential.

Results of optimization of radiation dose and image quality are presented in Figs. 6 and 7. Figure 6 shows the radiation dose versus thickness of filters at different tube potential. The result shows that as the thickness of an additional filter increases, the dose gradually decreases from 80 to 120 kVp.

Figure 6 illustrates that the reduction of radiation dose of 72.77 % can be achieved with 0.2 mm thickness of Copper filter. With 0.4 mm thickness of Copper filter, the dose reduces about 33.34 %. For 0.6 mm thick, the dose reduces over 25 % compared to 0.4 mm thick.

For Aluminum filter, there is only small dose reduction compare to Copper filter. The reduction of dose only denoted 6.82 %, which is ten times lesser than Copper

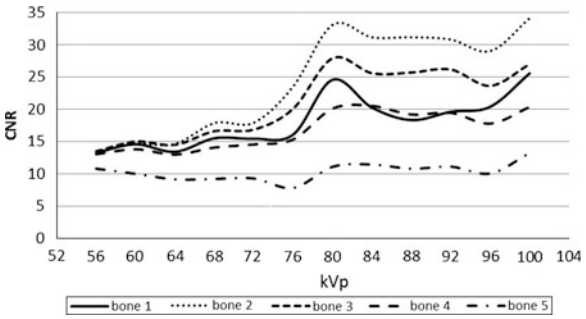


Fig. 3 The CNR of bone stepwedges against tube potential

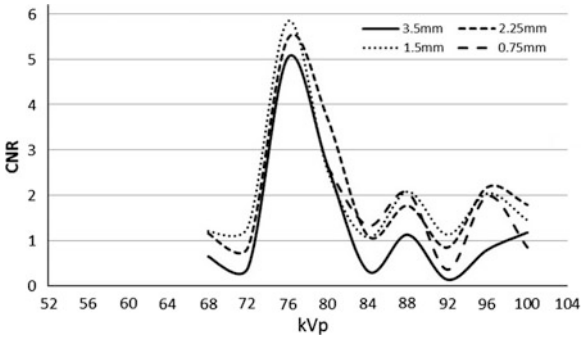


Fig. 4 The CNR of tumors of different diameters versus tube potential

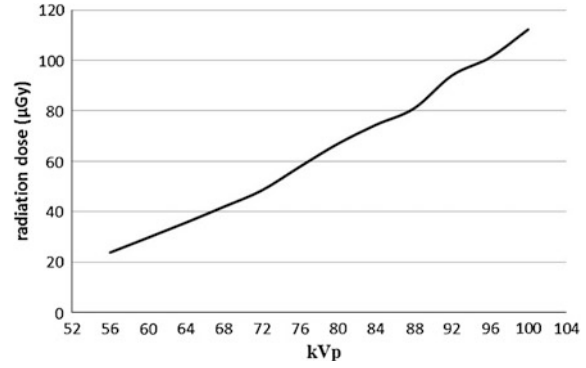


Fig. 5 The absorbed dose versus tube potential

filter. Copper filtration shows gradual decrease in dose while Aluminum filter only shows a moderate fall in dose reduction.

The CNR is measured to determine the image quality. Higher CNR shows that the signal of the image is better compared to background's signal, which will denote a better image quality and vice versa.

Figures 7 and 8 demonstrate the CNR of bone and fat versus tube potential for different thickness of Copper and Aluminum filters. Based on Figs. 7 and 8, both

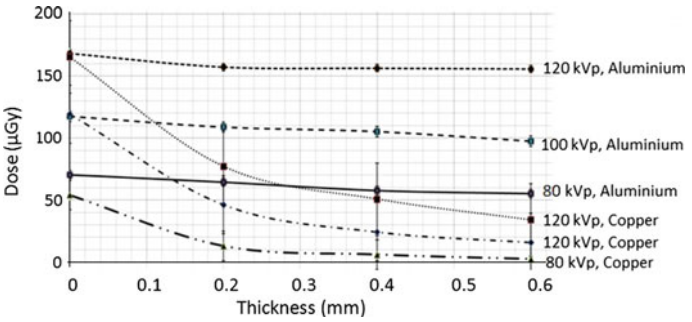


Fig. 6 The measured radiation dose versus thickness of Copper and Aluminum filters for 80, 100 and 120 kVp

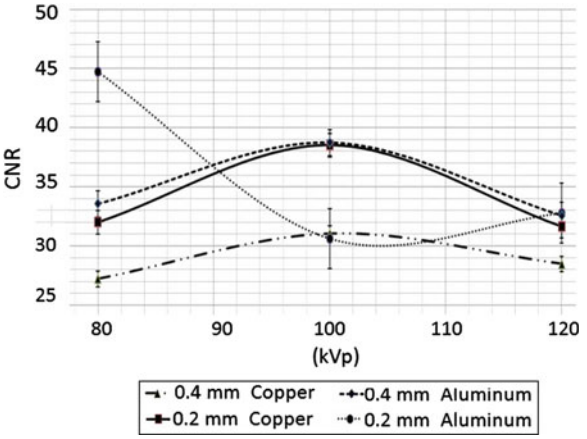


Fig. 7 The CNR bone versus tube potential with Copper and Aluminum filters at thickness of 0.2 and 0.4 mm

bone and fat have highest CNR with 0.2 mm of Copper filter at 100 kVp. On the other hand, the CNR of bone and fat is highest with 0.2 mm thickness of Aluminum filter at 80 and 100 kVp respectively.

4 Discussions

This study indicates that 80–120 kVp is the best range of tube potential for chest imaging. The highest CNR of fat, bone and tumor in the chest phantom can be observed in this range of tube potential. Increasing the tube potential will increase the radiation dose. The radiation dose gradually increases when the tube potential increases.

Thus, it is not necessary to use higher kVp to obtain the best quality image. However, the tube potential selected should be appropriate for CNR required and

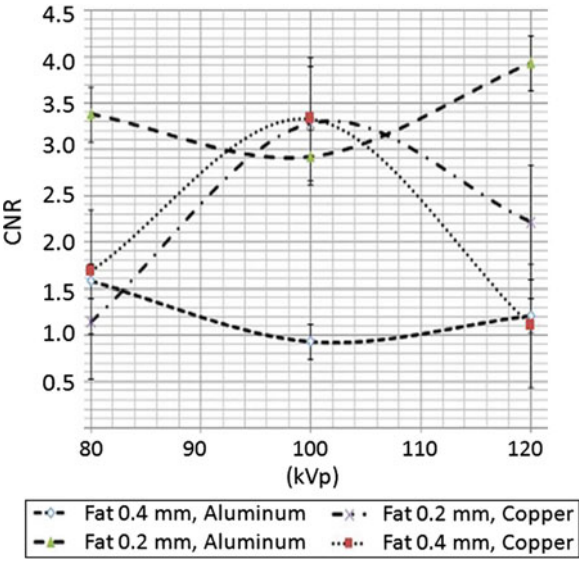


Fig. 8 The CNR fat versus tube potential with Copper and Aluminum filters at thickness of 0.2 and 0.4 mm

part of the body being imaged [8–10]. Selecting low tube potential does not contribute any information on the image as the details are hardly can be seen. It only gives absorbed dose on patient but does not contribute information of image.

Variation of optimization factors should be considered in the formation of images. To attain the correct balance between patient dose and image quality it is important to know appropriate parameters selected that affect image quality and radiation dose [11, 12]. Generally, added filtration such as Copper and Aluminum should absorb all low energy and transmit all high energy photons. For diagnostic radiology, Aluminum ($Z = 13$) and Copper ($Z = 29$) filters are usually selected for filtration [1].

Copper with an atomic number ($Z = 29$) is a better filter for absorbing a higher proportion of lower energy photons than Aluminum for most diagnostic X-ray beams. Optimization in x-ray radiographic imaging requires best balance between the CNR in the image and the applied radiation dose in order to optimize the kVp and filter settings. The most significant dose reduction is achieved through the use of Copper filter compare to Aluminum filter. This is because Copper filter has higher atomic number than Aluminum filter.

This study shows that 0.2 mm thickness of Copper filter gives better image quality and optimal dose at 100 kVp for chest imaging. The CNR of bone is 38.52 at 100 kVp compare to 32.01 at 80 kVp and 31.65 at 120 kVp. Fat also has greater CNR at 100 kVp compare to other parameters. Higher kilovoltage is preferred for high density structures as greater penetration will give a smaller range of beam intensities transmitted through the object. Calculations of CNR and dose for different features in chest phantom using different filter thickness and tube potential produce opportunity to examine the changes in radiation quality.

5 Conclusion and Future Work

In conclusion, reduction of dose can be achieved by varying the tube potential without changing the quality of the image. Optimization of x-ray radiographic imaging is also achievable with the use of Copper filter at 100 kVp with 0.2 mm filter thickness. For all imaging tasks, selection of suitable radiation quality is important as it will affect both radiation dose and image quality. Future work can be done by adding more filtration types and thickness to evaluate the radiation dose and image quality in x-ray imaging.

Acknowledgments The authors would like to thank Universiti Sains Malaysia for providing the travel fund to attend and present this work in the conference. The authors would like to thank Advanced Medical and Dental Institute, Universiti Sains Malaysia and School of Physics, Universiti Sains Malaysia for providing facilities for this research work and kind support throughout the project.

References

1. Zainon R, Abu Talib NS, Abu Bakar SN, Moner NHM, Aziz NAA (2014) Assessment and optimization of radiation dosimetry and image quality in x-ray radiographic imaging. In: Proceedings of the world congress on engineering, WCECS 2014. Lecture notes in engineering and computer science, San Francisco, USA, pp 508–511, 22–24 Oct 2014
2. Siebert JA, Shelton DK, Moore EH (1996) Computed radiography x-ray exposure trends. Acad Radiol 3:313–318
3. Cunnigham IA (1999) Image quality and dose. In Siebert JA, Filipow LJ, Andriole KP (eds) Practical digital imaging and PACS. American association of physicists in medicine monograph 25. Medical Physics Publishing, Madison, Wis, pp 225–228
4. Gökçe SD, Gökçe E, Coşkun M (2012) Radiology residents' awareness about ionizing radiation doses in imaging studies and their cancer risk during radiological examinations. Korean J Radiol 13(2):202–209
5. FDA (2010) Initiative to reduce unnecessary radiation exposure from medical imaging, center for devices and radiological health, U.S. Food and Drug Administration
6. The International Commission on Radiological Protection, The 2007 Recommendations of the International Commission on Radiological Protection: ICRP Publication 103, The International Commission on Radiological Protection
7. Frush D, Denham CR, Goske MJ, Brink JA, Morin RL, Mills TT, Butler PF, McCullough C, Miller DL (2012) Radiation protection and dose monitoring in medical imaging: a journey from. J Patient Saf 8:232–238
8. Gray VK (2010) Reducing radiation exposure in disagnostic imaging. Healthc Technol Online
9. Amis SE, Butler PF, Appelgat KE, Birnbaum SB, Brateman LF, Hevezi JM, Mettler FA, Morin RL, Pentecost MJ, Smith GG, Strauss KJ, Zeman RK (2007) American college of radiology white paper on radiation dose in medicine. Am Coll Radiol 4(5):272–284
10. Ron E (2003) Cancer risks from medical radiation. Health Phys 85(1):47–59
11. Breaking the trend of increased radiation exposure to patients through dose monitoring. http://www.sectra.com/medical/dose_monitoring/articles/breaking-the-trend-of-increased-radiation-exposure-to-patients-through-dose-monitoring-2013–07-05/
12. Fazel R, Krumholz HM, Wang Y, Ross JS, Chen J, Ting HH, Shah ND, Nasir K, Einstein AJ, Nallamothu BK (2009) Exposure to low-dose ionizing radiation from medical imaging procedures. N Engl J Med 361(9):849–857

Hybrid Computation Models for High Performance Biological Sequence Alignment on a Cloud System

Taylor Job and Jin H. Park

Abstract As a convenient high performance computation system, cloud system is more and more popularly used in the field of bioinformatics. We develop and examine several hybrid computation models for high performance biological sequence alignment on a cloud system. In our practice, Smith-Waterman and CloudBurst alignment algorithms are evaluated for performance with the computation models, which are built from combinations of current technologies including Hadoop, SHadoop and Fair Scheduler, on the Amazon Elastic Compute Cloud (EC2) system. In our experiment with relatively small data sets, the computation model with SHadoop showed the best performance for both Smith-Waterman and CloudBurst algorithms, i.e., speedup of 1.86× and 1.19×, respectively, over the baseline model (with Hadoop). For relatively large data sets, the computation model with SHadoop showed the best performance with Smith-Waterman algorithm (1.03×) and the computation model with SHadoop plus Fair Scheduler showed the best performance with CloudBurst algorithm (1.15×).

Keywords Biological sequence alignment · Cloudburst · Cloud system · Fair scheduler · Hadoop · High performance computing · SHadoop · Smith-Waterman algorithm

1 Introduction

The concept of Big Data and it's use is now everywhere in today's world. From business to academia, various organizations are now utilizing current technologies to access and manipulate information in the sea of ever-changing data.

T. Job · J.H. Park (✉)
Computer Science Department, California State University, Fresno, CA 93740, USA
e-mail: jpark@csufresno.edu

T. Job
e-mail: taylorjob@mail.fresnostate.edu

H.K. Kim et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-94-017-7236-5_26

Bioinformatics is a closely related field of manipulating Big Data, which encompasses a wide variety of biological data such as DNA, RNA and protein data from a wide variety of species. Most of the biological data are available in the form of very large text files, e.g., FASTA and FASTQ, and could take hours to days to process depending on the application and the scope of the operation.

Some significant ideas about analyzing large amounts of raw data come from Google's MapReduce [1] concept, which splits up the job and executes the parts on a cluster of computing nodes in a parallel manner. Apache Hadoop [2], the open source version of MapReduce, is a popular software tool used to implement this concept and is what we use in our practice described in this paper. Besides, cloud computing environment is a convenient way of accessing service, data and storage through internet, and more and more widely used in bioinformatics computing to achieve high performance. In the cloud system environment, users are able to scale the number of nodes needed for the job using a tool like Hadoop. Amazon's Elastic Compute Cloud (EC2) [3] is one of the most widely used commercial cloud computing services and has a good user interface which makes it easier to setup, strip down, or add more computing nodes to the cluster. In our practice, we use EC2 for the computing cluster.

While the usage of Hadoop on a cloud computing system does speed up the processing of bioinformatics operations, there have appeared some approaches of improving performance. Since Hadoop's default scheduler is based on a FIFO mechanism, the resulting throughput is low when many jobs are submitted. Apache Fair Scheduler [4] is a pluggable MapReduce scheduler and helps multiple users or multiple job submissions by supporting the mechanism of executing shorter jobs without delay and thus, increases the system throughput. SHadoop [5] is a fully compatible, optimized version of Hadoop, and its primary goal is reducing the execution time of MapReduce jobs, especially short jobs. This is done by optimizing the setup and cleanup operations of a job and by introducing an instant message system between the job-tracker and the task-trackers that speeds up the communication of scheduling information between them. Reducing the execution time of MapReduce is definitely beneficial when running an exhaustive bioinformatics sequence alignment application. CloudBurst [6] is a parallel read-mapping algorithm used to map DNA-sequence data to the reference genome. It is implemented on a Hadoop based cluster and aims to optimize the parallel execution. CloudBurst's creators claim that as the number of processors increases on a user's cluster, the speedup is near linear [6].

In this paper, we build hybrid computation models from the aforementioned technologies to exploit high performance on bioinformatics computing and our practice is limited to a couple of sequence alignment algorithms, Smith-Waterman [7] and CloudBurst [6] algorithms. In our practice, all experimentations are done with a Hadoop based cluster on Amazon EC2 cloud service as we did in our primitive work [8]. In particular, our experiments include evaluating Smith-Waterman and CloudBurst applications in four computation models, which are based on Hadoop, Hadoop plus Fair Scheduler, SHadoop and SHadoop plus Fair Scheduler.

The rest of this paper is organized as follows. In Sect. 2, a brief review of related work is presented. In Sect. 3, our proposed computation models and implementation methodologies are described. In Sect. 4, experimental results and performance analysis are presented, and Sect. 5 concludes the paper.

2 Related Work

In this section, we briefly review some recent approaches on running bioinformatics applications on cloud computing systems.

As the field of bioinformatics expands, so has the number of different ideas and approaches that use current data-driven technologies, such as a variety of cloud services as well as parallel frameworks and applications, to aid research and development. Research work in [9] examines performance from using Hadoop on a cloud computing system. The authors describe how to obtain an increase in performance by utilizing Hadoop on a cloud computing service. They explore different alignment tools and applications that perform sequence alignment including CloudBurst. A similar work is described in [10] in which the idea of using a parallel platform for executing the bioinformatics tool, dotplot, in a cloud environment is presented. In this work, Microsoft's Azure software is used to parallelize the execution of the tasks, instead of using Hadoop.

In the research described in [11], Google App Engine computing platform is used as the computational resource. The authors introduce the method of building the computer generated protein models used in the protein structure prediction. The proposed protein model comparator is their solution to the problem of large-scale model comparison and can be scaled for different data sizes.

A comparison of two bioinformatics applications on two cloud technologies is described in [12]. In this work, a pairwise Alu sequence alignment application and an Expressed Sequence Tag (EST) sequence assembly program are evaluated. Performance analysis is done with Apache Hadoop, Microsoft DryadLINQ and traditional MPI implementation. Authors also examine the effect of inhomogeneous data on the cloud technologies' scheduling mechanisms.

The research work described in [13] reviews existing cloud-based services in bioinformatics and classifies them into four groups, i.e., Data as a Service (DaaS), Software as a Service (SaaS), Platform as a Service (PaaS) and Infrastructure as a Service (IaaS). Based on some analysis work, authors conclude with their perspective on the adoption of cloud computing in bioinformatics.

Cloud BioLinux, a publicly accessible Virtual Machine, is described in [14]. This VM allows researchers to enable on-demand infrastructures for high performance bioinformatics computing on a cloud platform. A vast array of pre-configured command line and graphical software applications are included in the VM. Some of the applications include sequence alignment, clustering, assembly, display and editing.

3 System Components and Computation Models

Our series of computational experiments are conducted in two separate rounds, one for relatively small data set and the other for relatively large data set. The first round (round1) is conducted with a 6-node cluster (5 DataNodes and 1 NameNode) and the second round (round2) is conducted with an 11-node cluster (10 DataNodes and 1 NameNode) provided by Amazon's EC2 cloud service. We also utilize Amazon's Simple Storage Service (S3) to store and easily access data used in the computation. Figure 1 shows the structure of the system used in the round1.

3.1 Technologies Incorporated

The baseline technology used in our computation models is Hadoop [2] and we describe additional technologies, which are SHadoop and Fair Scheduler, in more detail in this section.

3.1.1 SHadoop

SHadoop [5] provides computational efficiency based on a couple of ideas. The first idea is that it optimizes given tasks that initialize and complete a MapReduce job,

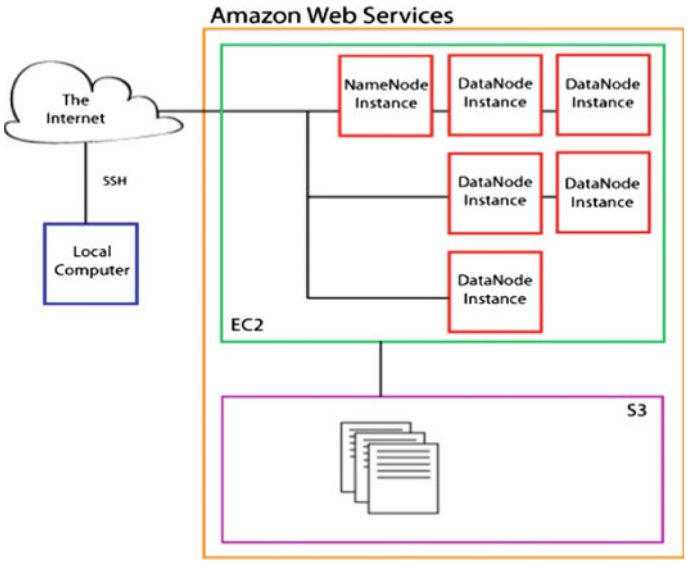


Fig. 1 Computing cloud platform

which in turn reduces the time taken to execute it. The second idea is that it uses an instant message communication system to relay all messages from JobTracker to TaskTrackers. This mechanism accelerates the performance-sensitive task scheduling as well as execution. It is reported that SHadoop can reach stable performance improvement of $\sim$25 % in average from testing benchmarks including WordCount, Sort, Grep, and Kmeans [5]. The performance improvement with SHadoop on standard exhaustive searching operations gave us the motivation of using it to improve the performance of data intensive computations in bioinformatics.

3.1.2 Fair Scheduler

Apache Fair Scheduler [4] is a pluggable MapReduce scheduler, which creates an interface for multiple job submissions to a single computing cluster to share CPU resources evenly. A single job, which is currently running, is allocated with the entire cluster, but when other jobs are submitted to the same cluster idle task slots are assigned to new jobs. This allows idle tasks to help parallelizing the execution. The Fair Scheduler organizes jobs into pools and allocates available resources equally among them. This optimization in scheduling improves the system throughput since smaller jobs can finish earlier without waiting for the completion of heavy jobs. The default scheduler used in Hadoop is based on the FIFO mechanism and thus, is inefficient comparing to the Fair Scheduler when multiple jobs are submitted to a single cluster. In the Fair Scheduler, there are certain number of parameters, which can be configured to allow even more customization, including preemption of jobs in other pools, limiting the number of jobs in a pool, etc.

3.2 Applications

3.2.1 Smith-Waterman Algorithm in MapReduce

We implemented Smith-Waterman algorithm [7], which is a non-heuristic local sequence alignment algorithm, in Python with MapReduce. Since the alignment score is computed based on the computation of $n * m$ score table entries, where n and m represent lengths of reference and pattern strings, respectively, we exploited a coarse-grained parallelism by partitioning the score table region into pieces. For the sake of simplicity, we used linear gap scoring scheme in our practice. The recurrence relation of computing the entries of the score table (H) is shown below and Fig. 2 illustrates the concept of exploiting the parallelism by partitioning the score table entries into coarse-grained blocks for MapReduce implementation.

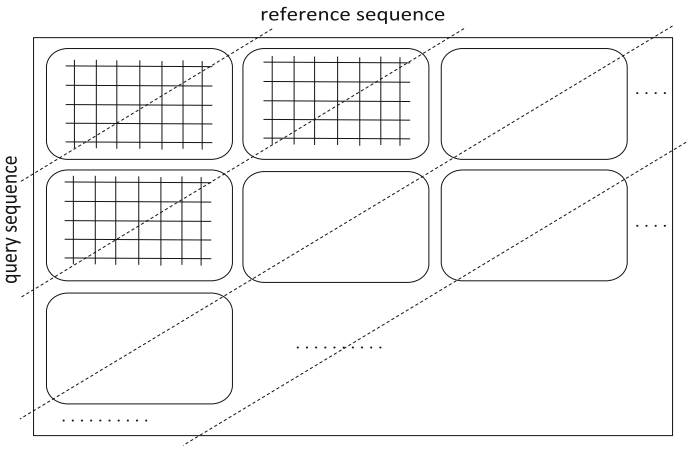


Fig. 2 Coarse-grained parallel processing on H: blocks on a diagonal at a time

$$
\begin{aligned}
H_{i,j} = \max\{ & H_{i-1,j-1} + \delta(a_i, b_j), \quad //\delta \textit{ is similarity score} \\
& \max_{k \geq 1}\{H_{i-k,j} - W_k\}, \quad //W_k \textit{ is weight of } k \textit{ indels} \\
& \max_{l \geq 1}\{H_{i,j-l} - W_l\}, \quad //W_l \textit{ is weight of } l \textit{ indels} \\
& 0\}, \quad \text{where} \quad 1 \leq i \leq n \text{ and } 1 \leq j \leq m;
\end{aligned}
$$

3.2.2 CloudBurst

CloudBurst [6] is a seed-and-extend based algorithm, which maps short query sequences (reads) to the reference genome, and implemented in Hadoop (MapReduce) to exploit parallelism. Compared with RMAP [15], which is an early day's short read mapping tool, it is reported that CloudBurst achieves speedup of up to 30 times faster and it reduces the execution time of a job from hours to minutes in a large cluster with 96 cores [6]. It is also reported that the near linear speedup is achieved as the number of processors increases in the system [6]. CloudBurst is available as open-source, and Fig. 3 (from [6]) illustrates the operational overview of the MapReduce mechanism used in CloudBurst.

3.3 Proposed Computation Models

In our practice, we developed four computation models and tested biological sequence alignment algorithms, which are Smith-Waterman and CloudBurst, on Amazon EC2 cloud system for performance comparison. Figure 4 shows the

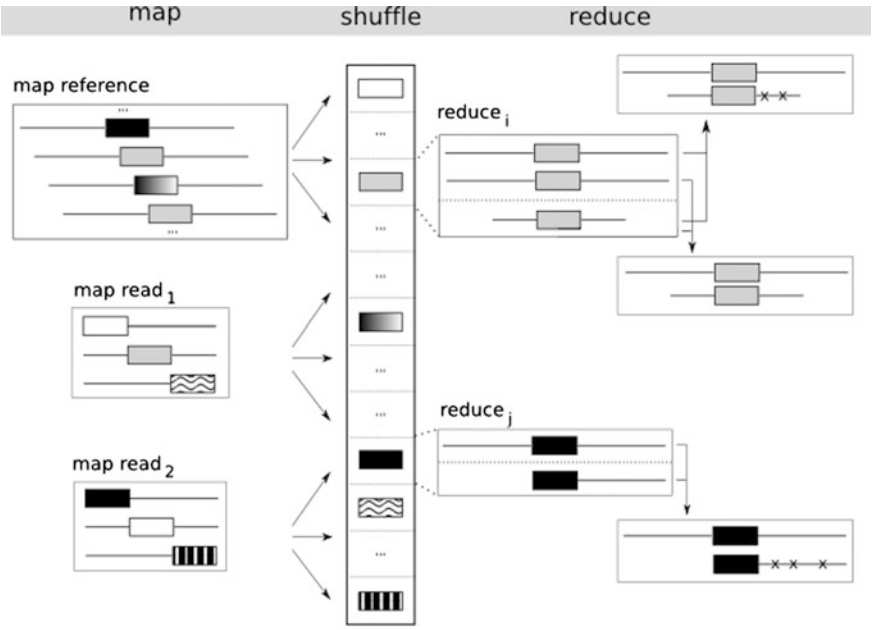


Fig. 3 Overview of CloudBurst algorithm

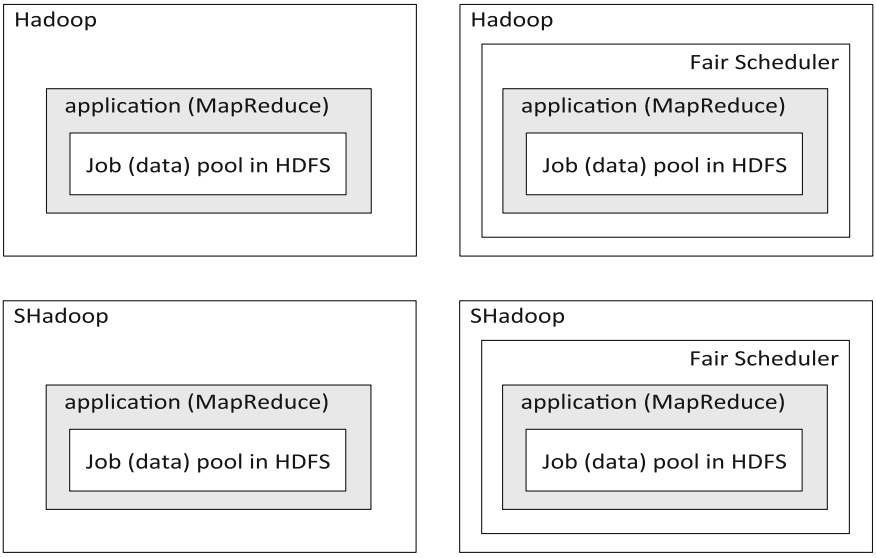


Fig. 4 Computation models

computation models, which we built from combinations of current technologies including Hadoop, SHadoop and Fair Scheduler. The model with Hadoop (without Fair Scheduler) is the baseline model for the purpose of performance comparison. Data sets of an application are copied from S3 storage to HDFS (Hadoop File System) before execution. Our interest is observing performances of the models with SHadoop and/or Fair Scheduler.

The computation models are portable for different types of applications. Since we use two applications, i.e., Smith-Waterman and CloudBurst, in our practice the computation models are implemented in eight different ways—four models for each application.

4 Experimental Results

For the implementation of the computation models that we proposed, we use 6 Large Instances (5 DataNodes (slaves) and 1 NameNode (master)) for the round1 experiment and 11 Large Instances (10 DataNodes (slaves) and 1 NameNode (master)) for the round2 experiment. A Large Instance (m3.large) in EC2 has 2 virtual cores and 7.5 GB of storage. Software installation/setup are done successfully with the following packages, which we accessed from Internet sources: Hadoop, SHadoop, Fair Scheduler and CloudBurst. As described in the previous section, the Smith-Waterman algorithm is implemented in Python with MapReduce mechanism to achieve high performance.

To configure the master and slaves during Hadoop cluster setup we used the following operations, and we skip describing detailed installation/setup steps of the software in this paper.

Configure Master and Slaves:
On NameNode:

$ sudo vi masters
Copy in Public DNS from NameNode
$ sudo vi slaves
Copy in Public DNS from every DataNode

On each DataNode:

$ sudo vi masters
Keep blank
$ sudo vi slaves
Copy in Public DNS from current DataNode

To measure the performance of the computation models (refer Fig. 4) with the selected two applications, we conducted two rounds (round1 and round2) of testing, one with relatively small prokaryotic genomes and the other one with relatively big eukaryotic chromosomes. Data used in the round1 include an EColi.fa (5.6 MB),

Salmonella.fa (4.9 MB) and Streptococcus_suis.fa (1.9 MB). Data used in round2 include HG19 chr6.fa (174.5 MB), chr13.fa (117.5 MB) and chr17.fa (82.8 MB). In each round, to observe the performance differences among different computation models we ran the 3 jobs in the order of the largest to the smallest and in the order of the smallest to the biggest and computed the average execution time of the two trials. The graphs shown in Figs. 5 and 6 illustrate the results of the round1 and round2, respectively.

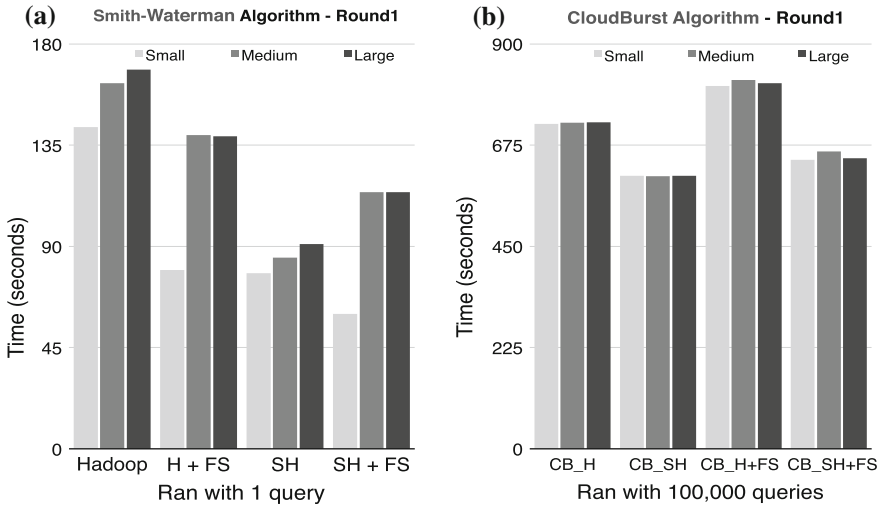


Fig. 5 Performance of computation models—round1

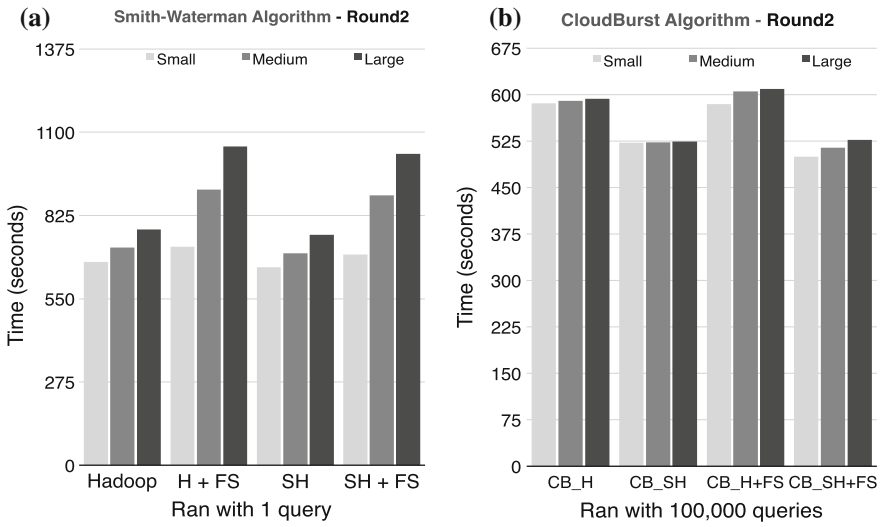


Fig. 6 Performance of computation models—round2

In Fig. 5a, Hadoop represents the baseline model (Hadoop with default FIFO scheduler) and H + FS, SH and SH + FS represent Hadoop with Fair Scheduler, SHadoop, and SHadoop with Fair Scheduler, respectively. As shown in Fig. 5a, for Smith-Waterman algorithm, using Fair Scheduler with Hadoop achieves reasonable performance gain over simply using Hadoop, but it does not guarantee the gain with SHadoop. In fact, using only SHadoop shows the best performance overall. Figure 5b shows performances with CloudBurst algorithm. In the figure, CB_H represents CloudBurst with Hadoop (baseline model) and CB_H + FS, CB_SH and CB_SH + FS represent CloudBurst (with Hadoop) with Fair Scheduler, CloudBurst with SHadoop, and CloudBurst with SHadoop and Fair Scheduler, respectively. The best performance gain is observed with the model of CloudBurst with SHadoop (CB_SH). With this application, using Fair Scheduler shows worse performance than without using it.

With the round2 experiment (refer Fig. 6), we observed that the model with only SHadoop shows the best performance with Smith-Waterman algorithm thought the performance gain is minimal comparing to the baseline model (Hadoop only). In the round2, which manipulates relatively large data set, it is observed that CloudBurst application yields the best performance with the computation model with SHadoop and Fair Scheduler.

In conclusion, using SHadoop achieves the best performance with both Smith-Waterman (1.86× speedup over baseline model) and CloudBurst (1.19× speedup over baseline model) algorithms in our round1 practice, and also shows the best performance with Smith-Waterman algorithm (1.03×) in the round2 practice. An exception is that the model with SHadoop and Fair Scheduler shows the best performance with CloudBurst algorithm (1.15×) in the round2 practice.

5 Conclusion and Discussion

In this research, we built four hybrid computation models on a cloud computing system to accelerate biological sequence alignment applications, and tested for their performances. Although our practice is limited with a couple of sequence alignment algorithms we observed that using SHadoop yields the best performance for both Smith-Waterman and CloudBurst algorithms with relatively small data set. In the case of manipulating relatively large data set, SHadoop shows the best performance for Smith-Waterman algorithm and the combination of SHadoop and Fair Scheduler shows the best performance for CloudBurst algorithm. Different from our initial guess, adding Fair Scheduler to SHadoop did not always show the best performance, except the case with CloudBurst with relatively large data set. We plan to test more diversified bioinformatics applications with the proposed computation models with various size reference databases to yield more comprehensive results.

References

1. Dean J, Ghemawat S (2008) MapReduce: simplified data processing on large clusters. Commun ACM 51(1):107–113
2. Apache Hadoop. http://hadoop.apache.org. Accessed on Mar 2014
3. Amazon Elastic Compute Cloud. http://www.amazon.com/ec2/. Accessed on Mar 2014
4. Apache Fair Scheduler. http://hadoop.apache.org/docs/r2.3.0/hadoop-yarn/hadoop-yarn-site/FairScheduler.html. Accessed on Mar 2014
5. Gu R, Huang Y, Sun Y et al (2014) SHadoop: improving MapReduce performance by optimizing job execution mechanism in hadoop clusters. J Parallel Distrib Comput 74 (3):2166–2179
6. Schatz M (2009) CloudBurst: highly sensitive read mapping with MapReduce. Bioinformatics 25(11):1363–1369
7. Smith TF, Waterman MS (1981) Identification of common molecular subsequences. J Mol Biol 147:195–197
8. Job T, Park JH (2014) Exploiting high performance on bioinformatics applications in a cloud system. In: Proceedings of the world congress on engineering and computer science, WCECS 2014. Lecture notes in engineering and computer science, San Francisco, USA, pp 563–566, 22–24 Oct 2014
9. Quan Z, Wen-rui J, Xu-bin L, Yi J (2012) Hadoop applications in bioinformatics. In: Proceedings of the 7th open cirrus summit, pp 48–52
10. Cano M, Karlsson J, Klambauer G, et al (2012) Enabling large-scale bioinformatics data analysis with cloud computing. In: Proceedings of the 10th IEEE international symposium on parallel and distributed processing with application, pp 640–645
11. Widera P, Krasnogor N (2011) Protein models comparator: scalable bioinformatics computing on the Google App Engine platform. Comput Res Repository 1:1–8
12. Ekanayake J, Gunarathne T, Qiu J (2011) Cloud tecnologies for bioinformatics applications. IEEE Trans Parallel Distrib Syst 22(6):998–1011
13. Dai L, Gao X, Guo Y, Zhang Z (2012) Bioinformatics clouds for big data manipulation. Biol Direct 7(43):1–7
14. Krampis K, Booth T, Chapman B, Tiwari B, Bicak M, Field D, Nelson K (2012) Cloud BioLinux: pre-configured and on-demand bioinformatics computing for the genomics community. BMC Bioinf 13(42):1–8
15. Smith AD et al (2008) Using quality scores and longer reads improves accuracy of Solexa read mapping. BMC Bioinf 9:128

Modeling the Impact of International Travellers on the Trend of the HIV/AIDS Epidemic

Ofosuhene Okofrobour Apenteng and Noor Azina Ismail

Abstract This study aimed to determine the impact of international travellers (tourists) on the prevalence of HIV/AIDS. A compartmental epidemic model was extended to include the effect of tourists on the spread of HIV/AIDS. Nonlinear ordinary differential equations were used to describe the rate of change in the number of individuals in the respective compartments. Tourists are defined as travellers who leave their country for a period of time. The reproduction number was calculated to provide new insight into the spread of HIV/AIDS. The results indicated that the constant flow of tourists into a country increases the spread of these diseases, although it does not change the status of the HIV/AIDS spread from endemic to epidemic.

Keywords Basic reproduction number · Epidemiology · HIV/AIDS epidemic · Mathematical model · Numerical optimization · SIA model · Sexual transmitted disease · Simulation · Stability · Tourists

1 Introduction

Tourism is one of the fastest growing industries in many countries. Tourism enables successful business activity and generates considerable revenue and publicity for many countries worldwide. The World Tourism Organization (UNWTO) defines tourism as *the activities of persons travelling to and staying in places outside their usual environment for not more than one consecutive year for leisure, business and other purposes* [1]. The tourism sector is frequently associated with casual sex and

O.O. Apenteng (✉) · N.A. Ismail
Department of Applied Statistics, Faculty of Economics and Administration, University of Malaya, 50603 Kuala Lumpur, Malaysia
e-mail: oapenten@siswa.um.edu.my

N.A. Ismail
e-mail: nazina@um.edu.my

H.K. Kim et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-94-017-7236-5_27

is thus linked to an increased risk of human immunodeficiency virus (HIV) transmission. It is estimated that 50 % of travellers engage in sexual practices while abroad [2]. For example, in 2004, the news agency Cable News Network (CNN) reported that an estimated 16,000–20,000 people are child sex victims in Mexico, largely in border, urban, and tourist areas [3]. In the United Kingdom, a study was conducted involving 258 heterosexual travellers who attended the Hospital for Tropical Diseases in London. The study indicated that the rate at which foreigners acquired HIV infection was much higher than that of locals [4]. In Europe, there has been an increase in new AIDS cases among immigrants [5]. This increase is no different from that in Asia, India and Africa or among migrant laborers, long-distance truck drivers, and commercial sex workers. HIV epidemics travel to other countries for green pasture.

The relationship between HIV/AIDS and tourism is not well studied, and most conclusions drawn regarding this relationship were based on speculation and unreliable evidence. Many travellers find vacations or trips abroad as an occasion to be involved in sexual activity and possibly drug use, as these activities become more available and affordable [5–9] abroad than at home. International tourists who engage in unprotected sex with sex workers, hotel workers and others in the local population may create a bridge for HIV/AIDS to be transferred between the tourist's home country and the tourist's destination.

The implications of human immunodeficiency virus (HIV), which causes acquired immunodeficiency syndrome (AIDS), control efforts for international travellers are a growing concern throughout the world [10]. There have been several reports about the modeling of HIV transmission to identify factors that might contribute to its spread. Unfortunately, there is little understanding of how to model the impact of tourism on the spread of HIV in the public health field. In view of this knowledge gap, we proposed a mathematical model to study the impact of the constant inflow of tourists on the HIV epidemic. Two scenarios were considered. First, we considered the constant influx of tourists into all three population compartment levels, and second, we considered no direct inflow of tourists into any of the three population compartment levels. We used data from Malaysia to test the proposed model because in the absence of well conducted studies, the available literature provides a broad sense of the risk and prevention practices associated with the contribution of tourism to the spread of HIV epidemics. This work is modified version of paper presented at International Conference in Modeling Health Advances 2014 (ICMHA) [11].

2 Model Description and Analysis

The model describes three components: the number of susceptible $S(t)$, infected (HIV) $I(t)$, and AIDS cases who are exhibiting AIDS symptoms $A(t)$ at time t. It was assumed that at any time, new sexually active locals and international tourists T_s, T_1 and T_A enter the susceptible, HIV and AIDS compartments, respectively. It

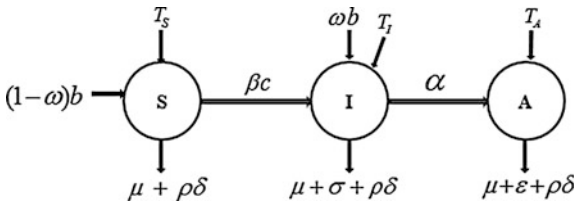


Fig. 1 Model of the effect of tourism on HIV/AIDS

was assumed that new locals of birth rate ωb enter the HIV class. A proportion ω of these individuals were assumed to be infected with HIV (categorized in the I class) and a complementary proportion $(1-\omega)b$ was assumed to be in or move to the susceptible class, S. The natural death rate was assumed to be the same in each class, $\mu > 0$. AIDS patients had an additional disease-related mortality rate, $\varepsilon > 0$. Moreover, c_i represented the rate at which an individual acquired new sexual partners, where the subscript i corresponds to the presence or absence of tourism. We further assumed that c was 0.05 % of the tourists. Additionally, δ represented the proportion of tourists who left the population, and ρ was the proportion of the actual visit time to the maximum allowable visit time by a tourist. The probability of transmission from an individual in category $S(t)$ to the HIV-infected category $I(t)$ was represented by β. A description of the model is shown in Fig. 1.

Based on the above assumptions and the description of the model, along with the demographic parameters (birth and death), the model consisted of three nonlinear ordinary equations:

$$\frac{dS}{dt} = (1-\omega)bS + T_S - \beta cIS - (\mu + \rho\delta)S \tag{1}$$

$$\frac{dI}{dt} = T_I + \omega bI + \beta cIS - (\mu + \alpha + \sigma + \rho\delta)I \tag{2}$$

$$\frac{dA}{dt} = T_A + \alpha I - (\mu + \varepsilon + \rho\delta)A \tag{3}$$

where $c = 0.05\,\%T$.

From (1)–(3), the total population was given by $N = S + I + A$, where $T_T = T_S + T_I + T_A$. Therefore, solving for $\left(\frac{dN}{dt}\right)$ and substituting S in the form of N and I into (2) produced

$$\frac{dN}{dt} = (1-\omega)bN + 2\omega\, bI - bI - (1-\omega)bA + T_T - \mu N - \rho\delta N - \sigma I - \varepsilon A \tag{4}$$

$$\frac{dI}{dt} = T_I + \omega\, bI + \beta cI(N - I - A) - (\mu + \alpha + \sigma + \rho\delta)I \tag{5}$$

$$\frac{dA}{dt} = T_A + \alpha I - (\mu + \varepsilon + \rho\delta)A \tag{6}$$

where $N(0) = N_0 > 0$, $I(0) = I_0 \geq 0$, and $A(0) = A_0 \geq 0$.

The derivatives of (4)–(6) implied that the model was well constructed for $N > 0$. We assumed that all dependent variables and parameters of the model were non-negative. To solve these equations, we analyzed models (4)–(6) using the stability theory of differential equations.

3 Equilibrium and Stability Analysis

In this section, we analyzed the stability of the points of equilibrium in models (4)–(6).

3.1 Equilibrium of the Model

The model did not show a disease-free equilibrium due to the direct influx of tourists into all the three population compartment levels. There would be an endemic equilibrium $E^* = (N^*, I^*, A^*)$. Equating (4)–(6) to 0, we produced

$$[(1-\omega)b - \mu - \rho\delta]N^* + (2\omega b - b - \sigma)I^* - [(1-\omega)b - \varepsilon]A^* + T_T = 0 \tag{7}$$

$$T_I + \omega b I + \beta c I^*(N^* - I^* - A^*) - (\mu + \alpha + \sigma + \rho\delta)I^* = 0 \tag{8}$$

$$T_A + \alpha I^* - (\mu + \varepsilon + \rho\delta)A^* = 0 \tag{9}$$

By solving (7)–(9) simultaneously, we obtained

$$I^* = \frac{(\mu + \varepsilon + \rho\delta)\{[(1-\omega)b - \mu - \rho\delta]N + T_T\} + T_A[\varepsilon - (1-\omega)b]}{\alpha[(1-\omega)b - \varepsilon] - (\mu + \varepsilon + \rho\delta)(2\omega b - b - \sigma)} \tag{10}$$

$$A^* = \frac{1}{(\mu + \varepsilon + \rho\delta)}\{T_A + I^*\} \tag{11}$$

which are all positive when

$$N < \frac{T_T}{[(1-\omega)b - \mu - \rho\delta]}.$$

To search for steady states in the absence of tourism, we solved Eqs. (1)–(3) by setting $T_T = 0$ and $c_1 = 0$.

$$(1-\omega)bS - \beta IS - \mu S = 0 \tag{12}$$

$$\omega bI + \beta IS - (\mu + \alpha + \sigma)I = 0 \tag{13}$$

$$\alpha I - (\mu + \varepsilon)A = 0 \tag{14}$$

From (13) and (14) we obtained, respectively

$$I = 0 \text{ or } S = \frac{(\mu + \alpha + \sigma) - \omega b}{\beta} \tag{15}$$

$$A = \frac{\alpha I}{\mu + \varepsilon}, \tag{16}$$

respectively. From (15), we defined the reproduction number as

$$R_0 = \frac{\beta}{\mu + \alpha + \sigma - \omega b}.$$

The total population was given by the following equation.

$$S_e + I_e + A_e = N_e \tag{17}$$

Thus,

$$I_e = \frac{(\mu + \varepsilon)[\beta N - (\mu + \varepsilon + \sigma)] + \omega b}{\beta(\mu + \varepsilon + \sigma)} \tag{18}$$

$$A_e = \frac{\alpha(\mu + \varepsilon)[\beta N - (\mu + \varepsilon + \sigma)] + \omega b}{\beta(\mu + \varepsilon + \sigma)} \tag{19}$$

These equations led to a unique endemic equilibrium given by

$$E_e = \left\{ \frac{(\mu+\alpha+\sigma)-\omega b}{\beta}, \frac{(\mu+\varepsilon)[\beta N-(\mu+\varepsilon+\sigma)]+\omega b}{\beta(\mu+\varepsilon+\sigma)}, \frac{\alpha(\mu+\varepsilon)[\beta N-(\mu+\varepsilon+\sigma)]+\omega b}{\beta(\mu+\varepsilon+\sigma)} \right\}.$$

4 Local Stability and Points of Equilibrium

To determine whether the disease continued to spread, we needed to find the stability of the point of disease-free equilibrium. The reproduction number provides a vivid description of the average number of secondary infections caused by each infected individual throughout that individual's period of infection, thereby

providing new insight into the probable course of the infection [12–15]. The reproduction number is the expected number of secondary cases produced by a typical infection in a completely susceptible population [15]. The disease cannot persist if $R_0 < 0$. The condition $R_0 < 0$ fits both the local and global stability of the disease-free equilibrium. Thus, the disease cannot continue to spread if $R_0 < 0$. Therefore, to determine whether the disease would continue to spread, we examined the dynamics for $R_0 > 1$. If the disease persists in the system $R_0 > 1$, then either the system is stable near the interior equilibrium or a periodic attractor potentially destabilizes the system. To assess these possibilities, we used the above conditions to calculate the reproduction numbers corresponding to various scenarios. For all time $t > 0$, all parameters in the model system described by (4)–(6) were assumed to be non-negative. This model had a disease-free equilibrium at $E_0 = (N, 0, 0)$, in which no disease is present at the initial stage, implying that $N = \frac{T_T}{\mu+\rho\delta-(1-\omega)b}$ when $I(0) = I_0 \geq 0$ and $A(0) = A_0 \geq 0$.

Considering no flow of tourism, we found the reproduction number R_0 for the entire model. Using the notations by van den Driessche and Watmough [15], we let $F_i(x)$ be the rate of appearance of new infections in compartment i and let $V_i(x)$ represent the rate of movement from one compartment to another. Thus, the right-hand sides of (4)–(6) were solved as $F - V$, where

$$F = \begin{pmatrix} 0 \\ \beta cI(N - I - A) \\ 0 \end{pmatrix} \tag{20}$$

$$V = \begin{pmatrix} \mu N + \rho\delta N - (1-\omega)bN + bI - 2\omega bI + \sigma I + (1-\omega)bA + \varepsilon A \\ -\omega bI + (\mu + \alpha + \sigma + \rho\delta)I \\ -\alpha I + (\mu + \varepsilon + \rho\delta)A \end{pmatrix} \tag{21}$$

Then, we considered the Jacobian matrices associated with F and V.

$$J_F = \begin{pmatrix} 0 & 0 & 0 \\ 0 & \beta cN & 0 \\ 0 & 0 & 0 \end{pmatrix} \tag{22}$$

$$J_V = \begin{pmatrix} \mu + \rho\delta - (1-\omega)\, b & b - 2\omega b + \sigma & (1-\omega)\, b + \varepsilon \\ 0 & -\omega b + \mu + \alpha + \sigma + \rho\delta & 0 \\ 0 & -\alpha & \mu + \varepsilon + \rho\delta \end{pmatrix} \tag{23}$$

The basic reproduction numbers obtained from (4)–(6) were contained in a spectral radius of the matrix $J_F \times J_V^{-1}$ at the disease-free equilibrium, described as

$$R_{01} = \frac{\beta cN}{\mu + \alpha + \sigma + \rho\delta - \omega b}. \tag{24}$$

To find the reproduction number of the entire model (including tourism), we substituted (25) into (24). This method implied the following equations.

$$N = \frac{T_T}{\mu + \rho\delta - (1-\omega)b} \tag{25}$$

and

$$R_{01} = \frac{\beta c T_T}{(\mu + \rho\delta + \omega b - b)(\mu + \alpha + \sigma + \rho\delta - \omega b)} \tag{26}$$

.

5 Analysis and Results

The first reported cases of HIV/AIDS in Malaysia were in 1986. Between 1986 and 2010, 91,362 men, women and children have been observed to be infected with HIV, and 12,943 individuals have died of HIV/AIDS [16]. In 2010, there were 16,452 Malaysian people with AIDS, and the Malaysian population aged between 15 and 49 years was recorded as 16,081,300 individuals [17]. Thus, 15,973,486 individuals were included in the susceptible class. In the same year, a total of 24,577,196 tourists entered the country [18]. Based on the above information, the following sets of initial conditions were established, as presented in Table 1. These conditions were used to calculate the reproduction numbers.

The results of this analysis showed that if $R_0 > 1$, the disease-free equilibrium was unstable and HIV/AIDS infected the host population. However, if $R_0 < 1$, then the disease-free equilibrium was stable. Based on the analysis of the model, the reproduction number was used to determine whether the disease-free equilibrium was stable or unstable. We found that $R_{01} = 0.1875$ for the case in which there was

Table 1 Values inserted for each parameter

Parameter	Value	Source
μ	0.0046	[16]
b	0.0176	[16]
δ	0.8596852	[18, 19]
α	0.1461574	[19]
ρ	14 days	–
σ	0.2132871	[19]
ω	0.6	–
β	0.0033	[16]
ε	0.79867129	[16]

a constant flow of tourists and $R_0 = 0.0239$ for the case in which there was no tourism. These results implied that tourism does not disturb the equilibrium status of the disease, although the reproduction number slightly increased in the presence of tourism.

6 Conclusions and Future Work

The primary objective of this study was to formulate and analyze a deterministic mathematical model of the impact of international travel (tourism) on the trend of the HIV epidemic in the human population using a system of nonlinear ordinary differential equations. The results showed that the constant flow of tourists into a country does not disturb its disease-free equilibrium, although its reproduction number slightly increases.

Further investigation is necessary to obtain realistic values for each parameter. Considering these challenges, our results require further investigation before being used as a guide for the early treatment of infected individuals. However, these results suggest that the government should make further efforts to curb illegal activities performed by tourists. It is critical to encourage greater awareness regarding the personal risk of contracting and spreading HIV/AIDS among tourists and those in contact with them. Finally, the proposed model will be tested against real-world data to test its fitness and to evaluate its accuracy.

Acknowledgments This work was supported by the University Malaya Research Grant RP004 J-13ICT: Demographic Network Modelling of the Spread of Infectious Diseases, under the Equitable Society Research Centre, University of Malaya.

References

1. Ugurlu T (2010) Definintion of tourism (UNWTO definition of tourism)/what is tourism ?
2. Ericsson CD et al (2001) Sexually transmitted diseases in travelers. Clin Infect Dis 32 (7):1063–1067
3. Courson P (2004) Japan put on sex-trade watch list. 2004 (cited 2014 20/3/14). Available from: http://edition.cnn.com/2004/US/06/14/trafficking.report/
4. Hawkes S et al (1992) HIV infection among heterosexual travellers attending the hospital for tropical diseases London. Genitourin Med 68(5):309–311
5. Sinka K et al (2003) Impact of the HIV epidemic in sub-Saharan Africa on the pattern of HIV in the UK. Aids 17(11):1683–1690
6. Abdullah ASM et al (2002) Risk factors for sexually transmitted diseases and casual sex among chinese patients attending sexually transmitted disease clinics in Hong Kong. Sex Transm Dis 29(6):360–365
7. Abdullah ASM et al (2004) Sexually transmitted infections in travelers: implications for prevention and control. Clin Infect Dis 39(4):533–538
8. Ocha W, Earth B (2013) Identity diversification among transgender sex workers in Thailand's sex tourism industry. Sexualities 16(1–2):195–216

9. Wilson A (2010) Post-fordist desires: the commodity aesthetics of Bangkok sex shows. Feminist Legal Stud 18(1):53–67
10. Lewis ND (1989) AIDS and tourism: implications for Pacific island states, vol 1. Pacific Islands Development Program, East-West Center
11. Apenteng OO, Ismail NA (2014) Modelling the Impact of International Travellers on the Trend of HIV Epidemic. In: Proceedings of the world congress on engineering and computer science 2014, WCECS 2014. Lecture notes in engineering and computer science, San Francisco, USA, 22–24 October, 2014. .pp 776–780
12. Gran JM et al (2008) Growth rates in epidemic models: application to a model for HIV/AIDS progression. Stat Med 27(23):4817–4834
13. Roberts M, Heesterbeek J (2007) Model-consistent estimation of the basic reproduction number from the incidence of an emerging infection. J Math Biol 55(5):803–816
14. Hethcote HW (2000) The mathematics of infectious diseases. SIAM Rev 42(4):599–653
15. van den Driessche P, Watmough J (2005) Reproduction numbers and sub-threshold endemic equilibria for compartmental models of disease transmission. Math Biosci 180(1–2):29–48
16. UNICEF (2011) AIDS in Malaysia. 2011 (cited 2014 20/03.14). Available from: http://www.unicef.org/malaysia/aids.html
17. *Department of Statistics: Population Projections Malaysia 2010-2040*. 2012
18. Tourist Arrivals by Country of Residence, Malaysia 2000–2011. 2012, Malaysian Tourism Promotion Board
19. Health Facts 2012 (2012) Ministry of health, pp 1–8

Antibodies of HCV

Bhagwan D. Aggarwala

Abstract We present a mathematical model which describes the development of HCV, and its resistant variants, in a patient. We assume that, apart from the variants that are already in the patient's blood stream, it requires only one more mutation at a specific nucleotide for an HCV virus to become resistant to the antiviral drug being administered, i.e. for u_0 (virus, together with all its variants, present when the treatment starts) to change into u_1, virus which is resistant to the drug. We assume that, in the presence of drug pressure, it is easier for u_0, to change to u_1 than the other way around. The Model will say that there are exactly two outcomes of treatment: either the patient has a REBOUND of virus or SVR, sustained viral recovery. The model will also outline the important role of a patient's immune system and say that if the immune system of the patient is strong enough, then HCV does not take hold. Finally, we shall also study how sensitive the results of our model are to changes in the treatment regime and/or to changes in the numerous parameters in the system.

Keywords Differential equations · Hepatitis C virus · Immune system · Rebound · Robustness · Sustained virologic response

1 Introduction

One hundred and seventy million people are infected with HCV worldwide [1]. In the United States, more than five million people are supposed to be living with HCV [2]. Approximately 30,000 new cases are diagnosed each year. This situation is likely to get worse as the number of people infected with HCV from blood transfusions before 1990 come to be newly diagnosed. This is because, before

B.D. Aggarwala (✉)
Department of Mathematics and Statistics, University of Calgary, Calgary T2N 1N4, Canada
e-mail: aggarwal@ucalgary.ca

H.K. Kim et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-94-017-7236-5_28

1990, there was no screening of blood against HCV, so that millions of patients must have been infected through blood transfusions. These cases are now coming to light.

Presently, there is no vaccine against HCV, and the standard treatment consists of weekly doses of peginterferon alpha and daily doses of ribavirin along with some protease inhibitor. In the beginning, cases of HCV were treated with ribavirin only but with very little success [3]. When peginterferon alpha was added, a sharp drop in virion was observed within a couple of days. However, even this treatment is unsuccessful in over half the patients, and many of these non responders go on to develop cirrhosis and then liver cancer. If the liver is transplanted in such patients, chances of their getting infected again are relatively high. The treatment is incredibly costly, which must leave some marginalized sections of the society suffering from the disease without any hope of cure. Government sponsored, universal health care schemes, like the one in Canada, should help.

Like the Human Immunodeficiency virus (HIV), HCV can stay dormant for twenty years and more while attacking the liver all this time. This accounts for HCV cases transmitted through blood transfusions before 1990 now coming to light. HCV mutates easily which makes for a large number of mutant viruses. There are six known genotypes (numbered 1 through 6) and more than 50 subtypes (e.g., 1a, 1b, 2a…) [4]. Because HCV mutates easily, some mutated virus is observed in patients who have never been treated, so that HCV exists in infected patients as HCV quasi-species.

While HIV has received major attention from the medical community in recent years, HCV is just as serious. While it is true that HIV positivity was a death sentence before the discovery of HAART, and is a manageable illness now, HCV is still a death sentence for a large percentage of people that get infected. It has been suggested that, apart from the liver, which is the main target of the virus, HCV may also affect the nervous system [5]. The genotype 1 of HCV is responsible for most of the infections in North America.

We present a mathematical model which describes the development of HCV, and its resistant variants, in a patient. It is known that, in an HCV virus, some virus mutations are hundreds of times more effective against the drug being administered than others. As an example, it has been reported that the variant V36A/M confers $\sim$3.5-fold resistance, whereas A156V/T confers $\sim$466-fold resistance to telaprevir [6]. Ignoring the mild resistance, we assume that, apart from the variants that are already in the patient's blood stream, it requires one more mutation, **at a specific nucleotide**, for an HCV virus to become resistant to the antiviral drug being administered, i.e. for u_0 (virus, together with all its variants, present when the treatment starts) to change into u_1 (virus which is resistant to the drug being administered). We assume that, in the presence of drug pressure, it is easier for u_0 to change to u_1 than the other way around, so that we assume that the probability of u_1 changing to u_0 is much smaller than the one of u_0 changing to u_1. We also assume that u_0 changes to u_1 after **one specific mutation at a given nucleotide**. HCV has approximately 9600 nucleotides, and its copying mechanism is error prone at the rate of 1 in about 10,000. The virus lives for 2–3 h outside a cell, so that new

viruses are being produced inside the infected cells at about the same rate. On average, it replicates about ten times in a day. The probability of its mutating at any **given** site in 9 replication cycles comes out to be 9.37031×10^{-8} and 1.04109×10^{-7} in 10 such cycles. We take this probability to be 10^{-7} which is the value of Q_1 in our model.

We also consider the effect of antibodies in our model. The antibodies are produced in response to the presence of the virus and decay when they encounter a virus. We show that if the rate of production of antibodies is high enough, the virus does not develop a chronic state. This says that, if the (adaptive) antibody response of the host is strong enough, an HCV infection does not take hold. We speculate that this may be the reason why a significant number of HCV infected patients do not develop the chronic state of the disease.

Our model will also say that, depending upon the antibody response of the host, there is a chronic state of the disease in an untreated patient. This state is an equilibrium state of our model in the absence of any treatment. However, there is also an (unstable) equilibrium state in a treated patient when the drugs have taken effect and reduced the virus count but the virus has not developed any resistance yet (so that $u_1 = 0$), and also an equilibrium state when all the virus has become resistant to the drugs being administered (i.e. when $u_0 = 0$ after a very long time). The former state (with $u_1 = 0$) is unstable in a patient under treatment, because the virus is slowly developing resistance, i.e., u_0 is slowly turning into u_1. As for the third state (the equilibrium state with $u_0 = 0$), if this state is stable, we have a rebound, otherwise we have SVR.

2 The Model

2.1 Set up

We take one day as the unit of time and write

$$F_1(x_1, u_0, u_1, y_1) = A_1 - A_2 x_1 - (1 - e_1) x_1 (A_3 u_0 + A_6 u_1) \tag{1}$$

$$\begin{aligned} F_2(x_1, u_0, u_1, y_1) = {} & A_9 A_4 (1 - e_1 e_2)[(1 - Q_1) p_0 x_1 u_0 + Q_2 p_1 x_1 u_1] \\ & - c_1 u_0 - c_3 y_1 u_0 \end{aligned} \tag{2}$$

$$\begin{aligned} F_3(x_1, u_0, u_1, y_1) = {} & A_9 A_4 (1 - e_1 e_2)[Q_1 p_o x_1 u_0 + (1 - Q_2) p_1 x_1 u_1] \\ & - c_1 u_1 - c_3 y_1 u_1 \end{aligned} \tag{3}$$

$$F_4(x_1, u_0, u_1, y_1) = A_{10}(u_0 + u_1) - c_2 (u_0 + u_1) y_1 \tag{4}$$

with $x_1' = F_1$, $u_0' = F_2$, $u_1' = F_3$ and $y_1' = F_4$;

In these equations, the quantity x_1 stands for the number of susceptible cells in one unit of volume, which cells are attacked by the viruses u_0 and u_1 at the rates A_3

and A_6 respectively. Generally $A_3 > A_6$ because of the higher fitness of the 'wild-type' virus u_0. Since HCV is not (or is only mildly) cytopathic, we have not included a separate equation for the infected cells as is done in most models of HIV where the virus is highly cytopathic. It should be noted that half-life of infected cells is not very different from those of susceptible cells and they may also multiply like the susceptible cells. The life cycle of infected cells is, therefore, very much like that of susceptible cells. The infected cells produce both wild type and resistant virions at the rates p_0 and p_1 respectively. Since, eventually, most of the virions produced will be of the resistant type, we assume that $p_1 > p_0$. The antibodies y_1 are produced in response to the presence of both u_0 and u_1 at the rate A_{10} $(u_0 + u_1)$ and are neutralized at the rate $c_2(u_0 + u_1)y_1$ as and when they encounter a virus. The parameter $A_4 < A_3$ accounts for a protease inhibitor.

The value of A_1, the rate at which the susceptible cells are being created, has been estimated at anywhere from one to 180,000/mL in the literature while the value of A_1/A_2, the equilibrium value of total number of cells, has been estimated to be anywhere from 4 million to 13 million cells/mL [7]. We take this value to be one million in an (appropriate) one unit of volume.

2.2 Equilibrium Points

We begin to analyze our system by taking an example. We take the equilibrium value of y_1 as (obviously) A_{10}/c_2. We assume that

$A_1 = 10$; $A_2 = A_1/1{,}000{,}000$; $A_3 = 0.00000001$; $A_4 = A_3$; $A_5 = 0.005$; $A_6 = 0.5A_3$; $A_9 = 1000$; $A_{10} = 0.1$; $Q_1 = 0.0000001$; $Q_2 = Q_1{}^*Q_1$; $p_0 = 0.9$; $p_1 = 0.99$; $e_1 = 0.1$; $e_2 = 0.9$; $c_1 = 8$; $c_2 = 0.0000001$; $c_3 = 0.0000001$; and solve our system numerically (on Mathematica 8.0). The result is the three points $\{(x_1, u_0, u_1) = (1{,}000{,}000, 0, 0), (989{,}011, 12.3456, 0), (899{,}101, 0, 249.383)\}$

Apart from the disease free solution, there are two other solutions. Notice that $u_0 = 0$ in one solution and $u_1 = 0$ in the other solution. What is happening? To see this, we solve the same system with the same values of parameters as above, but with $e_1 = e_2 = Q_1 = Q_2 = 0$, i.e. without any treatment. The result is the three points $\{(x_1, u_0, u_1) = (1{,}000{,}000, 0, 0), (900{,}000.0, 111.111, 0), (818{,}181.8, 0, 444.444)\}$.

The relevant solution without any treatment is the one with $u_1 = 0$, which has $x_1 = 900{,}000$; and $u_0 = 111.111$. This is the so called chronic equilibrium point. It is intuitively clear that as the treatment starts, the number of virions should come down, and the body may reach another (unstable) equilibrium point where the treatment has reduced the virus count but no resistance has developed yet. Later on resistance may develop, which will result in a rebound. This is exactly what happens in our model. As the treatment starts, the number of healthy cells goes up and the system (i.e. the body) reaches another equilibrium point when the number of healthy cells has gone up (as expected, from 900,000 to 989,011), and the number of virions has come down (as expected, from 111.111 to 12.3456), but no resistance has developed yet. As the treatment continues, the resistance slowly develops, and

we reach the next equilibrium point, where the number of healthy cells has gone down (as expected, from 989,011 to 899,101), all the virions have changed to the resistant type (as expected), and the number of resistant virions has gone up (as expected, to 249.383).

Alternatively, we may calculate the two equilibrium points (one with $u_0 = 0$ and the other with $u_1 = 0$) by **assuming** that $u_0 = 0$ for one point and that $u_1 = 0$ for the other point. For $u_1 = 0$, the result is

$$\left\{(x_1, u_0) = \left(\frac{A_1}{A_2}, 0\right), \left(\frac{c_1c_2 + A_{10}c_3}{A_4A_9c_2(e_1e_2 - 1)p_2(Q_1 - 1)}, \frac{A_2(c_1c_2 + A_{10}c_3) - A_1A_4A_9c_2(e_1e_2 - 1)p_0(Q_1 - 1)}{A_3(c_1c_2 + A_{10}c_3)(e_1 - 1)}\right)\right\} \tag{5}$$

and for $u_0 = 0$, the result is

$$\left\{(x_1, u_1) = \left(\frac{A_1}{A_2}, 0\right), \left(\frac{c_1c_2 + A_{10}c_3}{A_4A_9c_2(e_1e_2 - 1)p_1(Q_2 - 1)}, \frac{A_2(c_1c_2 + A_{10}c_3) - A_1A_4A_9c_2(e_1e_2 - 1)p_1(Q_2 - 1)}{A_6(c_1c_2 + A_{10}c_3)(e_1 - 1)}\right)\right\} \tag{6}$$

The former equilibrium point (other than the disease free solution) with $u_1 = 0$ is seen to be unstable, if p_1 is sufficiently large compared to p_0.

For $e_1 = e_2 = Q_1 = Q_2 = 0$, (i.e. without any treatment), the value of u_0 (with $u_1 = 0$) turns out to be

$$u_0 = \frac{-A_2(c_1c_2 + A_{10}c_3) + A_1A_4A_9c_2p_0}{A_3(c_1c_2 + A_{10}c_3)}.$$

We write

$$R_0 = \frac{A_1A_4A_9c_2p_0}{A_2(c_1c_2 + A_{10}c_3)}.$$

If $R_0 < 1$, the corresponding value of u_0 is less than zero, and consequently, the chronic state will not develop. It follows that if

$$A_{10} > \frac{A_1A_4A_9c_2p_0 - A_2c_1c_2}{A_2c_3}, \tag{7}$$

the chronic state will not develop and the infection does not take hold. We speculate that this is the reason why a large number of people do not proceed to a chronic

state and self cure after being infected with HCV. Their adaptive immunity is just too strong. For the values of the parameters assumed above, we get the critical value of A_{10} as 1.0.

Noting that, with the values of the parameters as assumed above, (x_1, u_1) = (818,182, 444.444) which is what (6) gives and (x_1, u_0) = (900,000, 111.111) which is what (5) gives, we notice that these values coincide with those calculated directly (i.e. without the additional assumption that $u_0 = 0$ in one solution and $u_1 = 0$ in the other).

2.3 Positivity of the Solution

It is obvious that if a solution starts in $\{x_1, u_0, u_1, y_1\} > \{0, 0, 0, 0\}$, then it stays in that region. This is because at $x_1 = 0$, $x_1' > 0$. A similar argument applies to the other variables.

2.4 Boundedness of the Solution

The quantity x_1 is clearly positive and bounded by A_1/A_2. Also y_1 is positive and bounded by A_{10}/c_2. We have also assumed that $p_0 < p_1 < 1$. It follows that

$$\begin{aligned}(u_0+u_1)' &= -c_1(u_0+u_1) + A_4A_9(1-e_1e_2)(p_0u_0+p_1u_1)x_1 \\ &\quad - c_3(u_0+u_1)y_1 \\ &< (u_0+u_1)[A_4A_9(1-e_1e_2)p_1x_1 - c_1 - c_3y_1]\end{aligned} \tag{8}$$

For given values of u_0 and u_1, the equilibrium value of x_1 is

$$\frac{A_1}{A_2 + (1-e_1)(A_3u_0 + A_6u_1)},$$

which value is clearly a maximum. Also, this value is less than (or equal to)

$$\frac{A_1}{A_2 + (1-e_1)(u_0+u_1)A_6},$$

so that we get

$$(u_0+u_1)' < \frac{(u_0+u_1)A_1A_4A_9(1-e_1e_2)p_1}{[A_2 + (1-e_1)A_6(u_0+u_1)] - c_1 - c_3y_1}. \tag{9}$$

It follows that, for large enough values of $u_0 + u_1$, $(u_0 + u_1)'$ is negative for arbitrarily small values of c_1 and c_3.

This proves the boundedness of the solutions of our system.

2.5 Treatment Outcomes

Since the equilibrium point with $u_1 = 0$ is unstable, there are only two possible outcomes of treatment. They are either the rebound, i.e. the solution with $u_0 = 0$ obtained above, or SVR, the disease free solution.

We write

$$R_3 = \frac{A_1 A_4 A_9 c_2 p_1 (1 - e_1 e_2)(1 - Q_2)}{A_2 (c_1 c_2 + A_{10} c_3)}.$$

If $R_3 > 1$, we have REBOUND, otherwise we have SVR.

This says that if

$$A_{10} > \frac{c_2[-A_2 c_1 + A_1 A_4 A_9 p_1 (1 - e_1 e_2)(1 - Q_2)}{A_2 c_3},$$

we will have an SVR. Notice that the right hand side may be made as small as we like depending upon the parameters that define the treatment, so that for a sufficiently strong treatment, we should have SVR. But of course, we must be mindful of the side effects of a strong treatment. We also recall that if statement (8) holds, then the chronic state will not develop. All these results are in line with what actually happens in a real situation.

2.6 Examples

We shall now give some examples to illustrate our model. We take $A_1 = 10$; $A_2 = A_1/1{,}000{,}000$; $A_3 = 0.00000001$; $A_4 = A_3$; $A_5 = 0.005$; $A_6 = 0.5A_3$; $A_9 = 1000$; $Q_1 = 0.0000001$; $Q_2 = Q_1{*}Q_1$; $p_0 = 0.9$; $p_1 = 0.99$; $c_1 = 8$; $e_1 = 0.1$; $e_2 = 0.9$; $c_2 = 0.0000001$; $c_3 = 0.0000001$; and solve our model on Mathematica 8.0 for several values of A_{10}. The results are given in Figs. 1 and 2 where the values of $\mathrm{Log}_{10}[u_0(t) + u_1(t)]$ are indicated along the vertical axis against time (in days).

Decay of Virus:

It should be noted that in our examples, the virus decays in a bi-phasic manner over the first forty days. In actual studies, this decay of virus is noted to take place in tri-phasic ways, with virus coming down significantly during the first few days. A lot of attention has been paid in the literature to explain this tri-phasic delay. According to one opinion, "In such studies, the first phase is assumed to be an initial sharp decay related to the antiviral 'efficacy' of IFN in clearing of free virus by blocking viral production and secretion which occurred after a delay of about 8–9 h from the beginning of therapy. The second decay phase showed a more

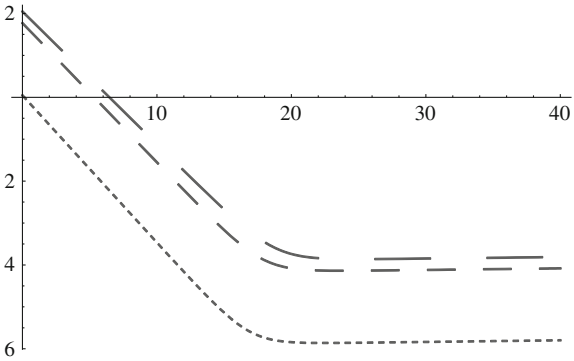


Fig. 1 The behavior of our model, showing drop in virus count with treatment, for several values of A_{10} for the first forty days **(a)** $A_{10} = 0.1$ (*long dashes*), **(b)** $A_{10} = 0.5$ (*median dashes*) **(c)** $A_{10} = 0.99$ (*short dashes*). Notice that higher values of A_{10}, the antibody production rate, result in lower values of u_0, the initial virus count at the chronic equilibrium point

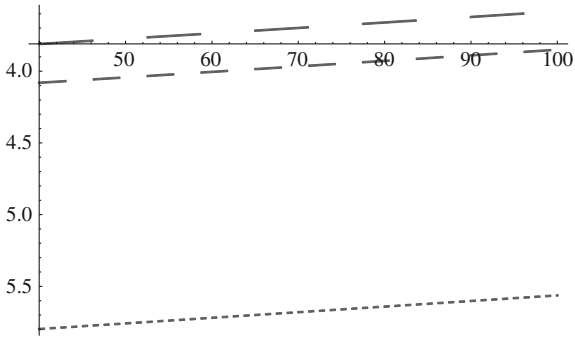


Fig. 2 The same case as in Fig. 1, showing the rebound of virus in each case during the next few weeks (from 40 to 100 days)

gradual slope in HCV RNA levels, thus representing the rate of killing, clearance of virally infected cells while the third phase of viral decay may be attributed to the effect of RBV that may be related to restoration of a previously suppressed cellular immune response." [8]. According to another author, "the slope of the "shoulder phase" in patients with tri phasic viral decay represents the pre-treatment death rate of infected cells and the third-phase slope represents the treatment-enhanced death rate of infected cells due to the immune modulatory effect of RBV." [9]. The 'shoulder phase' refers to the second phase of tri-phasic decay. If the half life of infected cells is reasonably long, these explanations appear suspicious.

We argue that this tri-phasic decay may happen because of decay in the effectiveness of the drug during the first few days. We change the values of e_1 in our model with time, and show the results as $\text{Log}_{10}[u_0(t) + u_1(t)]$. Figure 3 illustrates a situation where the effect of the drug is very high in the beginning and is maintained at a low level later on.

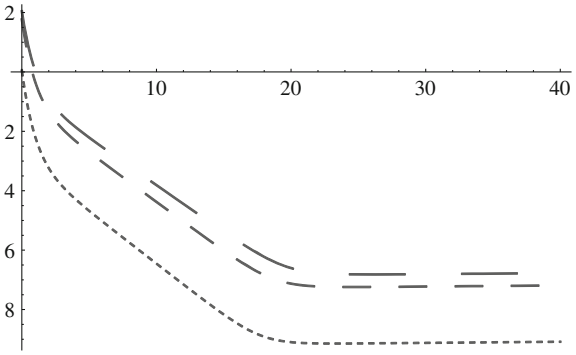


Fig. 3 Decay of virus when e_1 changes as $e_1(t) = 0.1 + 0.85e^{-t}$. All other parameters are the same as in Fig. 1. Tri-phasic decay of virus is clearly visible in all cases of $A_{10} = 0.1$(*long dashes*), 0.5 (*median dashes*), and 0.99 (*short dashes*)

2.7 Number of Virions

It is to be noted that the number of virions in the chronic state in our model is of the order of a few hundreds. In actual cases, this number is in millions. We introduce an appropriate scaling factor A_{11} in our model and rewrite it as

$$F_1(x_1, u_0, u_1, y_1) = A_1 - A_2x_1 - A_{11}(1 - e_1)x_1(A_3u_0 + A_6u_1) \tag{10}$$

$$\begin{aligned} F_2(x_1, u_0, u_1, y_1) = {} & A_9A_4(1 - e_1e_2)[(1 - Q_1)p_0x_1u_0 + Q_2p_1x_1u_1] \\ & - c_1u_0 - A_{11}c_3y_1u_0 \end{aligned} \tag{11}$$

$$\begin{aligned} F_3(x_1, u_0, u_1, y_1) = {} & A_9A_4(1 - e_1e_2)[Q_1p_ox_1u_0 + (1 - Q_2)p_1x_1u_1] \\ & - c_1u_1 - A_{11}c_3y_1u_1 \end{aligned} \tag{12}$$

$$F_4(x_1, u_0, u_1, y_1) = A_{10}(u_0 + u_1) - A_{11}c_2(u_0 + u_1)y_1 \tag{13}$$

with $x_1' = F_1$, $u_0' = F_2$, $u_1' = F_3$ and $y_1' = F_4$;

The solution of this model for $A_1 = 1000$; $A_2 = A_1/1{,}000{,}000$; $A_3 = 0.00000001$; $A_4 = 0.999A_3$; $A_5 = 0.005$; $A_6 = 0.5A_3$; $A_9 = 1000$; $Q_1 = 0.0000001$; $Q_2 = Q_1{*}Q_1$; $p_0 = 0.9$; $p_1 = 0.99$; $c_1 = 8$; $c_2 = 0.0000001$; $c_3 = 0.0000001$; $A_{10} = 0.1$; $e_1(t) = 0.1 + 0.85e^{-t}$; $e_2 = 0.9$; $A_{11} = 0.00001$ is shown in the Fig. 4, once again as $\mathrm{Log}_{10}[u_0(t) + u_1(t)]$. The curve closely follows the actual readings of a patient reported by Reluga et al. [7] over a span of 14 days.

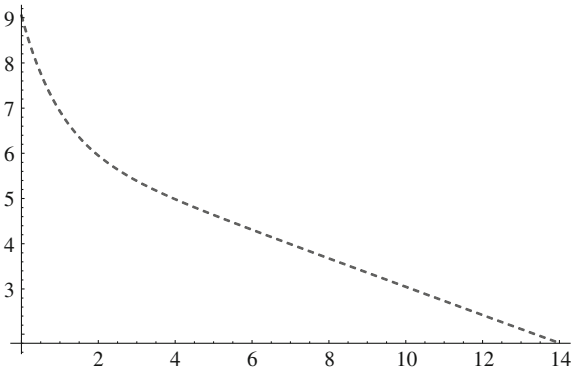


Fig. 4 Decay of virions in a particular case for the first 14 days. The *curve* closely follows the actual readings of a patient reported in the literature

2.8 Robustness of the System

Extending our previous discussion [10], we tried to see, whether our results were peculiar to the values of the various parameters that we happen to have chosen for our system. Since, the system depends on a large number of parameters, therefore, as preliminary measure, we tried to change the values of only one of them and see the effect of such a change. We chose the important parameter A_9 which gives the number of virions produced by an infected cell. This number was taken to be 1000 in our examples. We asked the computer twice to choose this number randomly anywhere between 1000 and 2000 and tested the effect of this change on our results. The numbers chosen by the computer were 1168 and 1214. The results for both these choices were very similar. We reproduce the results for A_9 = 1214 in this paper (Figs. 5 and 6).

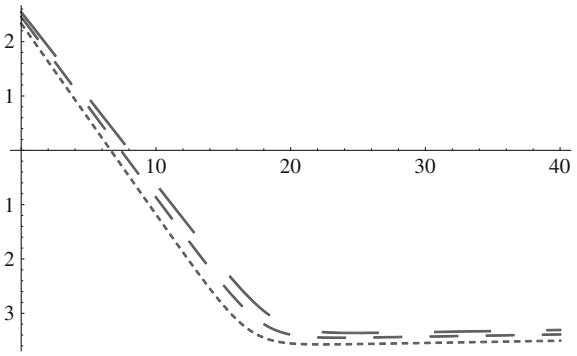


Fig. 5 Same as Fig. 1 except that A_9 has been changed to 1214

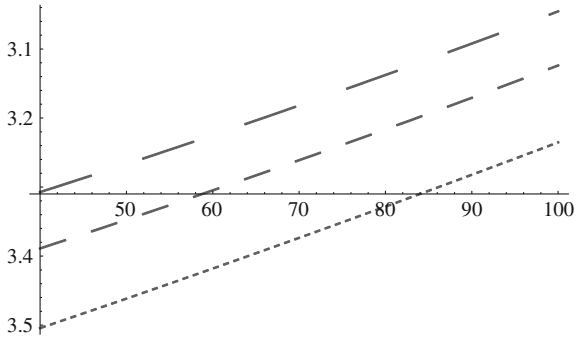


Fig. 6 Same as Fig. 2, except that A_9 has been changed to 1214. Notice the rebound as before

2.9 Sensitivity of the System

As against the robustness of our system to changes in the parameters like A_9, we noticed that our system is very sensitive to changes in the treatment variables like e_1 and e_2. We report this sensitivity here. As our Fig. 7 indicates, the rebound approach in our system changes to SVR as we change the value of e_1 from 0.1 to 0.11. This should be of some interest to medical practitioners (Fig. 8).

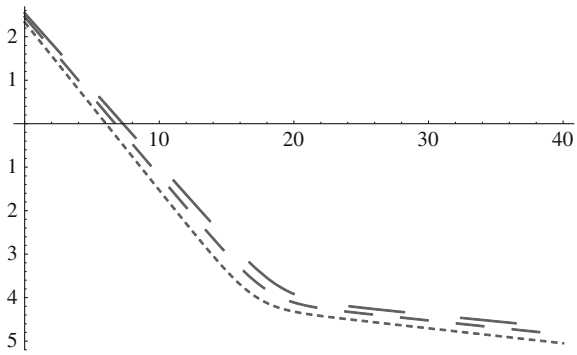


Fig. 7 Same as Fig. 5 except that e_1 has been changed to 0.11. As the *slope of the curves* near $t = 40$ indicates, the system is going towards SVR

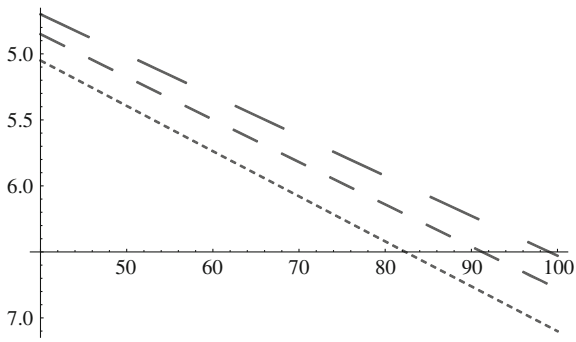


Fig. 8 Same as Fig. 6 except that e_1 has been changed to 0.11. Notice that the system is now going to SVR as against rebound in Fig. 6

Acknowledgment I wish to acknowledge the immense help of Dr. Rita Aggarwala in preparing this paper. This paper could not have been written without her help.

References

1. Dixit NM (2008) Advances in the mathematical modeling of Hepatitis C virus dynamics. J Indian Inst Sci 88:37–43
2. Chak E, Talal AH, Sherman KE, Schiff ER, Saab S (2011) Hepatitis C virus infection in USA: an estimate of true prevalence. Liver Int
3. Reddy KR, Nelson DR, Zeuzem S (2009) Ribavirin: current role in the optimal clinical management of chronic Hepatitis C. J Hepatol 50(2):402–411
4. Chen SL, Morgan TR (2006) The natural history of Hepatitis C virus (HCV) infection. Int J Med Sci 3(2)
5. Forton DM, Karayiannis P, Mahmud N, Taylor-Robinson SD, Thomas HC (2004) Identification of unique Hepatitis C virus quasispecies in the central nervous system and comparative analysis of internal translational efficiency of brain, liver, and serum variants. J Virol 78(10):5170–5183
6. Rong L, Dahari LH, Rebeiro R, Perelson AS (2010) Rapid emergence of protease inhibitor resistance in Hepatitis C virus. Sci Transl Med 2(30):30–32
7. Reluga T, Dahari H, Perelson AS (2009) Analysis of hepatitis C virus infection models with hepatocyte homeostatis. SIAM J Appl Math 69(4):999–1023
8. Ballesteros AL, Fuster D, Planas R, Clotet B, Tural C (2005) Role of viral kinetics under HCV therapy in HIV/HCV-coinfected patients. J Antimicrob Chemother 55:824–827
9. Dahari H, Rebero RM, Perelson S (2007) Triphasic decline of Hepatitis C virus RNA during antiviral therapy. Hepatology 46(1)
10. Aggarwala BD (2014) Effect of Antibodies on HCV Infection.In: Proceedings of the world congress on engineering and computer science 2014, WCECS, San Francisco, U.S.A., 22–24 October 2014 (Lecture notes in engineering and computer science), pp 771–775

Gas Transport Through Inorganic Ceramic Membrane and Cation-Exchange Resins Characterization for Ethyl Lactate Separation

Edidiong Okon, Habiba Shehu and Edward Gobina

Abstract Ethyl lactate is an important organic ester, which is biodegradable in nature and widely used as food additive, perfumery, flavor chemicals and solvent. Inorganic porous ceramic membrane has shown a lot of advantages in the equilibrium process of ethyl lactate separation. In this work, the transport characteristic of carrier gas including Nitrogen (N_2), Helium (He), Argon (Ar) and Carbon dioxide (CO_2), with α-Al_2O_3 inorganic ceramic membrane used for ethyl lactate separation was investigated, at the pressure drop of 0.01–0.09 bar and 298 K. The carrier gas flow rate was molecular weight dependent in the order: He > Ar > N_2 > CO_2 with respect to pressure drop. The membrane pore size distribution was analysed using Scanning electron microscope coupled with energy dispersive x-ray analyser (SEM-EDXA). The SEM surface of the commercial cation-exchange resin catalysts before esterification showed a defect-free surface.

Keywords Carrier gas · Carrier gas flux · Characterization · Cation-exchange resin · Esterification · Ethanol · Ethyl lactate · Inorganic membrane · Kinetic diameter · Permeability

Nomenclature

Symbols

A	Surface area of the membrane (m^2)
D	Diffusivity ($m^2\ s^{-1}$)
F	Permeability (mol $mm^{-2}\ s^{-1}\ Pa^{-1}$)
J	Flux (mol $s^{-1}\ m^{-2}$)

E. Okon · H. Shehu · E. Gobina (✉)
Centre for Integration and Membrane Technology (CPIMT), School of Engineering, The Robert Gordon University Aberdeen, Aberdeen AB10 7GJ, UK
e-mail: e.gobina@rgu.ac.uk

E. Okon
e-mail: e.p.okon@rgu.ac.uk

H. Shehu
e-mail: h.shehu@rgu.ac.uk

H.K. Kim et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-94-017-7236-5_29

$\bar{P}$ Permeance (mol m^{-2} s^{-1} Pa^{-1})
Knv Intrinsic permeability corresponding to Knudsen flow (m)
Kv The intrinsic permeability corresponding to viscous flow (m^2)
M Gas molecular weight (g/mol)
Pm Mean pressure (bar)
Q Gas flow rate (mol s^{-1})
R Gas molar constant (8.314 J mol^{-1} K^{-1})
S Solubility (mol m^{-3} Pa^{-1})
T Temperature (Kelvin)
ΔP Pressure drop

Greek Symbols
Å Angstrom
α Constant representing viscous flow (mol m^{-2} s^{-1})
β Constant representing Knudsen flow (m^{-2} s^{-1})
μ Viscosity (Pa^{-1} s)
$\bar{v}$ Mean molecular velocity (Pa s^{-1})
λ Mean free path (m)
r_p Membrane Pore radius (m)
δ Membrane thickness (m)

1 Introduction

Ethyl lactate is a biodegradable and non-toxic material with an excellent solvent property which could potentially replace halogenated and toxic solvents for a broad range of consumer and industrial uses, corresponding up to 80 % of worldwide solvent consumption [1]. It can also be used in the pharmaceutical industry as a dispersing/dissolving excipient for several biological compounds without destroying the pharmacological activity of the active ingredient [2]. Ethyl lactate can replace environmentally damaging solvents including toluene, acetone, N-methyl pyrrolidone and xylene [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]. Currently some studies have focused on the water-permeable membrane reactor which has to do with liquid-phase reversible reactions including esterification reactions [3]. Among the membranes considered, inorganic membranes have been found to be the perfect membrane for the esterification reaction process because they can allow heterogeneous catalysts to be deposited easily on the surface of the membrane; this results in an increase in the purity of products since side reactions and corrosion problems can be avoided [4].

Esterification reactions are usually limited by equilibrium and therefore do not reach completion. However, using a membrane can result in higher conversion by shifting the chemical equilibrium towards the formation of the product by the removal of water from the reaction mixture as well as driving the reaction towards completion [5]. Membrane-based separation technologies have shown a wide range of application in food, biotechnology, pharmaceutical and in the treatment of other industrial effluents [6]. Membrane can be classified into two groups including inorganic and organic membranes [7].

Inorganic membranes have attracted important attention in different fields including academic and industry [8]. These membranes can be prepared using different methods including sol-gel, chemical vapour and sintering processes [7]. However, inorganic ceramic membrane composed of two materials such as aluminum oxide (Al_2O_3) and zirconium dioxide (ZrO_2), other materials include titanium dioxide (TiO_2) and silicon dioxide (SiO_2) [7]. The transport behavior of gases through inorganic ceramic membrane can be explained using various mechanisms such as viscous flow, Knudsen diffusion, surface diffusion and molecular sieving mechanism.

In Knudsen diffusion mechanism, gas molecules diffuse through the pores of the membrane and then get transported by colliding more frequently with the pore walls [9]. Gas separation by molecular sieving mechanism takes place when the pore dimensions of the inorganic ceramic membrane approach those of the permeating gas molecules. In capillary condensation mechanism, separation can takes place in the pores of the membrane with mesoporous layer in the presence of condensable gas specie. Surface diffusion mechanism occurs when the adsorption of the permeating gas molecule adsorb on the pore surface of the membrane material there by increasing the gas transport performance. Viscous flow mechanism takes place if the pore radius of the membrane is larger than the mean free path of the permeating gas molecule [9]. Gas permeability can also be described in terms of solution-diffusion mechanism, i.e.

$$\text{Permeability(F)} = \text{solubility(S)} \times \text{diffusivity(D)} \tag{1}$$

The diffusivity explains the rate at which gases move across the membrane and solubility describes the interaction between the membrane surface and the permeating gas molecule [7]. The transport of gases across the pore space of membrane has been a subject of numerous studies in the development of separation process involving membrane [10].

2 Experimental

Figure 1 shows a schematic diagram of a simple gas permeation setup. The gas transport through a porous inorganic ceramic membrane was performed using single gases including carbon dioxide (CO_2) helium (He), nitrogen (N_2) and argon (Ar) at different gauge pressure (bar) and room temperature (298 K).

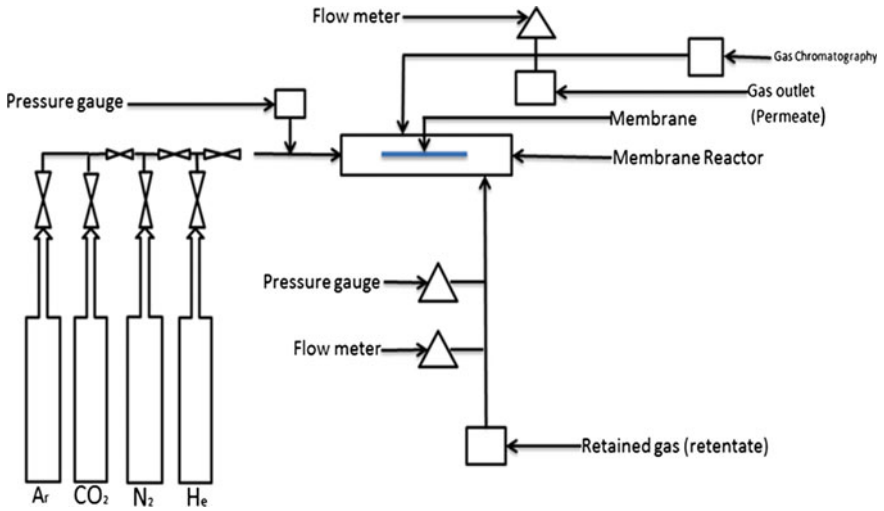


Fig. 1 Schematic diagram of gas permeation setup

The gas transport was performed at room temperature of 298 K between the pressure drops of 0.01–0.09 bar. The flow meter (cole Parmer model) was used to determine the gas flow rate. The inner and outer radius of the membrane was 7 and 10 mm respectively, whereas the effective length of the membrane was measured to be 36.6 cm. The membrane was prepared using a similar procedure as that proposed by Gobina [11]. The characterisation of the membrane pore size distribution was examined using SEM-EDXA to determine the morphology and the elemental composition of the membrane support before and after modification. The characterization of the resin catalysts was also carried out to determine the surface morphology of the resins before esterification reaction. Amberlyst 16, Amberlyst 15, Amberlyst 36 and Dowex 50W8x (obtained from Sigma Aldrich) were used as the resin catalysts.

3 Results and Discussion

The gas kinetic diameter was plotted against permeance to determine the flow mechanism of the single gases. Table 1 shows the different gases with their respective kinetic diameter (Å). The result obtained in Fig. 2 showed that helium gas with the lowest kinetic diameter recorded the highest permeability. If the membrane had any molecular sieving properties, then CO_2 would have been next to He. However, N_2 with the higher kinetic diameter exhibited a higher permeance than CO_2 indicating that the gas transport through the membrane was not based on molecular sieving mechanism, but there could be another flow mechanism that was

Table 1 Gas molecular weight and their respective kinetic diameter

Gases	Molecular weight (g/mol)	Kinetic diameter (Å)
Helium (He)	4	2.60
Argon (Ar)	40	3.43
Nitrogen (N_2)	28	3.64
Carbon dioxide (CO_2)	44	3.30

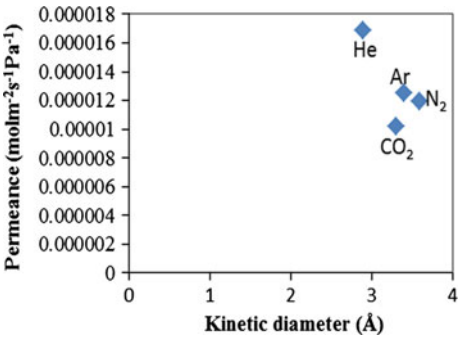


Fig. 2 Effect of kinetic diameter with the gas permeance at 0.03 bar gauge pressure

responsible for the flow of these gases. The order of the gas kinetic diameter is represented as N_2 > Ar > CO_2 > He as shown in Table 1.

It can be seen from Fig. 3a–d that the gas flux through the porous membrane increases linearly with gauge pressure at 298 K for all the gases. These results corroborate with a similar results by Tomita et al. [12]. The order of the gas molecular weight was represented as CO_2 (44) > Ar (40) > N_2 (28) > He (4). Helium gas has the lowest molecular weight but exhibited a higher flux with the gradient of 0.0123 mol s^{-1} m^{-2} bar^{-1} whereas CO_2 with the higher molecular weight exhibited a low flux with a low gradient of 0.0056 mol s^{-1} m^{-2} bar^{-1}. Although N_2 and Argon have different molecular weights, their respective flux and gradient were close. These also suggest that Knudsen mechanism of transport contributes to the gas flow through the porous ceramic membrane since this mechanism has a relationship with the gas molecular weight.

A linear equation was obtained for all the gases with the gradient of the graph being less than 1. R^2 values indicating good fit of data were obtained for the gases. The gas flux was calculated using the following equation:

$$J = \frac{Q}{A} \tag{2}$$

where J = flux (mol m^{-2} s^{-1}), Q = flow rate of the gases (mol s^{-1}), A = membrane surface area (m^2) and the gas permeance was obtained using the following equation:

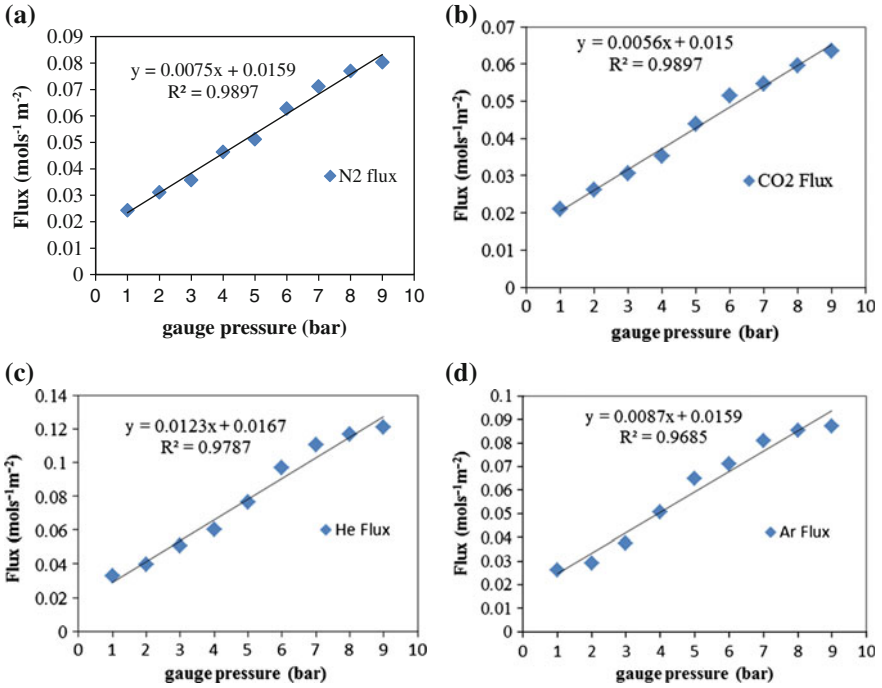


Fig. 3 **a** N_2 gas flux (mol m^{-2} s^{-1}) against gauge pressure (bar). **b** CO_2 gas flux (mol m^{-2} s^{-1}) against gauge pressure (bar). **c** He gas flux (mol m^{-2} s^{-1}) against gauge pressure (bar). **d** Ar gas flux (mol m^{-2} s^{-1}) against gauge pressure (bar)

$$\overline{P} = \frac{J}{\Delta P} \tag{3}$$

where ΔP is the pressure drop across the membrane (bar), J = flux (mol m^{-2} s^{-1}) and $\bar{P}$ is the permeance (mol m^{-2} s^{-1} Pa^{-1}).

Figure 4 shows the plot of permeance against the inverse of the square root of the gas molecular weight at 0.1 bar. The graph obtained was not a straight line graph as expected for Knudsen flow mechanism.

The flow rate was also plotted against the gauge pressure to further observe the flow mechanism. It can be seen from Fig. 5 that the gas flow rate increases with respect to gauge pressure and follows an 'S' curve. He gas showed the highest permeance at each gauge pressure compared to other gases. Which means; He gas was more permeable through the ceramic membrane compared to other gases. Considering the gas molecular weight as shown in Table 2, the result obtained showed that at higher gauge pressure, the gas flow rate was molecular weight dependent in the order; He > Ar > N_2 > CO_2 with respect to pressure, indicating Knudsen mechanism of transport at these higher gauge pressures.

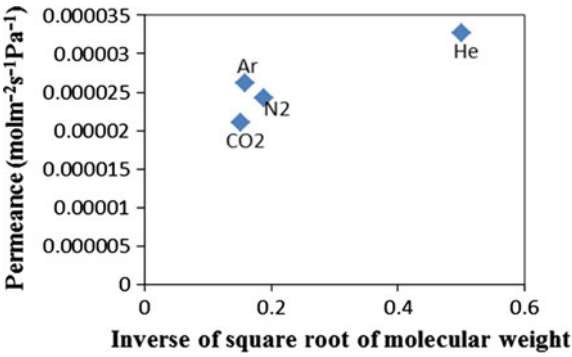


Fig. 4 Gas permeance (mol m^{-2} s^{-1} Pa^{-1}) against Inverse of square root of the gas molecular weight at 0.1 bar

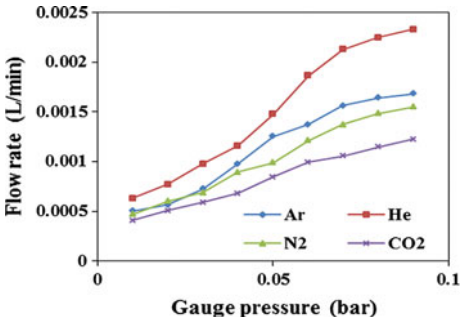


Fig. 5 Flow rate (L/min) of Ar, He, N_2 and CO_2 against gauge pressure (bar)

Table 2 Membrane pore radius (m) and mean free path (m) for the gases

Gas molecule	Mean free path (λ) m	Pore radius (m)
Ar	3.15E−04	4.72E−12
He	3.63E−04	1.09E−11
N_2	2.96E−04	4.45E−12
CO_2	1.11E−04	2.22E−12

The permeability of the gas through the membrane can be obtained from the equation:

$$F = \frac{J * \delta}{\Delta P} \tag{4}$$

where δ is the membrane thickness (m), F = permeability (mol m s^{-1} m^{-2} $Pa^{-1)}$, J = flux (mol m^{-2} s^{-1}) and ΔP is the change in pressure (bar).

The gas permeability (mol m s^{-1} m^{-2} Pa^{-1}) was plotted against the mean pressure (bar). From the results obtained for the straight line in Fig. 6, it was assumed that viscous flow was very low or approximately zero and as such viscous flow was neglected. However, Knudsen flow seems to be valid which indicate Knudsen

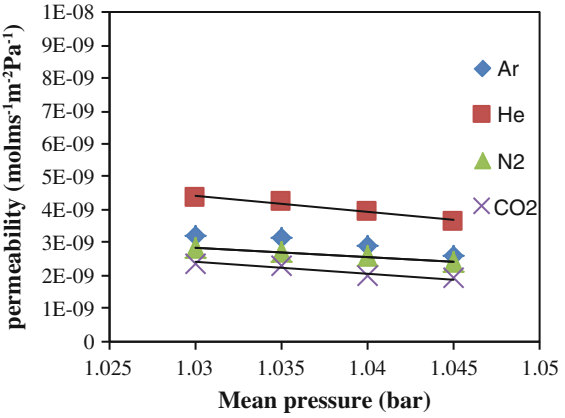


Fig. 6 Gas Permeability (mol m s^{-1} m^{-2} Pa^{-1}) against mean pressure (bar)

mechanism of transport. These results corroborate with a similar result by Julian et al. [13] but in their case, Knudsen number was very high and as such was neglected.

The straight line equation from the graph shown in Fig. 6 is represented as:

$$F = \alpha + \beta \cdot Pm \tag{5}$$

where F = permeability (mol m m^{-2} s^{-1} Pa^{-1}), α = constant representing viscous flow, β = constant representing Knudsen flow and P_m = mean pressure (bar) [14]. The constant representing Knudsen and viscous flow can be further calculated using the following equation [13]:

$$\alpha = \frac{k_v}{R \cdot T \cdot \mu \cdot L} \tag{6}$$

where α = constant representing viscous flow (mol m^2 s^{-1}), K_v = the intrinsic permeability corresponding to viscous flow (m^2), R = gas molar constant (J mol^{-1} K^{-1}), T = temperature (K), μ = gas viscosity (Pa.s), δ = membrane thickness (m)

$$\beta = \frac{4\nabla}{3 \cdot R \cdot T \cdot L} K_{nv} \tag{7}$$

where β = constant representing Knudsen flow (m^2/s), $\overline{v}$ = mean molecular velocity (Pa s^{-1}), T = temperature (K), R = gas molar constant (J mol^{-1} K^{-1}), δ = membrane thickness (m), K_{nv} = intrinsic permeability corresponding to Knudsen flow (m).

The membrane pore radius and the mean free path with the gases were also determined. From Table 2, the results obtained showed that the membrane pore radius for the gas was all smaller than the mean free path, indicating Knudsen mechanism of transport. These results corroborate with a report by Benito et al. [14], Pandey and Chauhan [15]. According to these authors, Knudsen diffusion is the dominant mechanism if the membrane pore radius is smaller than the mean free

path of the molecules and this is also significant for membrane with small pore (radius <10 nm) for a free-defect membrane. The results showed that the membrane pore radius was less than 10 nm indicating a free-defect membrane and Knudsen flow as the dominant mechanism of transport.

The membrane pore radius was calculated using the formula [13]:

$$\tau\rho = \frac{16 \cdot A \cdot \mu}{3 \cdot B} \sqrt{\frac{8RT}{\pi \mathrm{M}}} \tag{8}$$

where r_p = membrane pore radius (m), α = constant representing viscous flow from the permeability graph, β = constant representing Knudsen flow from the permeability graph, μ = gas viscosity (Pa.s), M = gas molecular weight (g/mol), $\pi = 3.141$.

4 Morphology of Membrane and Cation-Exchange Resin Catalysts

The SEM image of the modified membrane was also obtained as well as the EDXA spectra of the membrane. The image was focused at 200 μm. Figure 7 shows the surface image of the modified α-Al_2O_3 ceramic membrane. The pore of the

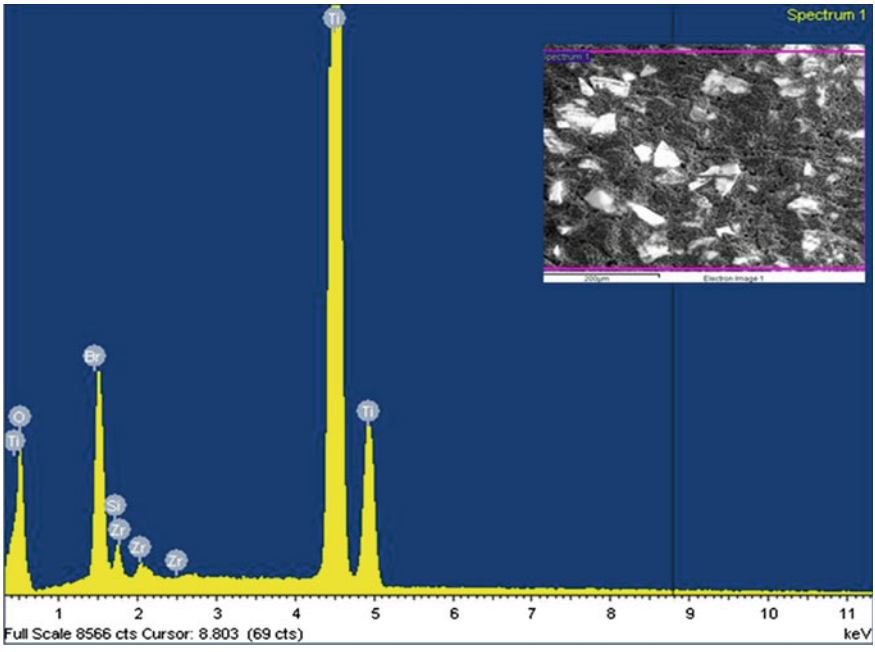


Fig. 7 SEM images and the EDXA spectra of the modified α-Al_2O_3 support

membrane was found to reduce after modification [16]. This could indicate the effect of support on the pores of the membrane. This result corroborates with a similar study by Tomita et al. [12]. The EDXA results obtained showed that the elemental composition of the modified membrane consists of elements such as silicon (Si), titanium (Ti), bromine (Br), oxygen (O) and zirconium (Zi). However, Ti and O showed a higher concentration. This could indicate that the membrane support was initially coated with TiO_3 and subsequently coated with SiO_2 and ZrO_2.

The SEM of the cation-exchange resin catalysts was also analysed before employing the catalysts in esterification reactions. The different cation-exchange resin used for the analysis include: Amberlyst 16, Amberlyst 15, Amberlyst 36 and Dowex 50W8x. The resin catalysts were analysed at the magnification of 100X and the scale of 10 µm. It can be seen that the surfaces of Amberlyst 16 (8a), Amberlyst 36 (8b) and Dowex 50xw8 (8c) 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. 8d, it was also observed that Amberlyst 15 exhibited small cracks on the surface. This was suggested to be the effect of sulfonic acid group in which the structure of the solid catalyst is made up of thereby indicating a strong catalytic effect.

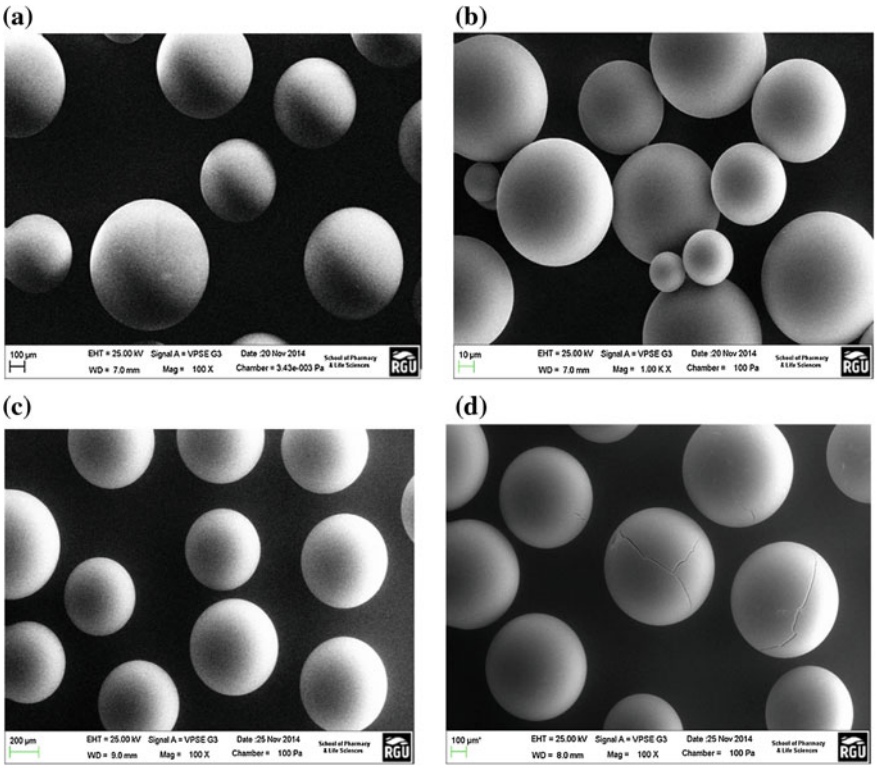


Fig. 8 **a** (Amberlyst 16), **b** (Dowex 50W8x), **c** (Amberlyst 36) and **d** (Amberlyst 15): SEM surface morphology of the fresh resin catalysts before esterification reaction

5 Conclusion

The permeation tests to determined the characterisation of gases with ceramic membrane for ethyl lactate production was achieved using Knudsen flow mechanism. The gas flux through the membrane increases linealy with gauge pressure at 298 K indicating a good fit of the data. Nitrogen with the higher kinatic diameter exhibited a higher permeance than CO_2 suggesting that the gas flow was not based on molecular seiving mechanism. The SEM of the cation-exchange resin catalysts before esterification showed a plain surface with samll cracks indicating a defect-free surface. The membrane pore radius was smaller than the mean free path indicating Knudsen mechanism of transport. The SEM of the membrane shows a decrease in size after modification while the EDXA showed that the ceramic membrane was initially coated with TiO_3 and subsequently with SiO_2 and ZrO_2.

Acknowledgment The conference was sponsored by IDEAS Research Institute, The Robert Gordon University, Aberdeen, United Kingdom. The Authors of this paper acknowledge the center for Process Integration and Membrane Technology at RGU for providing the research infrastructure.

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. Pereira CS, Silva VM, Rodrigues AE (2009) Fixed bed adsorptive reactor for ethyl lactate synthesis: experiments, modelling, and simulation. Sep Sci Technol 44(12):2721–2749
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. Dassy S, Wiame H, Thyrion FC (1994) Kinetics of the liquid phase synthesis and hydrolysis of butyl lactate catalysed by cation exchange resin. J Chem Technol Biotechnol 59(2): 149–156
5. Delgado P, Sanz MT, Beltrán S (2007) Isobaric vapor–liquid equilibria for the quaternary reactive system: ethanol water ethyl lactate lactic acid at 101.33 kPa. Fluid Phase Equilib 255 (1):17–23
6. Calvo JI, Bottino A, Capannelli G, Hernández A (2008) Pore size distribution of ceramic UF membranes by liquid–liquid displacement porosimetry. J Membr Sci 310(1):531–538
7. Mulder M (1996) Basic principles of membrane technology, 2nd edn. Kluwer Academic Publishers, The Netherlands
8. Li H, Schygulla U, Hoffmann J, Niehoff P, Haas-Santo K, Dittmeyer R (2014) Experimental and modeling study of gas transport through composite ceramic membranes. Chem Eng Sci 108:94–102
9. Lee H, Suda H, Haraya K (2005) Gas permeation properties in a composite mesoporous alumina ceramic membrane. Korean J Chem Eng 22(5):721–728
10. Kanellopoulos NK (2000) Recent advances in gas separation by microporous ceramic membrane. 1st edn. Elsevier science B.V, Amsterdam
11. Edward G (2006) Apparatus and method for separating gases. United state Patent No.: US 7,048,778 B2. Robert Gordon University, Aberdeen, UK

12. Tomita T, Nakayama K, Sakai H (2004) Gas separation characteristics of DDR type zeolite membrane. Microporous Mesoporous Mater 68(1):71–75
13. Julian A, Juste E, Chartier T, Del Gallo P, Richet N (2007) Catalytic membrane reactor: multilayer membranes elaboration. In: Proceedings of the 10th international conference of the European ceramic society, pp 718–722
14. Benito JM, Conesa A, Rubio F, Rodríguez MA (2005) Preparation and characterization of tubular ceramic membranes for treatment of oil emulsions. J Eur Ceram Soc 25(11):1895–1903
15. Pandey P, Chauhan R (2001) Membranes for gas separation. Prog Polym Sci 26(6):853–893
16. Okon E, Shehu H, Gobina E (2014) Gas transport and characterization of inorganic ceramic membrane for lactic acid esterification. In: Proceedings of the world congress on engineering and computer science 2014, WCECS 2014, 22–24 Oct 2014. Lecture notes in engineering and computer science, San Francisco, USA, pp 590–594

Gasification of Wood and Plastics in a Bubbling Fluidised Bed: The Crucial Role of the Process Modelling

Maria Laura Mastellone and Lucio Zaccariello

Abstract Gasification is a thermochemical process that aims to convert solid fuels into a synthetic gas that can be addressed to an end-use apparatus to produce electric energy and heat or can be further refined to be transformed in chemical valuable products. Many organic materials can be gasified under different operating conditions (pressure, temperature, reactants) in different kind of reactors. A technology widely utilized to gasify several solid materials is the bubbling fluidized bed reactor: this technology can ensure a very high heating rate but its performance is strongly affected by the material properties and its interaction with the bed during the process. This study aims to correlate the main process modeling outputs and experimental evidences obtained for the gasification of a commodity plastic waste, polyethylene, and a pinewood chip, to the hydrodynamics of a bubbling fluidized bed gasifier. To this end the rate controlling stage of the primary cracking stage has been determined in order to evaluate the reaction time for both the materials under typical gasification conditions and evaluate which is the characteristic time length of their primary cracking in the dense bed of the bubbling fluidized bed.

Keywords Fluidised bed · Gasification · Hydrodynamics · Modeling · Plastics · Wood

M.L. Mastellone (✉) · L. Zaccariello
Department of Environmental Biological and Pharmaceutical Sciences and Technologies, Second University of Naples, Via Vivaldi 43, 81100 Caserta, Italy
e-mail: mlaura.mastellone@unina2.it

L. Zaccariello
e-mail: lucio.zaccariello@unina2.it

H.K. Kim et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-94-017-7236-5_30

1 Introduction and Scope

Gasification is a thermochemical process that transforms a carbon-based material into a gaseous mixture of low molecular weight species. What's left is a clean "synthesis gas" that can be converted into valuable products and electricity. Gasification has been used on commercial scale for more than 75 years by the chemical, refining and fertilizer industries and for more than 35 years by the power industry [1]. First application of gasification was related to coal and derivative (i.e., char produced from coal) to obtain petrochemical product knowledge and combustible gases [2]. The utilization of coal gasification to obtain alternative fuels by means of the Fischer-Tropsch process dates back to the period of World War II when Germany and Japan produced 15,000 m^3/day of liquid fuels. The Fischer-Tropsch process is virtually applicable to a syngas produced starting from any waste with a sufficient carbon amount. The real challenge is the processing of syngas in order to eliminate particulate, adjust the syngas composition (CO/H_2 ratio) and remove hydrogen sulfide and carbon dioxide.

Gasification technology is expected to play a key role in the clean energy production from fossil and alternative fuels because of the possibility to transform the organic part of them into a synthesis gas (syngas/producer gas). In particular, gasification of waste is an economical and environmental viable solution to produce cleaner energy together with a remarkable waste weight reduction. The change of the fuel characteristics as well as the change of the main aim of the process requires a deepening of the effect that the specific waste has on the process performance. Waste gasification is in fact a not yet wide commercial utilization process because of conversion efficiency losses and syngas cleaning concerns [3]. The reactor technology has also an important effect on the gasification performance as well as the reactants type and the operating conditions (temperature and pressure) [4, 5]. The understanding of the process development, strictly related to the fuel structure and kinetic behavior under gasification conditions, is crucial to correctly operate the gasifier in the case of single fuel feeding as well as under co-gasification conditions [6, 7]. Typical undesired by-products of gasification process are heavy hydrocarbons condensable at ambient temperature (tar) and carbonaceous particle with very small mean size (few microns). The gasification of solid materials can be represented by a series-parallel physical modifications and chemical reactions including: heating; drying; primary cracking (solid → volatiles) often called devolatilization; secondary cracking; tar formation; homogeneous chemical reactions (volatiles involving); heterogeneous chemical reaction (char involving).

This study aims to compare the different processes' behavior of a specific wood (white pine) and a widely used plastic (polyethylene) during the gasification in a bubbling fluidized bed with specific reference to the primary cracking stage i.e. to the conversion of solid materials to volatiles and char. During this stage the solid material properties change by affecting the hydrodynamics of the dense bed. This influence can be inconsequential, if its duration is limited or, on the contrary, it can promote a fluidization worsening if the time length is sufficiently large. To establish if the

primary cracking of the solid materials tested affects in a crucial way the gasifier performance, a series of experiments and modeling calculations have been made and reported in this chapter. The controlling reaction rate (if any) of the primary cracking stage has been then evaluated to measure the characteristic length of the process by calculating the dimensionless numbers useful to this end: Biot number (Bi), Pyrolysis number (Py) and a combination of these latter, Py′ = Bi • Py [8, 9].

2 Brief Description of the Bubbling Fluidised Bed Hydrodynamics

A bubbling fluidized bed can be described by means of three schemes, useful for different purposes: the graphical scheme (Fig. 1a) shows a visual representation of the typical aspect of a bubbling fluidized bed (BFB) characterized by a zone where a bed made by particles is suspended by the fluidising gas (emulsion phase) coming from the bottom distributor and crossed by bubbles streams (bubble phase) characterized by a size and mean velocity distribution along the bed height due to the coalescence and by a zone where the voidage increases, bubbles disappear,

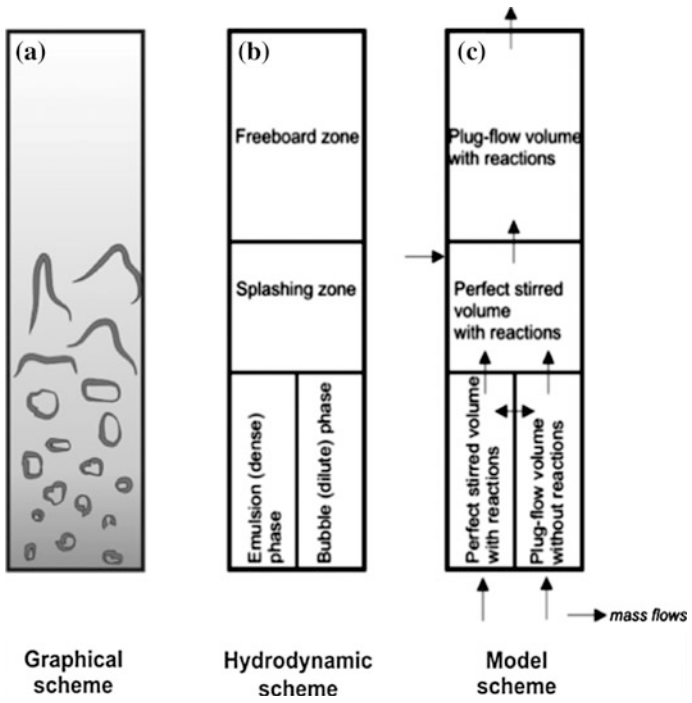


Fig. 1 Schemes of the bubbling fluidised bed gasifier utilised as reference for the model

turbulence decreases (freeboard). The changeover from the dense zone to the dilute zone is enough quick and discontinuous by creating a sort of "transition" zone that can be identified with the name of "splashing". In the case of a typical BFB, operated under conditions of a true bubbling regime, the hydrodynamics can be described by means of a model applied to system boundaries as reported in Fig. 1b and the kinetic modelling can be carried out by referring to a combination of ideal reaction volumes as reported in the same Fig. 1c. This representation highlights the peculiarity of a BFB reactor that can be seen as a combination of four zones: emulsion zone in the dense bed, bubble phase in the dense bed, splashing zone above the dense bed and freeboard zone between the splashing and the reactor exit. These zones have hydrodynamic regimes different from each to other and are connected by mass and energy exchange flows.

3 Description of the Process Reactions Occurring in the Dense Bed

Any thermal treatment, when applied to a solid particle, is characterized by similar first steps consisting in: progressive heating of the particle due to the bulk convective transport (external heating transfer); internal thermal diffusion; drying; primary cracking; volatiles releasing. Each material has a different thermal behaviour that has to be studied by specific experiments and modelling. Generally, at low temperature (100–200 °C) the moisture and volatile absorbed inside the material are released; at intermediate temperatures (350–600 °C) the weaker bonds are broken and a pyrolysis gas is released by leaving a carbonaceous amorphous solid called char; at higher temperatures (>800 °C) the degradation rate and gas conversion efficiency becomes higher but oils/tar are produced due to the aromatization induced by higher temperatures. During the heating and the releasing of moisture and volatiles the solid size, the apparent and particle density, porosity and chemical structure and composition changes. To evaluate in which way the materials change their properties during the heating-cracking period experimental observations, analytical measurements and mathematical modelling have been used and illustrated.

3.1 The Polyethylene Primary Cracking Modeling

The peculiar aspects correlated to the thermal treatment of plastic waste are: the very low content of ashes; the very low content of char; the low specific heat value of the solid polymer; the stickiness of the molten polymer; the very high primary cracking rate due to the low energy necessary to break the C–C and C–H bonds (Table 1).

Table 1 Energy values of the most important chemical bonds

Bond	Energy, kcal/mol
H–H	104
C–H	99
C–C	83
C=C	146
C≡C	200
O–O	35
O=O	119
O–H	111
C–O	86
C=O	177
H–F	135
H–Cl	103

Another physical property of polymers that affects the process in a bubbling fluidized bed is the low value of the melting energy (Table 2) and the high viscosity of the molten polymer.

Just after the injection into the hot fluidized bed, a very fast heat transfer mechanism leads the polymer pellet external surface up to the softening temperature [10]. The time necessary to reach this state can be evaluated by means of the dynamic energy balance on the single plastic pellet.

$$m_{fuel} \cdot c_p \cdot dT/dt = A \cdot h_{bed} \cdot \left(T_{bed} - T_{melting}\right) - r_{melting} \cdot \Delta H_{melting} \quad (1)$$

where m_{fuel} is the mass of the fuel, c_p is the specific heat capacity of fuel, T is the time-depending temperature of the pellet surface, A is the external pellet surface, h_{bed} is the heat transfer coefficient between the bed and the pellet, T_{bed} is the bed temperature, $T_{melting}$ is the temperature at which the pellet melts, $r_{melting}$ is the rate

Table 2 Input and output data for the energy balance

Parameter	Value	Units
Temperature (T)	850	°C
Gas viscosity (μ)	4.91E−05	Ns/m^2
Gas density (ρ)	0.315	kg/m^3
Inert particle diameter (d_{bed})	0.00035	m
Inert particle density (r_{bed})	2600	kg/m^3
Gas conductivity (k)	0.016	cal/s m°C
Fuel diameter (d_{fuel})	0.005	m
Density of fuel (ρ_{fuel})	950	kg/m^3
Specific heat of fuel (c_p)	0.55	kcal/°C kg
Melting heat of fuel ($\Delta H_{melting}$)	23.8	kcal/kg
Softening temperature of fuel	135	°C
Conductivity of fuel (k_{fuel})	0.46	cal/s m °C

of melting, $\Delta H_{melting}$ is the latent heat of melting. The term related to the melting process absorption can be neglected compared with the heat transferred by the fluidized bed that can be evaluated by means of the following relationship proposed:

$$h_{bed} = Nu \cdot k_{fuel} / d_{fuel} \tag{2}$$

where Nu is the Nusselt number, k_{fuel} is the fuel conductivity and d_{fuel} the fuel pellet diameter. By setting the d_{fuel} at 6 mm, the reactor temperature at 850 °C, the conductivity of the fuel as reported in Table 2 and by calculating the Nusselt number accordingly to Leckner et al. [11], the value of h_{bed} is 272 W/m^{2}K.

The general solution of the energy balance is reported in Eq. 3.

$$T(t) = T_{bed} + (T^{\circ} - T_{bed}) \cdot \exp(-h_{bed} \cdot A/m \cdot c_p) \tag{3}$$

With reference to data reported in Table 2, the estimated time necessary for the softening of the external surface, as calculated by Eq. 3, is about 1 s. This means that the agglomeration between the plastic pellets and the bed particles starts just after the injection in the fluidized bed. Once the pellet surface molten (i.e. very rapidly at bed temperatures greater than 450 °C), several sand particles stick on the plastic surface, forming an *aggregate* that has the external shell made of sand particles and the internal core made of polymer not yet molten (Fig. 2). The molten polymer flows throughout the bed particles forming the external shell, so forming a uniform coating over and between them and by promoting an adhesion increasing of other bed particles that are close to the aggregate. Moreover, the polymer flowing throughout the bed particles promotes the adhesion of several layers of inert material, until the polymeric mass completely flowed throughout the sand. The hydrodynamic variables, like the size and density of the bed particles and the fluidization velocity having a key role in determining the overall performance of the reactor [12], dramatically changes if agglomerates lifetime exceeds few seconds. At bed temperature as high as 850 °C the formation and crumbling of the aggregates are almost undistinguished [13]. This is due also to the parallel beginning of the primary cracking of the carbon–carbon bonds of the polymer chain (occurring at a

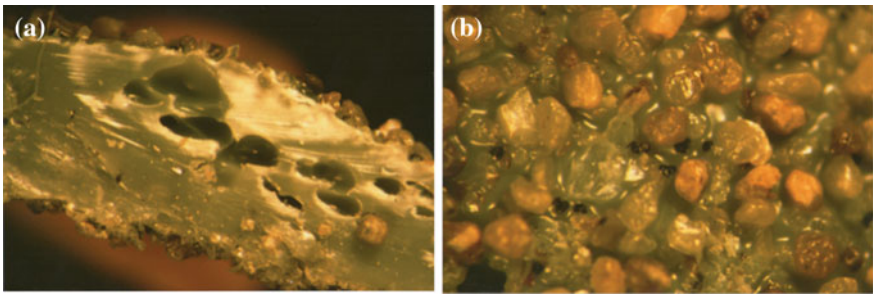


Fig. 2 Photos of a PE pellet kept 15 s in BFB at 450 °C. Transversal section (**a**); external surface (**b**)

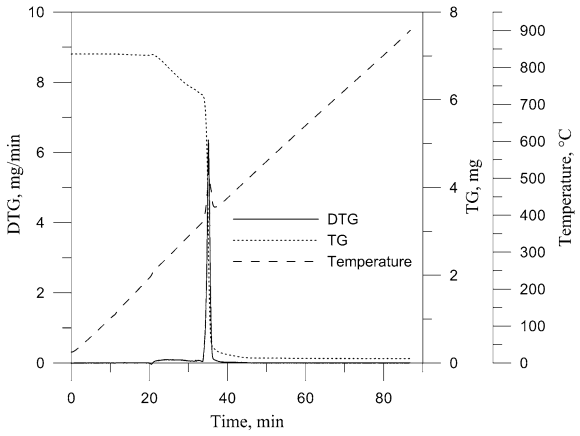


Fig. 3 Thermogravimetry and differential thermal analysis of polyethylene under inert atmosphere

mass temperature of about 400 °C), i.e. to the beginning of the primary cracking process that starts when the polymer has already covered the bed particles. In this situation the value of h_{bed} is as high as 1100 W/m^{2}K. By using Eq. 3, for a 10 μm thickness layer, the starting time for degradation (350–400 °C for thermoplastic polymers, Fig. 3) is about 5 s at 850 °C and 30 s at 450 °C (in accordance to experimental findings reported in Fig. 2).

As above reported, the crumbling occurs when the polymer forming adhesive bridges between the particles undergo primary cracking i.e. thermal cracking.

By assuming the internal polymer temperature as constant, the rate of mass loss results independent on the internal temperature profile and the solution of the related equation can be strongly simplified. Given m the mass of the fuel at the generic time t and temperature T, m_0 the mass of the initial fuel, m_∞ is the mass of solid residue, A is the pre-exponential factor, E is the apparent activation energy, T is the fuel temperature, the general equation of mass loss is:

$$-\frac{\partial m}{\partial t} = A\exp(-E/RT)(m - m_\infty)^n \tag{4}$$

The kinetic parameters and the reaction order n can be calculated by the mass loss data (TG) and by obtaining the df/dt versus t (DTG), with $f = (m_0 - m)/m_0$. The DTG reported in Fig. 3 shows one peak corresponding to 420 °C. The data of f, df/dt and the related temperature (T) corresponding to the data interval of the peak can be utilized to fit the kinetic equation (Eq. 4) and find the A and E parameters by means of the Arrhenius plot (5).

$$\ln\left[\frac{1}{(1-f)}\frac{df}{dt}\right] = \ln A - E/RT \tag{5}$$

By fitting the experimental data reported in the previous by means of Eq. (5) the kinetics parameters, A and E, for a 1-order kinetic equation, have been calculated:

Table 3 Comparison between the dimensionless numbers as evaluated for PE and pinewood

	Biot (Bi)	Pyrolysis number (Py)	Py · Bi	Control
PE (pellet)	4.9	5.9×10^5	2.4×10^6	–
PE (layer)	0.3	1.5×10^9	5.1×10^8	Kinetic

the frequency factor results to be 2000 s^{-1}, the E/R equal to 2.9×10^4. These data can be used to evaluate the dimensionless numbers above recalled in order to determine which is the controlling stage of the process is the external heating, internal heating of intrinsic chemical kinetics. The resulted values are reported in Table 3 and indicate that for the PE layer on the bed particles of a fluidised bed the control is the external heating.

According to Bates and Ghoniem [14], the kinetically controlled regime occurs for small Biot number (<1) and large Pyrolysis numbers (>10) so indicating the controlling reaction stage for the PE under form of layer is the kinetic one. Under this condition the material has a uniform temperature, equivalent to the reactor temperature, and the reactions occur uniformly throughout the particle.

A graphical resume of the described process model is reported in Fig. 4 that shows how the polyethylene particle undergoes different physical modifications before undergoing the chemical cracking and volatile gasification. The described steps are those occurring in the dense bed.

The main feature of the polyethylene gasification in the bubbling fluidised bed is then the very fast primary cracking rate and the good mixing of the fuel in the dense

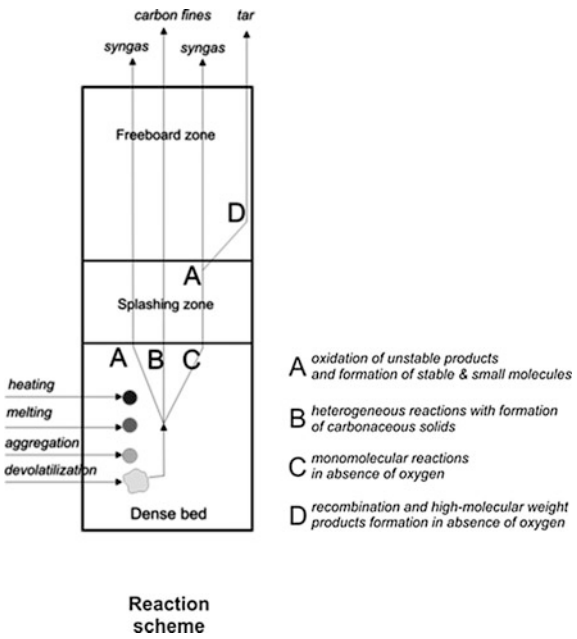


Fig. 4 Graphical representation of the reactions' steps occurring in a fluidized bed fed by polyethylene particles

bed. No segregation occurs at the bottom of the bed (heavy particle) or at the top (floating particle) because of the fast crumbling of agglomerates at temperature typical of gasification.

3.2 The Wood Primary Cracking Modeling

Wood is a natural polymer and its chemical structure is different by that of the polyethylene. While this latter is a repeating linear sequence of methyl monomer, wood has a heterogeneous aromatic structure. Pyrolysis and gasification of biomass have been extensively studied and a lot of data are available from both scientific literature and industrial applications [15]. The process modelling is completely different by that of polyethylene: the interaction between the wood chips and the bed material is in fact completely different by that of the polymer because there is no melting and, consequently, no aggregation phenomena. The heating process does not induce any formation of bed material-fuel agglomeration but promotes a continuous change in the apparent density of the wood chip due to the progressing of primary cracking inside the bed by following a "shrinking volume" model, in consideration that ash has a low content and are not cohesive.

The reason why the behaviour of a wood chip is different by that of a polyethylene pellet is the chemical structure of the wood components and the physical properties of intermediates. The heterogeneity of wood is clearly deduced by the TG curves and DTG of Fig. 5 that shows three peaks (other than the moisture loss) representing the hemicelluloses, cellulose and the lignin content, each of this having different degradation kinetics. By applying Eq. (4) and Eq. (5) to these data the kinetic parameters of Table 4 have been found.

The external surface temperature of a wood chip has been obtained by means of Eq. 3 and reported in Fig. 6 together with that calculated for PE. Even in this case,

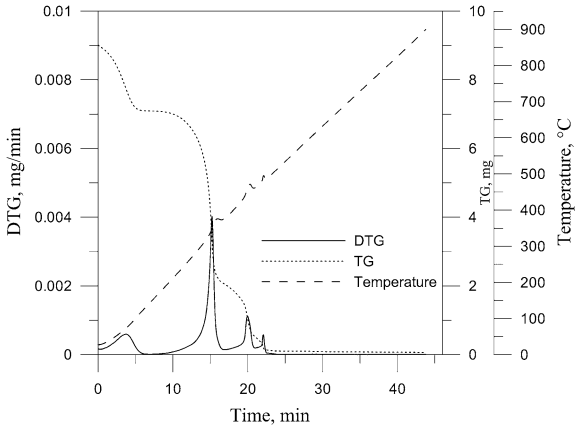


Fig. 5 Thermogravimetry and differential thermal analysis of wood (inert)

the controlling stage has been evaluated by using the approach of Pyle and Zaror [8] and reported in Table 5.

The controlling resistance is therefore the internal thermal diffusion as calculated with reference to the initial wood chip size and density. With the progressing of heating these properties change and, as a consequence, also the controlling stage can varies. In order to verify how these properties change with the heating and primary cracking processes, a series of experiments have been made on whole wood chips. The main objective of the experiments was to measure the shrinkage of the particle and the variation in the apparent density with the temperature.

Data of Fig. 7 report the resulting weight loss as obtained after a given residence time (15 min) at increasing values of the reactor temperature. In this case, for each wood chip the final weight, the size and the apparent density have been evaluated.

Table 4 Kinetic parameters as evaluated for the pine wood

	E, J/mol	k_0, s^{-1}
Peak1	1.81E+04	2.03E−01
Peak2	7.86E+04	1.58E+03
Peak3	1.79E+04	2.06E−02

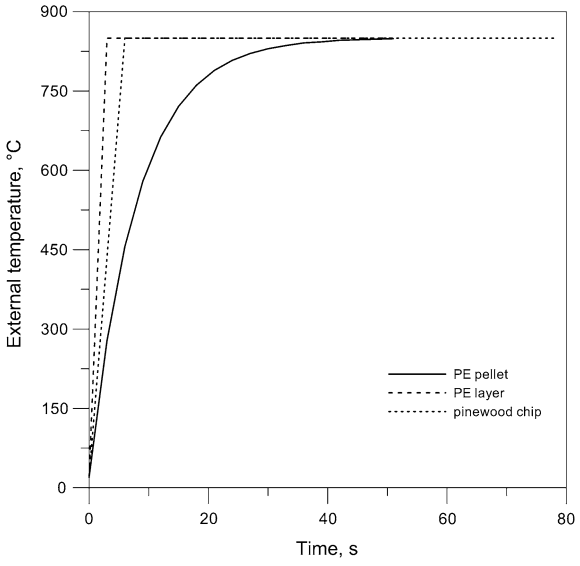


Fig. 6 Heating curves related the external surface of a PE pellet, PE layer, pinewood

Table 5 Comparison between the dimensionless numbers as evaluated for PE and pinewood

	Biot (Bi)	Pyrolysis number (Py)	Py * Bi	Control
Pinewood chip	73	3.6×10^{-4}	2.7×10^{-2}	Internal

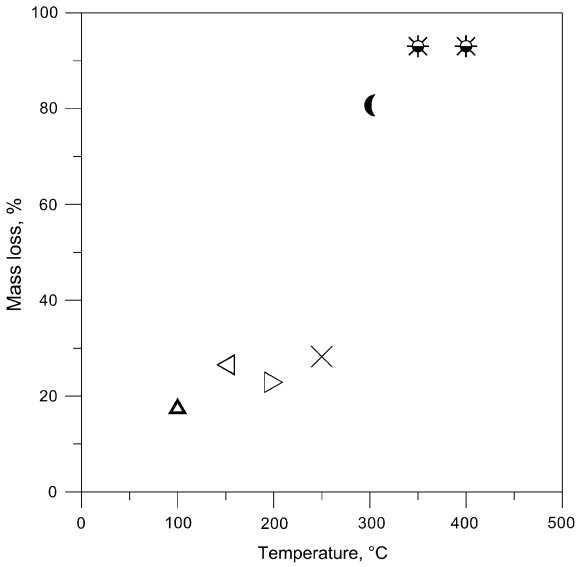


Fig. 7 Experimental data of mass loss of whole wood chips

Figure 7 shows that the temperature of 250 °C is the onset of the mass loss for the pinewood chip accordingly with the TG-DTG data.

The mass loss of a wood chip is accompanied by a change in the apparent density and, after a sufficient time, by a change of size due to the comminution phenomena. Figure 8 shows how the value of the ratio between the final apparent density and the initial one, dramatically changes when primary cracking begins.

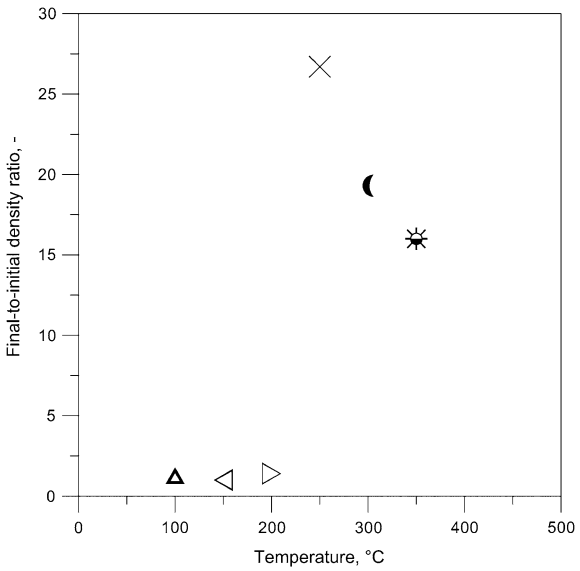


Fig. 8 Ratio between the final and the initial apparent density of the wood chips

The increasing of the apparent density induces a segregation effect in the bubbling fluidized bed of the reacting particles that cannot be fluidized and drops at the bed bottom until the comminution produces small fragments or the cracking is completed. The segregation of a large number of particles at the bed bottom creates a pyrolysis zone in this part of the bed. This feature, together with the considerations reported in the following paragraph, can explain why the BFB gasification of wood generally produces a large amount of tars. In fact, the oxygen content in the emulsion phase of a dense bed is much lower than that expected and calculated by the given oxygen-to-fuel ratio due to the segregation effect in the bubble phase. In the next paragraph the oxygen profile in the bed is evaluated under typical conditions of FB gasification.

4 Oxygen Distribution in the Dense Bed

The extension of oxidizing reactions in the dense bed, or, better, in the emulsion phase of dense bed, depends on the effective availability of oxygen in that zone. A crucial aspect of BFB gasification modelling is related to the real availability of oxidant in the same zone of volatiles releasing. The unavailability of oxygen, or any other oxidising agent, determines the occurrence of molecular rearrangement of hydrocarbons fragments and radicals into stable and heavy molecules (pathways B and C in Fig. 4). These latter are precursors of the most undesired by-products of gasification of plastics: carbonaceous particles and, overall, tars (pathway D in Fig. 4).

In a BFB the largest part of the oxygen is present as bubble phase because of the well-known two-phase theory for which the only fraction of fluidizing gas present in the emulsion phase is related to the minimum fluidization velocity. The oxidant flow rate determines a superficial gas velocity that is generally 5–10 times of the minimum fluidization velocity so that the fraction of oxygen actually present in the emulsion phase is in the inverse ratio. The oxygen passes throughout the bubbles surface to the emulsion phase along the rising the bed. The oxygen mass balance that allows calculating the oxygen concentration in the bubbles as a function of z is:

$$\frac{dC_{O_{2,b}}}{dz} = -k_m/U_b \cdot \left(C_{O_{2,b}}\big|_z - C_{O_{2,e}}\big|_z\right) \tag{6}$$

In the case of complete consumption of oxygen in the emulsion phase due to the oxidation reactions the oxygen concentration is 0 and the equation can be solved as in the following:

$$C_{O_{2,b}}(z) = C_{O_{2,b}}(0) \cdot \exp(-k_m/U_b \cdot z) \tag{7}$$

where k_m is the mass exchange coefficient.

Following equation puts into relation the single bubble rising velocity (U_b) with the superficial gas velocity (U_g), minimum fluidization velocity (U_{mf}), bed diameter (D_{bed}) and bubble diameter (D_b):

$$U_b(z) = 0.35 \cdot (32.2 D_b/0.3048)^{0.5} \left\{ \tanh[3.6(C_{b,z}/D_b)^{0.5}]^{1.8} \right\}^{1/1.8} \quad (8)$$

The bubble diameter can be obtained by:

$$D_b = \frac{0.62}{32.2^{0.2}} \left[\left(\frac{U_g - U_{mf}}{0.3048} \right)^{0.4} \right] \cdot \left\{ \left[z + 3.37 \left(\pi D_{bed}^2/4 \right)/N_{nozzles} \right]^{0.5} \right\}^{0.8} \quad (9)$$

This model allows correlating the concentration of oxygen in the bubble phase at different level of the dilute phase of the bed with the main operating conditions and geometrical parameter of the reactor.

The following figures report the oxygen concentration profile and the transfer efficiency of the oxygen along the dense bed as evaluated by using (6). It is evident that the bed height must be fixed in such a way to obtain the completeness of oxygen transfer in order to have the desired equivalence ratio in the dense bed (bed diameter has been fixed to 1 m).

The maximum concentration of oxygen in the bubble phase is present at the bottom of the bed and, at the same z level, the oxygen amount in the emulsion phase is very low. This means that the pinewood chips dropping at the bed bottom undergo the primary cracking in a zone with an oxygen-to-fuel ratio much lower than necessary to have gasification. This feature can enhance the production of tars due to the cyclization of radicals (Figs. 9 and 10).

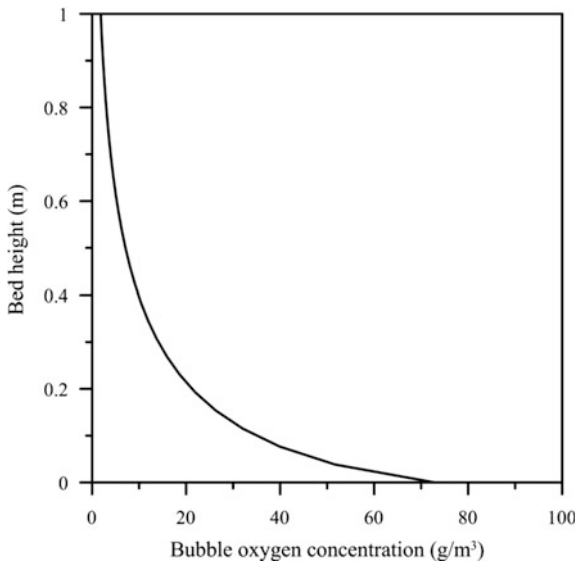


Fig. 9 Oxygen concentration profile (D_{bed} = 1 m)

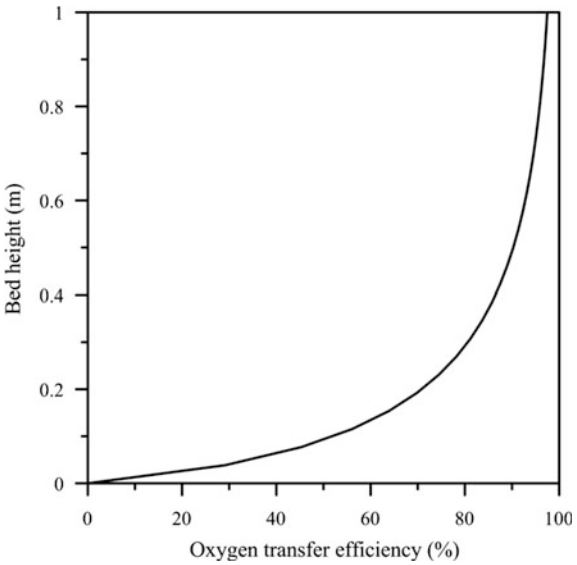


Fig. 10 Transfer oxygen efficiency (D_{bed} = 1 m)

5 Conclusions

The bubbling fluidised bed gasification of the polyethylene, as representative of thermoplastic polymers, and that of a woodpine chips, has been characterized by looking at the material interaction with the dense bed. The materials undergo a series of physical modifications of their structure and composition due to: the heating of the external surface, the primary cracking controlled by the intrinsic kinetic, the internal heat diffusion inside the particle. The controlling stage has been determined in order to understand which is the process rate on the basis of that the BFB must be appropriately designed.

In the case of polyethylene, the time necessary to complete the process from the beginning of heating (just after the injection) to the completion of primary cracking can be estimated by considering the intrinsic kinetic as limiting the overall process rate.

The wood pine chips undergo a reacting process controlled by the internal heat transfer rate whose modelling is complicated by the variation of size and density of the particle during the reaction itself. Experiments have been associated to the modelling to evaluate the variation of density and size with temperature increasing.

The complex hydrodynamics of the bubbling fluidized bed and the segregation of the oxygen in the bubble zone favour the monomolecular as well as the recombination reactions of volatiles in the emulsion zone. The correct height of the bed, correlated to several geometrical and operating parameters, is crucial in order to guarantee a fast and complete mass interphase exchange of oxygen from bubbles to emulsion zone. In any case the woodpine chips segregate at the bottom under typical fluidising velocity of BFB until the primary cracking is concluded. In this

case, the volatiles released by the reacting particles are only partly oxidised due to the limited fraction of oxygen available for reactions in this zone. The formed PAHs are generally to stable to be oxidised in the successive passage in other oxidising zones of the reactor so forming the tar.

References

1. GTC (ed) (2014) Gasification—an investment in our energy future. Arlington
2. Speight JG (ed) (2013) Coal-fired power generation handbook. Scrivener Publishing/Wiley, USA
3. Arena U, Zaccariello L, Mastellone ML (2009) Tar removal during the fluidized bed gasification of plastic waste. Waste Manag 29:783–791
4. McLendon TR et al (2004) High-pressure co-gasification of coal and biomass in a fluidized bed. Biomass Bioenergy 26:377–388
5. Mastellone ML, Zaccariello L (2013) Metals flow analysis applied to the hydrogen production by catalytic gasification of plastics. Int J Hydrogen Energy 38:3621–3629
6. Mastellone ML, Zaccariello L, Arena U (2010) Co-gasification of coal, plastic waste and wood in a bubbling fluidized bed reactor. Fuel 89:2991–3000
7. Mastellone ML, Zaccariello L (2014) The crucial role of the process modelling during the design of a bubbling fluidised bed gasifier of plastics. In: Proceedings of the world congress of chemical engineering and computer science 2014, WCECS 2014. Lecture notes in engineering and computer science, 22–24 Oct 2014, San Francisco, USA, pp 618–623
8. Pyle DL, Zaror CA (1984) Heat transfer and kinetics in the low temperature pyrolysis of solids. Chem Eng Sci 39:147–158
9. Carrasco JC et al (2014) Observed kinetic parameters during the torrefaction of Red Oak (*Quercus rubra*) in a pilot rotary kiln reactor. BioResources 9(3):5417–5437
10. Mastellone ML, Arena U (2004) Bed defluidization during the fluidized bed pyrolysis of plastic waste mixtures. Polym Degrad Stab 85:1051–1058
11. Leckner B, Palchonok GI, Andersson BA (1992) Representation of heat and mass transfer of active particles, in IEA mathematical modelling meeting, Turku, Apr 1992
12. Kunii D, Levenspiel O (1991) Fluidization engineering, 2nd edn. Butterwotrh-Heinemann, Boston
13. Arena U, Mastellone ML (2000) Defluidization phenomena during the pyrolysis of two plastic wastes. Chem Eng Sci 55:2849–2860
14. Bates RB, Ghoniem AF (2012) Biomass torrefaction: modeling of volatile and solid product evolution kinetics. Bioresour Technol 124(11):460–469
15. Miller RS, Bellan J (1997) A generalized biomass pyrolysis model based on superimposed cellulose, hemicellulose and lignin kinetics. Combust Sci Technol 126(1):97–137

Impact of Some Agro Fluids on Corrosion Resistance of Mild Steel

Ayo Samuel Afolabi, Anthony Chikere Ogazi and Feyisayo Victoria Adams

Abstract The corrosion behavior of mild steel in apple, mango, grape, orange and the mixture of these agro fluids were electrochemically studied. Chemical compositions of both mild steel and the agro fluids were carried out to determine the corrosion mechanism for the reaction. Polarization behaviors of mild steel in the agro fluids were determined by Tafel extrapolation curves over the interval of five days for a sixty-day immersion period at a constant temperature of 27 ± 2 °C. The cathodic polarization curves were almost identical irrespective of variation in concentration of the various fluids while the anodic polarization curves exhibited varying active and passive corrosion behavior. Also, the corrosion rates of the alloy decreased with increase in immersion period which could be due to gradual decline in the concentration of the acidic level in the fluids within the given range of potentials. Hence, the evolution of hydrogen gas and reduction of oxygen molecules from the reacting system were presumed to be major factors decreasing corrosiveness of the solution involved. SEM and EDS analysis of the corroded mild steel showed the respective compositions of the mild steel after the electrochemical tests. The result obtained from the study showed that electrochemical corrosion rate over the duration of immersion was in a range of mixture of the fluids > orange juice > grape juice > mango juice > apple juice.

Keywords Agro fluid · Characterization · Corrosion rate · Electrochemical behavior · Exposure time · Mild steel

A.S. Afolabi · A.C. Ogazi
Chemical Engineering Department, University of South Africa, P/Bag X6, Florida, Johannesburg 1710, South Africa
e-mail: afolaas@unisa.ac.za

F.V. Adams (✉)
Department of Petroleum Chemistry, School of Arts and Sciences, American University of Nigeria, Yola PMB 2250, Adamawa State, Nigeria
e-mail: feyikayo@gmail.com

H.K. Kim et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-94-017-7236-5_31

1 Introduction

Corrosion exhibits significant effects on materials. It reduces the safe, consistent and effective equipment operations and structures which eventually lead to the loss of the affected material [1, 2]. Metals form the essential bases for modern technological civilization. One of such areas is in the agro industry, where metallic alloys are widely used in the industrial processing and packaging of fruit juices. Most metallic corrosion result from electrochemical effects exhibited by these juices on the metals. Shreir [3] described electrochemical corrosion as a heterogeneous redox reaction at metallic/non-metallic interface in which the metal is oxidized and the non-metal is reduced. When this occurs, corrosion is initiated by the flow of electrons between the electrode sites of different potentials in contact with aqueous electrolytic solution [4, 5].

Studies showed that metals can corrode when exposed to the atmosphere as well as in acidic solutions [6, 7]. According to these researchers, corrosion involves the transfer of electrons along the surface of the metal under the influence of a potential difference. Sharma and Sharma [8], observed that metallic alloys do not corrode in dry air or in water completely free of air but require oxygen and water to occur. Corrosion is accelerated by acids or by contact with less active metals such as copper or lead. Certain salt solutions also accelerate corrosion, not only because they are acidic by hydrolysis, but also because of specific catalytic effects or reactions of the anions. Therefore, there is effective collision of particles which affect corrosion rate [9].

According to Costescu et al. [10], fruit juices are liquid, non-alcoholic products with different degrees of clarity and viscosity, obtained through pressing or breaking up of fruits with or without sugar or carbon dioxide addition. Agro fruits, exhibit a high level of carboxylic acidity which would have a corrosive effect on metals at different rates. Organic acids directly play an important role in the growth, maturation and acidity of the fruit, and also affect the shelf life of the fruit by influencing the growth of microorganisms [11, 12]. Organic acids such as citric, malic, oxalic and tartaric acids ranging from 0.1 to 30 g/L were found in orange, grape, mango and apple juices. However, there was a considerable difference in the organic acid content found in various types and brands of fruit juice [13]. According to Toaldo et al. [14], analysis of grape juices from *Vitis Labrusca L.* showed it contains a significant amount of gallic acid, in addition to phenolics, monomeric anthocyanins and antioxidant from its seeds. Apple concentrate was found to have a higher amount of malic acid than other carboxylic acids [11]. Brae burn apples contained the highest amount of citric acid in apples; however, Granny Smith apples were the overall most acidic apples tested. High pressure liquid chromatography (HPLC) has been studied to be very efficient chromatographic technique for determining chemical composition of organic acids in agro juices although absolute precaution is required [12, 13, 15].

Environmental factors like oxygen concentration in water or atmosphere, the pH of the electrolyte, temperature, concentration of various salts solutions, and many

more in contact with the metal play a significant role in the rate of corrosion of metals even if such metallic materials are completely homogeneous in nature. Meanwhile, hydrogen evolution from an acidic environment is responsible for the sustenance of corrosion of metal [16]. Higher concentration of a solution will cause more hydrogen gas evolution. The stability of the halide in the surface complex determines the effect of corrosion kinetics of the metal/alloy. According to Marcus and Maurice [17], metal (M) corroding to M^{2+} ions at the anode in the presence of water would be reduced to hydroxyl ions and hydrogen at the cathode. The hydroxyl is readily oxidized by air to a hydrated compound during the corrosion process.

Metallographic examination is one of the most procedures used in failure analyzes of metals. It involves microstructural inspection of corroded metallic materials or their alloys to ascertain the extent of corrosion [18]. Highly precision electron metallographic equipment, such as the scanning electron microscope (SEM), transmission electron microscope (TEM), energy disperse spectroscopy (EDS) and X-ray diffraction (XRD) are used for such analyzes [18]. SEM is very useful to show surface morphology and it is widely applied in material science [19].

This work examines significantly the electrochemical corrosion behavior of mild steel in orange, mango, grape, apple juices and their mixture to determine the various rates of corrosion under given range of physical conditions; likewise to compare the corrosion mechanisms of the corroded mild steel samples in these agro fluids.

2 Materials and Methods

The commercial mild steel used for the electrochemical corrosion studies was supplied by ArcelorMittal South Africa and the percentage chemical composition of the mild steel is presented in Table 1.

Square-base mild steel test specimens (10 mm × 5 mm thickness) were machined from Buehler IsoMet 4000 (USA) linear precision metal cutting machine and mounted in cold-curing polyester resin to reveal flat surfaces in contact with the corrosion media. The terminals of the test specimens were linked by insulated stripes of copper wire. Mild steel specimens were abraded using 220, 600 and 1200 grit emery papers mounted on IMPTECH (20 PDVT) grinding and polishing machine at average speed/force of 300 rpm/30 N over the duration of 4 min according to Advanced Laboratory specifications. They were later polished with diamond abrasive pastes of 3, 1 μm and 50/nm grit sizes at average speed 150/rpm force of 25 N for 3 min.

Table 1 Chemical composition of mild steel [20]

Thickness (t) (mm)	Fe	C	Mn	P	S	Si
t < 4.5	98.48	0.150	1.000	0.035	0.035	0.300

The mild steel samples were then degreased with acetone rinsed with distilled water and dried at ambient temperature. Polarization curves at various immersion time were measured by the open circuit potential (OCP) and recorded potentiodynamically with scan rate (potential sweep) of 0.002 (V/s). Corrosion potential measurement commenced from −1.0 V and ended at 2.0 V. Corrosion current densities (I_{corr}) and corrosion potential (E_{corr}) were evaluated from the intersection of linear anodic and cathodic branches of the polarization curves in accordance with Tafel extrapolation method adopted by Poorsqasemi et al. [21] investigation. The pH values of the various agro juices were also taken before and after each exposure time using a standard portable MBI model 3D (Montreal, Canada) pH meter. Other corrosion parameters such as anodic and cathodic Tafel slopes analyzes and evaluation of corrosion properties based on ASTM 59, 96 and 159 standards were also considered from the polarization curves by Tafel extrapolation.

Corrosion media investigated include fruit species of freshly harvested orange (*citussinensis*), mango (*chok Anan*), grape (*vitisvinifera L.*), apple (*delicious*) juices and their mixture. The various agro fluid samples were prepared by extracting the juices from freshly harvested fruits using commercial blender and later kept in a refrigerator at 0 °C [20]. Organic acids are mostly responsible for corrosive effects of agro juices [22, 23]. However, there is also minute presence of phenolic content, fatty acids and amino acids in the agro fluids [11, 22]. Because of the variation in the concentration and composition of organic acid in different agro fluids, it became necessary to determine the chemical composition of these acids in each medium.

Analysis of the organic acid content in the agro fluids was studied using high pressure liquid chromatography (HPLC) method. The organic acids were identified and quantified with the aid of ultraviolet (UV) detector with a wavelength of 250 nm attached to a model K-2502 KNAUER equipment by comparing their retention times and peak heights with standard organic acid solutions. Potentials of hydrogen (pH) values of the agro fluids were recorded before and after the exposure time to establish differences in their acidic level. Table 2 shows the analysis of the agro fluids used for the study.

Microscopic analysis of the electrochemical corrosion results was performed on the corroded metallic samples. Comparisons were drawn from surface morphologies to determine the extent of corrosion of the metallic alloy in the various agro fluids. The techniques used to evaluate corrosion products after electrochemical

Table 2 Chemical composition of agro juices [20]

Apple (*delicious*) (g/L)	Grape (*vitisvinifera*) (g/L)	Mango (*chokAnan*) (g/L)	Orange (*citussinensis*) (g/L)
Citric 0.064	Citric 0.072	Citric 2.940	Ascorbic 0.652
Malic 2.840	Malic 3.500	Malic 7.520	Citric 14.012
Shikimic 0.021	Succinic 0.002	Tartari 0.980	Malic 1.525
Succinic 0.210	Tartaric 7.140	Succinic 2.090	Lactic 1.913
Tartaric 0.019			Tartaric 0.382
Quinic 0.611			Oxalic 0.105

studies were scanning electron microscopy (SEM) and energy dispersive X-ray spectroscopy (EDS). SEM analysis of the steel specimens was examined on the corroded surfaces considering the surface exposed to the air side. The SEM reading was taken at magnification of 1500X/20 μm to expose wider corroded region with more distinguishable characteristics of the metal. The spectrum processor of the EDS was set to depict distinction in the composition of the surface elements.

3 Results and Discussion

3.1 Electrochemical Behavior of Mild Steel in Agro Fluids

Tafel slope analysis presented in Fig. 1 shows relative polarization behavior of mild steel in various agro fluids at 5th and 60th days of immersion. It is observed from the nature of various curves that the concentration of the environments differs completely. Figure 1a reveals that all the agro fluids have insignificant passivation of their oxide layers on the 5th day except for orange juice which shows passivation at potentials between −0.3 and 0.2 V. However, a pseudo-passivity behavior was observed on the 60th day in all the fluids tested (Fig. 1b).

Comparison of electrochemical corrosion parameters (Table 3) shows that E_{corr} of the corroded mild steel is highest in apple juice both on the 5th and 60th day (−552 and −402 mV, respectively) while orange has the least in the value of −576 mV on the 5th day and mango has potential of −330 mV on the 60th day. The highest and lowest I_{corr} emerged from mango and grape media, respectively both on 5th and 60th days as observed from the data. The maximum anodic Tafel slope (βa) is exhibited in apple juice (0.169 V dec^{-1}) while the least is shown in grape juice (0.115 V dec^{-1}) on the 60th day. However, βa was highest in grape juice and lowest in the mixture of the fluids on the 5th day. Apple juice maintains highest cathodic Tafel slope (βc) of 0.153 V dec^{-1} while orange juice exhibits the least (0.062 V dec^{-1}) on the 60th day. The βc observed on the 5th day was reverse of βa (i.e. highest in the mixture and lowest in grape juice).

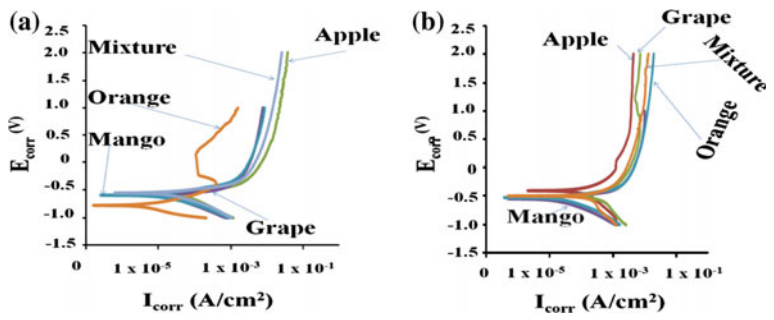


Fig. 1 Polarization curves of mild steel specimens on **a** 5th day, **b** 60th day immersion period

Table 3 Comparison of electrochemical corrosion parameters on the 5th and 60th days of immersion

Immersion period	Electrolyte	Parameter			
		βa ($V\ dec^{-1}$)	βc ($V\ dec^{-1}$)	I_{corr} ($\mu A\ cm^{-2}$)	E_{corr} (mV)
5th	Apple	0.150	0.065	2.102	−552
	Grape	0.194	0.044	1.247	−578
	Mango	0.127	0.090	6.621	−584
	Orange	0.127	0.064	2.017	−763
	Mixture	0.075	0.121	6.385	−573
60th	Apple	0.169	0.153	5.969	−402
	Grape	0.115	0.103	1.034	−499
	Mango	0.123	0.065	6.667	−530
	Orange	0.157	0.062	1.460	−515
	Mixture	0.118	0.084	5.624	−490

Corrosion rates (CR) of the mild steel specimens in the various agro fluids decreased over the duration of the study as seen from Table 4. According to Tran et al. [24], organic acids enhance corrosion rates of mild steel samples by accelerating cathodic reaction either through direct reduction at the metal surface or by means of buffering effect which involves dissociation of the hydrogen ions near the corroding surface.

The highest electrochemical corrosion rate was obtained from the mixture medium. It maintained the lead from 15th day (7.430 mm/year) to 60th day with 1.672 mm/year corrosion rate. Hence, it is possible that orange juice contains more acidic ions with lower pH value than mango, apple and grape lately. This, of course,

Table 4 Electrochemical corrosion rates of mild steel in the agro fluids [20]

Corrosion rate: CR (mm/year)					
Duration (days)	Apple	Grape	Mango	Orange	Mixture
5	6.946	9.684	7.759	8.056	9.224
10	5.435	7.278	7.705	7.435	7.571
15	3.131	5.809	6.750	7.014	7.430
20	2.447	5.068	6.080	5.154	7.296
25	2.275	3.632	5.706	4.562	6.545
30	2.069	3.469	4.215	2.972	6.031
35	1.825	2.275	3.063	2.352	5.944
40	1.741	1.933	2.387	2.347	3.578
45	1.567	1.752	2.305	1.700	3.090
50	1.383	1.560	1.669	1.593	2.307
55	1.344	1.451	1.356	1.558	1.899
60	1.301	1.403	1.339	1.530	1.672

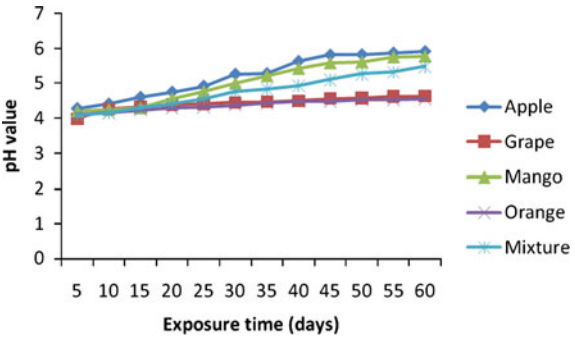


Fig. 2 pH variation of the agro fluids with time

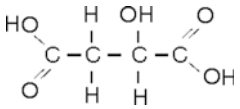
may have a great impact on the dissolution of passivating oxide layers on the metallic sample leading to more rapid corrosion rate. The mixture of these fluids, however, has more profound corrosion rate than any of the individual agro fluids as noted. Possibly presence of more acidic group in the fluid may have given rise to greater corrosive attack on the metallic specimen than other agro juices. There was significant decline in the potential of hydrogen (pH) values in the various corrosion fluids on the 60th day of immersion as indicated: apple 5.82, grape 5.45, mango 5.60, orange 5.31, and mixture 5.08, respectively.

The pH values of the corrosion media from electrochemical study is graphically presented in Fig. 2. The concentrations of acid in the various agro media decreased linearly with increasing immersion times possibly because of greater loss of hydrogen ions displaced from the various agro media due to corrosion of the mild steel samples.

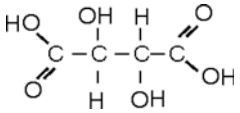
It is important to note that hydrogen ion [H^+] concentration in the agro fluids might have played an important role in enhancing corrosion rate. This is because CR is a function of concentration of the organic acids and may depend on the pH values of the reacting agro species as agreed by similar investigations [24]. Therefore, orange juice might have contained more acidic ions with low pH values than other media and subsequently exhibited more corrosive effect on the mild steel sample followed by grape juice, mixture medium, mango and apple juices, respectively.

3.2 Corrosion Mechanism of Mild Steel in Agro Fluids

The mechanism of corrosion processes involves the electrochemical interactions between the mild steel which serves as the anode and the agro fluids. The possible equations for chemical reactions during the electrochemical corrosion processes of the mild steel specimen in the agro fluids were proposed. The corrosion mechanisms of the mild steel sample were determined using dominant acid component of each agro fluids and were deduced as follow.



Scheme 1 Structural formula of malic acid ($C_4H_6O_5$)



Scheme 2 Structural formula of tartaric acid ($C_4H_6O_6$)

Apple and mango media contained malic acid as their dominant species. Malic acid is dicarboxylic (Scheme 1) and will dissociate partially to release two hydrogen ions [$2H^+$] per molecule of the acid. Subsequently one mole of hydrogen gas is evolved per mole of malic acid during electrochemical reactions (Eqs. 2 and 3).

Anodic reaction: oxidation of iron

$$Fe_{(s)} \rightarrow Fe^{2+}{}_{(aq)} + 2e^- \tag{1}$$

Cathodic reaction: dissociation of malic acid and subsequent evolution of hydrogen gas;

$$C_4H_6O_{5(aq)} \rightarrow C_4H_4O_5{}^{2-}{}_{(aq)} + 2H^+{}_{(aq)} \tag{2}$$

$$2H^+{}_{(aq)} + 2e^- \rightarrow H_{2(g)} \tag{3}$$

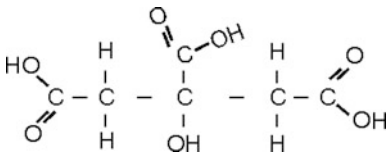
Overall reaction:

$$Fe_{(s)} + 2H^+{}_{(aq)} + 2e^- \rightarrow Fe^{2+}{}_{(aq)} + H_{2(g)} + 2e^- \tag{4}$$

Tartaric acid is also dicarboxylic as shown in Scheme 2 as the major dominant in grape juice; thus, two hydrogen atoms per molecule of the compound during electrochemical processes will be evolved. Hence, one mole of iron displaces two atoms of Hydrogen gas from the agro fluid to be oxidized to iron (II) oxide (Eqs. 5 and 6).

Anodic reaction: oxidation of iron

$$Fe_{(s)} \rightarrow Fe^{2+}{}_{(aq)} + 2e^- \tag{5}$$



Scheme 3 Structural formula of citric acid ($C_6H_8O_7$)

Cathodic reaction: dissociation of tartaric acid and subsequent evolution of hydrogen gas

$$C_4H_6O_{6(aq)} \rightarrow {C_4H_4O_6}^{2-}{}_{(aq)} + 2{H^+}_{(aq)} \tag{6}$$

$$2{H^+}_{(aq)} + 2e^- \rightarrow H_{2(g)} \tag{7}$$

Overall reaction:

$$Fe_{(s)} + 2{H^+}_{(aq)} + 2e^- \rightarrow {Fe^{2+}}_{(aq)} + H_{2(g)} + 2e^- \tag{8}$$

The structural formula of citric acid in orange juice (Scheme 3) shows three groups of carboxylic acid per molecule, this indicates that it is a tricarboxylic acid and exhibits more complex structure than malic and tartaric acids. The dissociation of citric acid causes displacement of 3 mol of hydrogen gas from 2 molecules of the acid as illustrated in Eqs. 10 and 11.

Anodic reaction: oxidation of iron

$$3Fe_{(s)} \rightarrow 3{Fe^{2+}}_{(aq)} + 6e^- \tag{9}$$

Cathodic reaction: dissociation of citric acid and subsequent evolution of hydrogen gas

$$C_6H_8O_{7(aq)} \rightarrow {C_6H_5O_7}^-{}_{(aq)} + 3H^+_{(aq)} \tag{10}$$

$$6{H^+}_{(aq)} + 6e^- \rightarrow 3H_{2(g)} \tag{11}$$

Overall reaction:

$$3Fe_{(s)} + 6{H^+}_{(aq)} + 6e^- \rightarrow 3{Fe^{2+}}_{(aq)} + 3H_{2(g)} + 6e^- \tag{12}$$

Distinctions can be drawn from the overall reaction mechanism (Eqs. 4, 8 and 12) to show which agro fluid would have most corrosive effect to the mild steel per molecule without considering other prevailing factors. One mole of Fe was displaced from malic and tartaric acids while 3 mol of Fe were lost in citric acid

medium. With respect to the number of evolved hydrogen atoms per molecule of each acid medium, citric acid possibly might have been more corrosive on the mild steel sample than malic and tartaric acids respectively. Therefore, orange juice was suspected to be corrosive more than other agro media as confirmed from the results of electrochemical corrosion study.

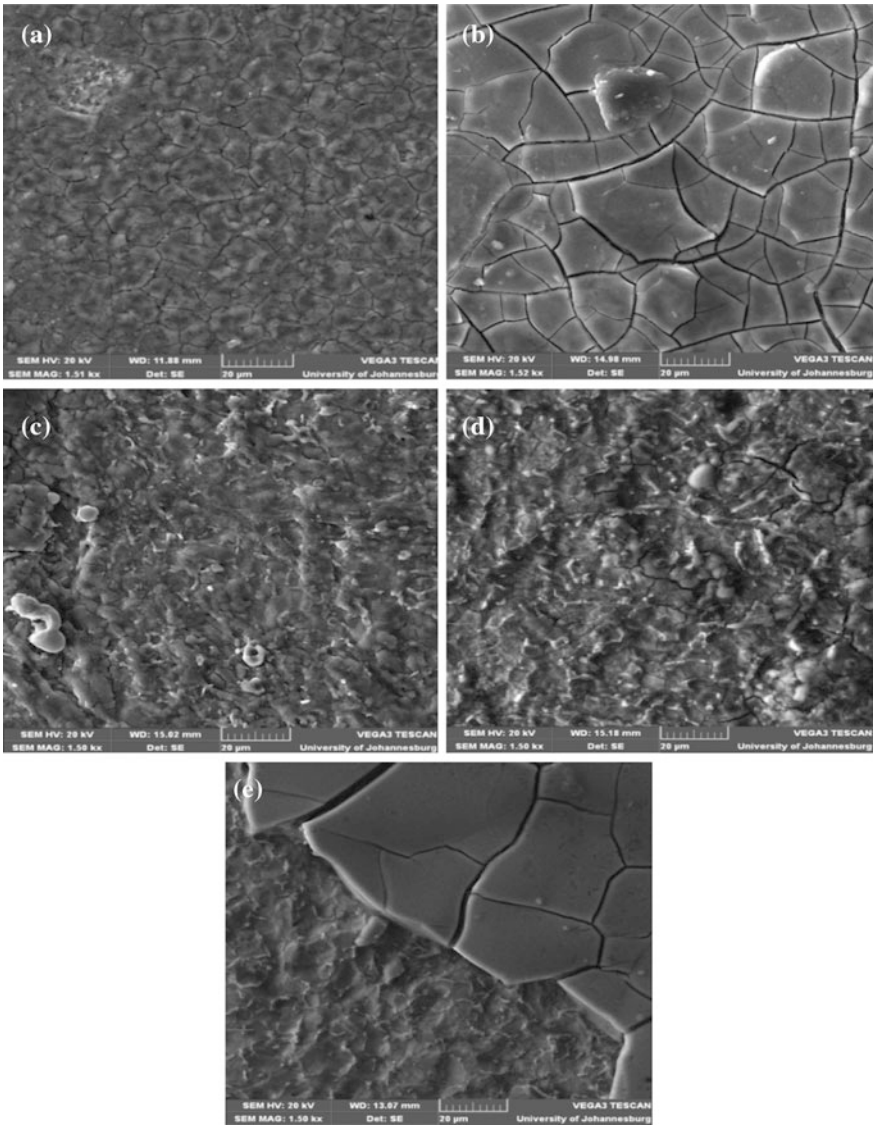


Fig. 3 Comparison of SEM surface morphologies of mild steel specimens immersed in **a** apple, **b** grape, **c** mango, **d** orange, **e** mixture juices at 20 μm

3.3 Scanning Electron Microscopic Analysis

SEM surface micrographs of the corroded samples indeed clearly depict varying degrees of oxide scales on the metallic samples. Corrosion mechanism and formation of crystal morphology on the metal specimens depend on corrosion products formed within the period of immersion [18, 25]. The various SEM surface morphologies of the corroded mild steel specimens show different degrees or sizes of passivating oxide films deposition. Intense deposited oxide layers were formed on the surfaces of mild steel specimens immersed in grape juice (Fig. 3b) and that of mixture medium (Fig. 3e), while specimen from mango juice (Fig. 3c) shows different structure of oxide formation across the surface. The micrograph of the metal sample immersed in orange juice (Fig. 3d) shows multiple film scales due to the nature of deposited oxide layers. The identified signs of rupture that occurred on the deposited oxides scales (Fig. 3d) may arise because of greater corrosiveness of the orange environment. Smaller sizes of oxide films with closed structures are observed on the surface of samples from apple (Fig. 3a) and mango juices (Fig. 3d) environments. The oxide film in Fig. 3a is much finer than Fig. 3d, possibly indicating better wear resistant to corrosion in apple juice than in mango juice.

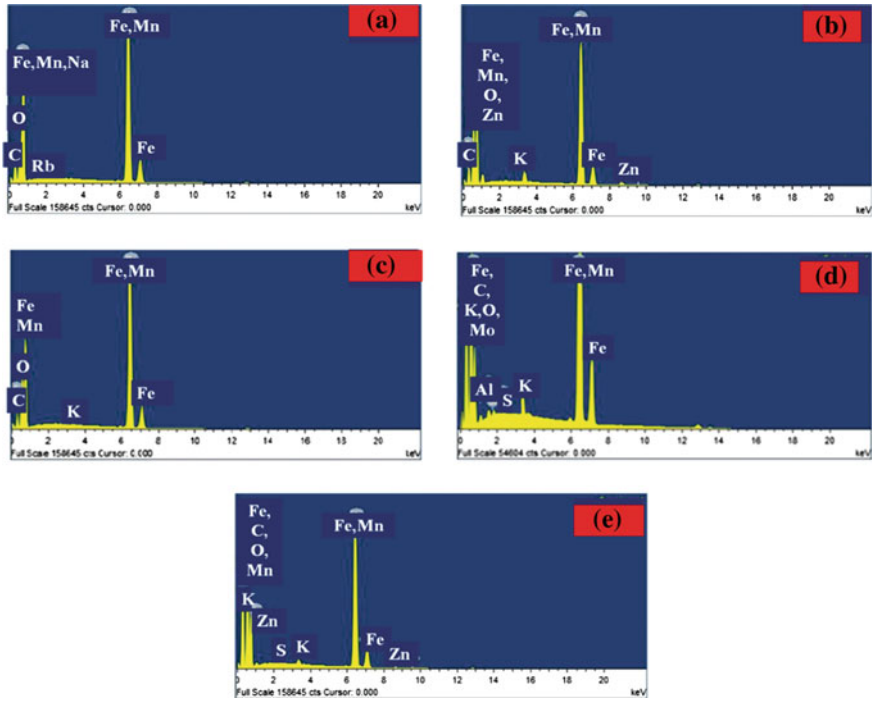


Fig. 4 EDS analyzes of mild steel in **a** apple, **b** grape, **c** mango, **d** orange, **e** mixture

Table 5 Correlation of CR and % weight of oxygen

Corrosion medium	Mixture	Orange	Grape	Mango	Apple
% weight of oxygen	24.29	20.01	19.47	14.44	11.25
CR (mm/year)	1.672	1.530	1.403	1.339	1.301

The elemental characterization of the corroded mild steel specimens in the agro juices after 60 days is viewed by energy dispersive X-ray spectroscopy (EDS). The results are presented in Fig. 4a–e.

EDS spectra in Fig. 4a–e show the various elemental composition shown on the surfaces of corroded mild steel specimens after sixty days of the immersion period. From the results, oxygen constitutes one of the major elements present in all the corroded samples. Oxygen may have been the main oxidizing agent in the agro juices environments which induced corrosive effect on the metallic specimens from the perspective of the study as supported by Porcayo-Calderon et al. [26]. The oxide content of the surface analysis suggests that a percentage weight of oxygen composition of the corroded specimen is proportional to the corrosion rate (CR) of the metallic alloy (Table 5).

Hence, the amount of oxygen deposited on each corroded specimens differs almost entirely. EDS analysis shows that percent weight composition of oxygen on the metallic surface from mixture medium is greatest (24.29 %), followed by metal sample immersed in orange juice with 20.01 % of oxygen deposit (Fig. 3d). Corroded metals in grape juice (Fig. 4b) and mango (Fig. 4c) contain 19.11 and 14.44 % weights of oxygen respectively. The least oxide deposit is observed on metal specimen in apple juice medium (Fig. 4a) with 11.25 % oxygen deposits.

4 Conclusion

Accelerated electrochemical corrosion behavior of mild steel in some selected agro fluids has been demonstrated. Corrosion rate (CR) of the mild steel in the agro fluids decreased progressively over the entire duration of the study. This was associated with greater formation of passivating oxide films. Reduction of hydrogen ions from the acid media and dissolved oxygen gas were also responsible for the decline in the rate of corrosion. Corrosion rate of the mild steel has the greatest impact in mixture medium (1.672 mm/year). But on individual basis orange juice was the most corrosive (1.530 mm/year) followed by grape juice (1.403 mm/year), mango juice (1.339 mm/year) and apple juice (1.301 mm/year), respectively at the last day of the immersion.

The significant increase in the pH values of the various agro fluids was a confirmation of a reduction in the acidity of all the test environments due to near or complete evolution of dissociable hydrogen gas from the corrosion media. The average pH values of various juices on the 60th day of immersion confirmed orange

juice might have exhibited highest corrosive effect than other individual agro fluids as a result of the additional acids present in the orange. Also, it can be associated with the greater amount of dissociable hydrogen ion in citric acid more than in malic and tartaric acids except for mixture juice which might have combined corrosive effect and subsequently corroding at greatest rate while apple juice had least corrosive effect. The result of microscopic analysis was proportionate to that of corrosion rates which showed that corrosion rate of the mild steel was greatest in mixture medium followed by orange juice, grape juice and mango juice while apple juice had the least corrosion rate.

Acknowledgments The authors gratefully acknowledge the technical supports rendered by the Metallurgical Engineering Department of the Tshwane University of Technology and Chemical Engineering Department of the University of Johannesburg, South Africa.

References

1. Hsu W, Yang C, Huang C, Chen Y (2005) Electrochemical corrosion studies on Co–Cr–Mo implant alloy in biological solutions. J Mater Chem Phys 93(2–3):531–538
2. Song G, Atrens D, Nairn J, Li Y (1997) The electrochemical corrosion of pure magnesium in 1 N NaCl. J Corros Sci 39(5):855–875
3. Shreir LL (2000) Basic concept of corrosion. In: Corrosion. Butterworth Heinemann. Great Britain, pp 74–95
4. Genesca J, Mendoza J, Duran R, Garcia E (2002) Conventional DC electrochemical techniques in corrosion testing. Technical report on corrosion, Department of Metallurgical Engineering, UNAN, City University 04510 Mexico D.F., and Instiuto Mexicano delPetroleo, Eje Central Lazaro Cardenas, 152.07730 Mexico
5. Frankel GS (2008) Electrochemical techniques in corrosion: status, limitations and needs. J ASTM Int 5(2):101–127
6. Bentiss F, Traisnel M, Gengembre L, Lagrenee M (1999) A new triazole derivative as inhibitor of acid corrosion of mild steel: Electrochemical studies, weight loss determination, SEM and XPS. J Appl Sci 152(3–4):237–249
7. ASM International (2000) Fundamentals of electrochemical corrosion (#06594G). Available: http://www.asminternational.org, 10 May 2013
8. Sharma KK, Sharma LK (eds) (1999) Physical chemistry, 9th edn. Viska Publishers, New Delhi
9. Khaled KF, Amin MA (2009) Corrosion monitoring of mild steel in sulphuric acid solutions in the presence of some thiazole derivatives—molecular dynamics chemical and electrochemical studies. J Corros Sci 51(9):1964–1975
10. Costescu C, Parvu D, Rivis A (2006) The determination of some physical–chemical characteristics for orange, grapefruit and tomato juices. J Agroalimentary Process Technol 12 (2):429–432
11. Wu J, Gao H, Zhao L, Liao X, Chen F, Wang Z, Hu X (2007) Chemical compositional characterization of some apple cultivars. J Food Chem 103(1):88–93
12. Niu L, Wu J, Liao X, Chen F, Wang Z, Zhao G, Hu X (2008) Physiochemical characteristics of orange juice samples from seven cultivers. J Agric Sci China 7(1):41–47
13. Ajila CM, Rao LJ, Prasada Rao UJS (2010) Characterization of bioactive compounds from raw and ripe *magnifera indica L.* peel extracts. J Food Chem Toxicol 48(12):3406–3411 (2010)

14. Toaldo IM, Fogolari O, Pimentel GC, Gois JS, Borges DL, Caliari V, Luiz MB (2013) Effects of grape seeds on the polyphenol bioactive content and elemental composition by ICP-MS of grape juice from vitislabrusca L. LWT-Food Sci Technol J 53(1):1–8
15. Li X, Yu B, Curran P, Liu S (2011) Chemical and volatile composition of mango wines fermented with different *Saccharomyces cerevisiae* yeast strains. S Afr J Enol Viticulture 32 (1):117–128
16. Fekry AM, Ameer MA (2011) Electrochemical investigation of corrosion and hydrogen evolution rate of mild steel in sulphuric acid solution. Int J Hydrogen Energy 36(17):11207–11215
17. Marcus P, Maurice V (1999) Structure of thin anodic oxide film formed on single-crystal metal surfaces. In: Wieckowski A (ed) Interfacial electrochemistry-theory, experiments and applications. Marcel Decker, New York, pp 541–558
18. Vander Voort GF (2002) Metallographic techniques in failure analysis, Buehler. J Metallographic Tech. 104738/6072, 3:1–17
19. Fonseca MP, Bastos LN, Caytuero A, Saitovitch EM (2007) Rust formed on Cannons of XVIII century under two environment conditions. Corros Sci 49(4):1949–1962
20. Afolabi AS, Ogazi AC, Adams FV, Abdulkareem AS (2014) Electrochemical behaviour of mild steel in some agro fluids, vol 2. Lecture notes in engineering and computer science: proceedings of the world congress on engineering and computer science, WCECS 2014, 22 Oct–24 Oct 2014, San Francisco, USA, pp 641–646
21. Poorsqasemi E, Abootalebi M, Peikari M, Haqdar F (2009) Investigating accuracy of the tafel extrapolation method in hcl solutions. J Corros Sci 51(5):1043–1054
22. Oluwole O, Olawale O (2010) Corrosion behaviour of nickel plated low carbon steel in tomato fluid. Leonardo Electron J Pract Technol 16:33–42
23. Oladele SK, Okoro HK (2011) Investigation of corrosion effect of mild steel on orange juice. Afr J Biotechnol 10(16):3152–3156
24. Tran T, Brown B, Nesic S, Tribollet B (2013) Investigation of the mechanism for acetic acid of mild steel. NACE international corrosion conference and expo 2013, Paper No: 2487, pp 1–12
25. Wang W, Jenkins PE, Ren Z (2012) Electrochemical corrosion of carbon steel exposed to biodiesel/simulated seawater mixture. J Corros Sci 57:215–219
26. Porcayo-Calderon J, Brito-Figueroa E, Gonzalez-Rodriguez JG (1999) Oxidation behavior of Fe–Si thermal spray coatings. J Mater Lett 38(1):45–53

Design and Characterization of a Model Fruit Juice Extracting Machine for Healthy and Vibrant Life in Today's Modern

Austin Ikechukwu Gbasouzor and Chika Anthony Okonkwo

Abstract The aim of this research is to design and fabricate pilot multi-tube boiler using a diesel fired burner $(C_{13}H_{25})_9$ to generate 80 kg of steam hour. The boiler tank is made of pure mild steel. Mild steel is used to fabricate the fire tubes and other parts such as the furnace, smokestack and return chamber that make up the boiler. The heating surface area was increased for sake of efficiency and fast steam generation by reversing the direction of the gas through a second and third parallel tube (three pass). The boiler (which is fired by a diesel burner) generates dry saturated steam at a pressure of 1.5 bars and temperature of 111.4 °C. It can be used for domestic and industrial purposes.

Keywords Conical resistor · Contamination · Conveyor · Detoxification · Extracting · Fruit juice · Potable · Slice · Sturdy · Unique

1 Introduction

Extraction is a process by which substance are removed for their original component or raw state. Hence, Extraction of juice (juice extraction) may be defined as the removal of juice from fruits; the juice is separated from the skin or chaff.

To achieve juice extraction, force is needed and the force depends on the biological nature and structure of the fruit from which the juice is to be extracted.

The pineapple fruit needs more force than other fruits for extraction because of its thickness of fruit skin. Fruits like banana need lesser force because of it light skin

A.I. Gbasouzor (✉)
Department of Mechanical Engineering, Anambra State University, P.M.B. 02, Uli, Nigeria
e-mail: unconditionaldivineventure@yahoo.com

C.A. Okonkwo
Xiamen University China and a Research Fellow in Electronics Development Institute (ELDI), Awka, Nigeria
e-mail: Okekenta2yes@yahoo.com

H.K. Kim et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-94-017-7236-5_32

nature. The fruit should be washed thoroughly and pretreatment should also be carried on fruit, proper filtration should be done on juice after extraction.

During the early age, juice extraction was done manually by the means of mouth sucking and hand squeezing. The storage of fruits does not last long due to lack of preventive measures; hence fruits were only available for consumption during harvest season. With this situation, there were no extraction and preservation of fruit juice for future uses. In today's world, pollution and chemicals come at us from all angles; our water contains toxic levels of metals and chlorine; our produce is tainted with pesticides; and our meat and meat products are pumped full of hormones and antibiotics.

It is small wonder that so many of us fall prey to chronic ailments and disease. So much so, that conditions like allergies, PMS, and migraines have become of common as a cold or the flu.

In order to live a healthy and vibrant life in today's modern, fast-paced world. It has become absolutely critical to drink, eat consciously biodynamic or organic whenever possible; freshly squeezed juice from fruits and vegetables are excellent source of minerals and vitamins that catalyst chemical reactions occurring in the body. These enzymes also produce the energy needed for digestion, absorption, and conversion of food into body tissues. An increased in take of fruit and vegetable juices ensures that the body will efficiently absorb more minerals and vitamins. Another helpful benefit of fruits and fruit juices is their ability to promote detoxification in human body. Fruit help to cleanse the body, especially those with high acid level. Tomatoes, pineapple, and citruses such as oranges, red grapefruits, and lemons are known for their detoxifying properties. While these fruits promote cleansing, they still provide the body with a high boost of vitamin C! Unfortunately, there are many people who do not eat sufficient quantities of fruit and other fresh produce on a daily basis- or even on a regular basis. This can lead to a lack of nutrients in the body that can have widespread ramifications on the whole body and your health. Fruit provides you with many great things for the body. Fruits have a lot of vitamins like A (especially apricots and cantaloupe) and vitamin C (especially citrus fruits like orange and grapefruit).

These two vitamins help heal cuts, assist night vision and create beautiful skin. They are also high in fibre. Fibre helps the stomach digest food and may help to reduce cancer. Most fruits have little fat.

The antioxidants present in the fruits also helps to protect the body from radical because high level of free radicals contribute to heart disease and some of the sugar present in these fruits are glucose, sucrose, and fructose with pre-dominant sugar varying in different fruit (Steminetz. K.A, Potter. J.D, vegetable fruits and cancer, epidemiologist mechanism, cancer causes).

All of these components are necessary in the building of optimum health and success, as well as combating and preventing illness and disease.

In the design and fabrication of the fruit juice extracting machine, careful selection of the most suitable materials are major concern. The skills knowledge and understanding of the importance of hygiene is also a valuable factor. These

Table 1 The sugar content of some ripe fruits

S/N	Fruits	Glucose	Fructose	Sucrose
1	Pineapple	2	1	8
2	Orange	2	2	5
3	Grape	8	8	0
4	Apple	2	6	4

factors call for selection of material, manufacturing and processing of the juice in order to increase the production quantity, reduce time and labour, also avoid food contamination (Table 1). The table below was adapted from Windowson E.M and P. A Malanie (1935) shows, the sugar content of some ripe fruits.

1.1 Problem Statement

Fruits are seasonal, they are not available throughout the year. Hence, the need to construct or fabricate a machine that can extract juice in order to enable availability and storage at all time. Establishing good and efficient design of the machine has been matter of concern. As the demand for fruit juice consumption increase from time to time around the globe, the need for designing and constructing a fruit juice extracting machine with high efficiency becomes necessary.

1.2 Significance of Research

Increase technology development and self-reliance of youths in Nigeria. Create self-development, self-worth and also enhance conversation of human energy, hence increase income.

Eliminates food poisoning and contamination by the use of corrosion resisting materials create employment for the skilled, semi-skilled and unskilled youths of this country.

1.3 Scope of Research

The scope of this research work ranges from:

- Proper material selection to ensure quality product.
- Proper measuring, cutting and welding of joints.
- Proper assembling of parts for construction of machines.
- Selection of juice yielding fruit to ensure optimum/proper running efficiency.
- Carrying out of machine analysis.

1.4 Limitations

This machine was designed mainly for the extraction of juice from succulent fruits like pineapple, orange etc. It cannot be used for extracting oil from oil bearing fruits or seeds like banana and mangoes etc. The machine is designed for squeezing of fruit not for pressing that is why fruits like mango and banana are not to be juice with it.

2 The Economic Importance of Fruit Juice Extractor

The citrus fruit juice extractor introduced in the early 40's commenced prior to the world war. The completion of the development was taken over by the central engineering laboratories in Tempe, Arizona. According to Akpan Justice. C. (2006) and Udomah Kingsly. S (2010). The first twenty four head rotary fruit extractor was completed in the mid 1946 after the completion it was experimented and tested with citrus fruit at the Sunkist exchange plant also in Tempe, Arizona. Operating problems encountered were noted and corrected and after the correction. The success obtained was encouraging.

In 1947 based on this encouraging performance, three more units were manufactured and operated commercially on orange at the Sunkist outaries California plant during the spring of 1947. Performance of these units was quite satisfactory and sufficiently encouraging.

In 1958, the pre-finishing arrangement was further perfected. This improvement involved the addition of the split into be used with the orifice tube, which provided pressure in the pre-finishing system. This improvement made possible if reduced diameter strainer tube internal bore during each stroke of the extractor, assuming maximum efficiency under operating conditions. Note the orifice discharge tubes at bottom which remove core membrane and seeds from juice at the moment of extraction. According to Akpan Justice. C. (2006) and Udomah Kingsley. S. (2010).

A model 49′5″ cup heated machine was produced in 1964 to handle the major portion of the grape fruit which was produced in 1958. Another extractor in the 91′ model series was introduced on 1972 this unit known (3-318′) cups suitable for processing the smaller size line, then in 1972 a plant in Brazil designed and constructed a 100″ extractor and that market the beginning of refining program. The previous models were redesigned to encompass the following changes.

The adoption of the top loading lower cutters to reduce the time required changing the components.

Whereby the covers were changed to fiberglass or stainless steel.

A stainless frame and cast stainless and two sides leg members. This change was made primarily to permit use of caustic cleaners to enhance cleaning and improve components life and all exposed aluminum parts where not changed to stainless cleaners.

The feed hopper designed was improved in 1976 to move efficiently transfer fruit, fruits from the tilted feed belt to a feed hopper while this modification was of importance to all model extractors. It was of greater value to the efficiency of high speed machines because of the grantor volume (up to recycle peel oil recovery systems) of being processed.

Extra improvements were done on the machine by using gear trains instead of the belt drive. As can be seen from the initial inline underwent three major model changes which were designated as the 91′ series than the 91A′ series and the 91B′ series. Even though some similarity in appearance to the original machine remains, it is obvious from the many mechanical changes that this extractor has been substantially improved from the data it was manufactured.

Again, the fruit juice extractor and sterilizer currently being designed with the aim of improving fruit juice extractor sterilizer. The machine durability, versatility and capacity were the main concentration and also they solve all defects and limitations forced by fruit juice extracting machines available in the market. In today's market, there are various brands or types of extractors available for the buyer or customers. From a market survey carried out the prices varying slightly different from market to market while departmental stores are more expensive.

2.1 Types of Fruit Juice Extractor Continuous Crushing Process Using Cam

The continuous crushing process was introduced during a science and engineering exhibition in Europe. It consists of the hopper, power unit, cam, follower and metal crusher with nail.

In order for the continuous crushing process machine to extract juice from orange or pineapple, it exerts impact shear and compressive force. The metal crusher incorporated in the machine crushes continuously the fruits that are fed into the machine by crushing it against a metal surface which separates the juice which is collected through a special opening. The wastes are forced out through the waste collector.

The motion of the metal crusher is achieved by the use of cam and follower, which is connected to an electric motor by the means of belt and pulley. The motion of the cam and the follower is such that it converts and transmit the rotating motion of the motor to reciprocating motion of the crusher. Its advance design is done using a gear instead of the belt and pulley.

2.2 The Basket Press

The basket press consists of a horizontal perforated cylinder, a hinged cover, piston with circular plate, paddle, a hydraulic system or mechanical press and a power unit.

The basket press needs to exerts impact shear or compressive force in order to extract juice, the incorporated paddle exerts both impact and shear force on the fruit. This is achieved as the paddle agitates the fruits, it hits and rotates them continuously, introducing a shear and impact force on the fruit. The mesocarp is then softened and separated from the skin. Compressive force can now be exerted by the hydraulic system. The piston force the fruits against the basket and the result is the juice being compressed out of the mesocarp.

2.3 The Continuous Screw Process

The continuous screw process consist of the conical barrel, the hopper, the screw shaft, thrust bearings, and the power unit. The force needed to extract juice from the fruits are applied by the screw shaft. The screw shaft also conveys the fruit and waste to the outlet hence, this process is a continuous process.

The screw shaft which is a mechanism device consist of a cylindrical or conical body around which is formed a spiral rib or thread impact and shear forces are applied on the fruits as the screw conveys them forward. Since the process is continuous, thee screw shaft units the fruits against the wall and also rotates them along. The fruits are moved forward in a manner similar to that in which prevented from rotating receive translator motion (advances) when the screw fitted into it is turned. In this case the fruits are prevented from being turned with the screw due to their weight and the friction between the fruit and the walls of the barrel.

Compressive forces are applied by the screw on the fruits against the wall. Due to the conical shape of the collector as the fruit moves down the length of the machine, thus the fruit movement shows down towards the outlet producing a compressive effect on the fruits due to gradual decrease in barrel diameter.

The continuous screw process is chosen in preference to all others because it is cheap to produce. Moreover it requires less labor since it operates on a continuous means.

2.4 Fruit Juice Extracting Machine Cocoa Juice Extractor

This semi-continuous cocoa pulp extractor has a horizontal cylinder. The rotation blades are mounted on the central axis and this remove pulp seeds. The machine has a through put of 99–180 kg per hour.

2.5 Fruit Presses

The fruits placed into the machine through the hopper. A handle attached to the machine if turned, presses the fruit and extract the juice.

2.6 Hydraulic Vine Presses

This is manually operated by a press and it extracts juice from the fruits through the hydraulic pressure.

3 Methodology Design of the Fruit Juice Extracting Machine

In this design of fruit extracting machine in this work many things were considered when analyzing the system.

3.1 Parts Design and Material Selection

Jain RK [3] described manufacturing processes as the processes involved in using various construction methods in producing the extracting machine. In manufacturing, the principal common characteristics 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.

The processes involves in producing the machine are as follows:

- Marking out operations or procedures
- Cutting operations or procedures
- Assembling operations
- Welding operation
- Machining operation

3.2 Marketing Out Operations/Procedures

This is done to get the required shape and size of the design according to our dimensions in order to meet out expectation or aim. It is done or carried out by using tapes, marker, squares, vernier caliper etc.

3.3 Cutting Operations/Procedures

- Power saw: for cutting of thick pipes and circular bars.
- Hacksaw: for cutting of rectangular plates and circular bars.
- Emery cloth: for smoothing and polishing of rough edges of wood.
- Chisel and hammer: for cutting of casing of the extractor unit.
- Guillotine machine: for cutting of mark out shit of stainless steel and mild steel into the required dimension or measurement.

3.4 Assembling Operations/Procedures

This aspect is bringing together of all required part components to form a unit or a complete machine.

3.5 Welding Operations/Procedures

This process is the system of using electric welding and electrode to join the art material into shape.

3.6 Machine Operations/Procedures

The operating shaft is machine using the lathe to get the required diameter of the shaft. The machining operation is also adopted for the manufacture of component. These are listed below:

- Grinding operation
- Turning operation
- Milling operation
- Drilling operation

3.7 Grinding Operation

This operation is a high precision metal cutting operation done with an abrasive grinding wheel, and the milling cutter. This operation is applied during the finishing of the work.

3.8 Turning Operation

Is the process of machining round stock on a turning machine, it is a process by which work piece rotates a fixes tool and the most common turning machine is the lathe. Turning operation turned main shaft, which is the extractor shaft to the required diameter in the lathe.

3.9 Milling Operation

The milling operation is used in machining which removes metals by feeding the work piece against the rotating cutter. This was carried out to the milling machine, and its application was made use of in manufacturing gears, main shaft, key ways, plunger's teeth etc.

3.10 Drilling Operation

The drilling operation is the process of making hole and also enlarging existing hole. Drilling is used to make hole in the filter chamber and some of the work frame and also the position of some parts like electric motor, collecting tank, and gears etc.

3.11 Material Selection and Components Used

The following are the materials selected for the components of the machine being produced.

3.12 Stainless Steel

The shaft, filter and outer chamber, hoppers and conveyor are made of stainless steel.

3.13 Properties of Stainless Steel

- It is hard and strong substance
- It is not a good conductor of heat and electricity
- It has high ductile strength

- It does not get oxidized easily
- It is highly resistant to corrosion
- It is capable of retaining its strength.
- It possess magnetic permeability

3.14 The Feed Hopper

This is where the fruits are being put into the machine, the feed hopper is made of stainless steel.

3.15 The Slicing Blade

The slicing blade is made of stainless steel and is used to cut the fruit into smaller portion before going to the conveyor or screw conveyor.

3.16 The Conveyor Shaft

It is a solid circular and made of stainless steel that carries the die (blade) that do the cutting through the rotation of the shaft by transmission of power from the electric motor through the help of gears. It has a length of approximately 560 mm and hinged at both ends by a house of bearings.

3.17 The Juice Collector

The juice collector is made of transparent glass and is used to collect juice from the filter chamber and it is located under the cylindrical chamber.

3.18 The Waste Collector

This is a channel where the chaff is being collected by opening the slab lock for the chaff to fall off into a collector chamber.

3.19 The Electric Motor

The motor supplies power to the gear which transmit power to the rotating shaft. The motor has a speed of 1250 rpm and a capacity of 5 kw.

3.20 Gear Train

The gear train to transmit power from the electric motor to the shaft that rotates the cutter and conveyor.

3.21 Advantages of Gear to Belt Drive

- It transmit exert velocity ratio
- They transmit large power than the belt
- It has high efficiency
- It has reliable service
- They have compact layout.

3.22 The Barrel

The filter inside the barrel/cylinder of the pressing section is used to separate the juice from the chaff when pressed.

3.23 The Ball Bearings

The bearings are used in order to prevent the shaft from wobbling. The bearings used are ball bearing and they are four in number, and coupled to the shaft at both ends.

3.24 The Main Frame

The main frame is made of mild steel. It supports and carries the weight of the machine and its components.

3.25 Outer Cylindrical Chamber

It is built or constructed with stainless steel and contains the shaft that carries the die extracting. It is the cylindrical chamber where the grinding is done and it has a length of approximately 700 mm.

3.26 Operation of the Extracting Machine

After the preparation of the fruits, which involves the peeling and slicing of the fruits, the machine is now being powered on through the control switch that powers the motor and the sliced fruit is being fed in through the hopper. The shaft powered by the gear train rotates anti-clock wise and the slicing blade will slice the fruit and send it to the conveyor where the squeezing will take place through the help of screw conveyor and the filter chamber helps filter the juice from the chaff.

The juice now passes through the hole of the cylindrical chamber of the conveyor and through the filter chamber then enters the juice collecting tank while the chaff moves to the barrel where they fail into the chaff collector.

The machine is a continuous process that is, the feeding of fruit, squeezing, extracting and drawing of chaffs are done continuously.

3.27 Fruits Juice Extractor Description

The juice extracting machine consists of various components and its will be great important to have the detailed description illustrated (Figs. 1 and 2) (Table 2).

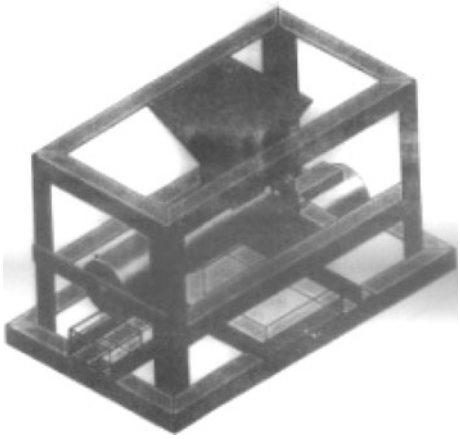


Fig. 1 The designed fruits juice extracting machine

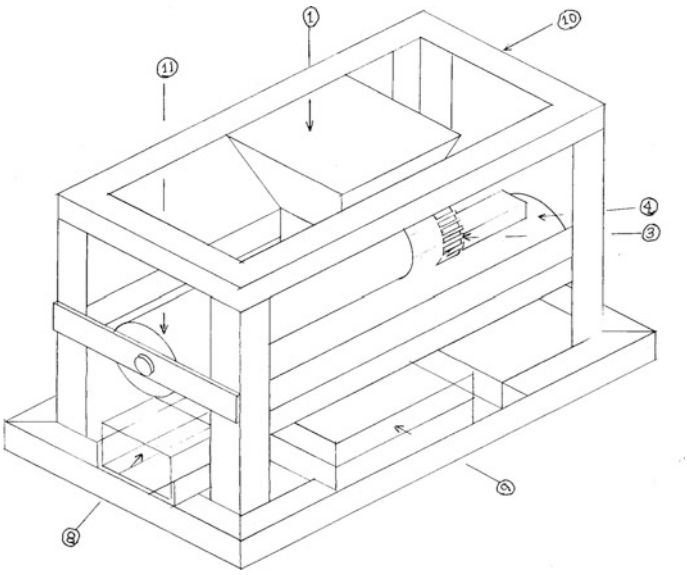


Fig. 2 The technical view of the fruits juice extracting machine

Table 2 Various components of fruit juice extracting machine

Item	Description	Item	Description
1	Feed hopper	2	Slicing blade
3	Conveyor shaft	4	Juice collector
5	Waste collector	6	Electric motor
7	Gear train	8	Barrel
9	Ball bearing	10	Main frame

3.28 Pre-Treatment of Fruits

The pre-treatment of the fruits are done to obtain an effective extraction of the juice. It is done to meet up the expectation quality and taste being needed. The processes involved in the pre-treatment are.

3.29 Fruit Selection

The fruits that are being selected are fresh fruits in order to obtain a high quality fruits juice. A small quality of bail fruit may change the taste of the whole batch of the juice extracted.

3.30 Washing

The fruits should be washed with clear pure or chlorinated water to prevent dirt or unseen micro-organisms before extracting for good healthy juice.

3.31 Fruit Preparation

This process involves the peeling and slicing of the fruit under the most hygienic conditions. Plastic or wooden table should be used to avoid contamination of the fruits by germs.

4 Design Analysis of the Cylinder Chamber

The proof of theorems and derivation of formulas are included in solving the problems. Allen et al. [1] State that

Given

$$\text{Length of the cylindrical chamber} = 25\,\text{cm}$$

$$\text{Diameter of cylindrical chamber} = 9\,\text{cm}$$

$$\text{Volume of the cylindrical chamber} = \text{R2L}$$

Consider

$$\begin{aligned} &\text{R} = \text{radius} \\ &\text{L} = \text{length} \\ &\text{R} = \text{d}/2 = 0.09\,\text{m}/2 = 0.045\,\text{m} \\ &\text{L} = 0.25\,\text{m} \end{aligned} \tag{1}$$

Substituting the value

$$\text{V} = \text{R2L} = 3.142(0.045)^2\text{R}0.25$$

$$\text{V} = 1.59^{-3}\ \text{m}^3$$

4.1 Drive Mechanism

Suppose

$$\text{Diameter of motor gear (D)} = 4.3\,\text{CM}$$

$$\text{DIAMETER OF MACHINE GEAR (Dc)} = 12\ \text{CM}$$

$$\begin{aligned}\text{Motor speed } (N_1) &= 1250\,\text{rpm}\\ \text{Machine speed } (N_2) &= X\end{aligned} \tag{2}$$

Recall that

$$N_2 = \frac{N_1 D_1}{D_2} = \frac{1250 * 43\,\text{mm}}{120\,\text{mm}}$$
$$N2 = 447.9\,\text{rpm} = 448\,\text{rpm}$$

$$\text{Torque developed between the rotating shaft} = Fc \times R$$

$$Fc = \text{centrifugal of rotation} = MW^2R \tag{3}$$

where

$$\begin{aligned}M &= \text{mass of rotation}\\ N &= \text{radius of rotation}\\ W &= \text{angular velocity of rotation}\end{aligned}$$

$$\begin{aligned}W &= \frac{2\pi N}{60} = \frac{2 \times 5.14 \times 448}{60}\\ &= 46.92\,\text{rad/sec}\\ M &= w/g\end{aligned}$$

where
G acceleration due to gravity
M mass of shaft, gear, and conical restriction on shaft

Total weight of these components are estimated at 75 kg
The weight of fruit processed is estimated at 5 kg

$$\begin{aligned}\text{Total weight} &= 7\,\text{kg} + 5\,\text{kg} = 12\,\text{kg}\\ M &= w/g = 12/9.81 = 1.22\,\text{kg}\end{aligned} \tag{4}$$

Recall that

$$\begin{aligned}Fc &= MN2R\\ Fc &= 1.22 \times (46.92)2 \times 60\,\text{mm}\\ &= 161148.8045\\ &= 161.15\ \text{N}\end{aligned}$$

$$\begin{aligned}\text{Torque T} &= \text{fc} \times \text{R} \\ &= \frac{161.15 \times 60}{1000} \\ &= 9.67\,\text{N/M}\end{aligned}$$

4.2 Power Developed

$$\text{P} = \text{T} \times \text{W} \tag{5}$$

Substitution

$$\begin{aligned}\text{P} &= 9.67 \times 46.92 \\ &= 453.7\,\text{watts} \\ \text{P} &= 453.7\,\text{watts or } 0.45\,\text{kW}\end{aligned}$$

Recall that

$$\begin{aligned}1\,\text{hp} &= 750\,\text{watts} \\ \text{X hp} &= \frac{453.7 \times 1}{750} \\ &= 0.6\,\text{hp}\end{aligned}$$

Motor power for our design is 0.6 hp.

4.3 Bending Moment and Shaft Design

The schematic representations of force acting on the shaft are as follows (Figs. 3 and 4).

4.4 Recall

$$\begin{aligned}&1\,\text{kg/F} = 10\,\text{N/M} \\ &12\,\text{kg/F} = 120\,\text{N/M} \\ &\text{Therefore } 50\,\text{N/M} = 45\,\text{N} \\ &125\,\text{n/m} = \text{XN} \\ &\text{XN} = \frac{120\text{N/M} \times 45\,\text{N}}{50\,\text{N/M}} \\ &\quad = 108\,\text{N}\end{aligned}$$

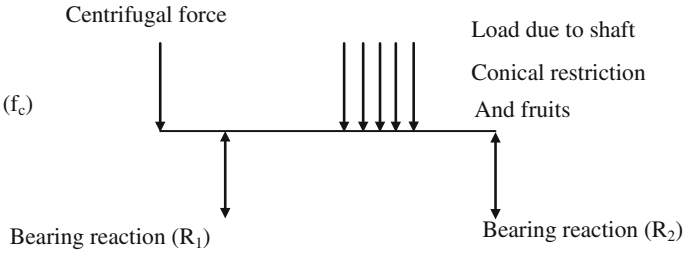


Fig. 3 Load due to shaft

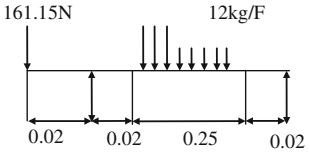


Fig. 4 Force acting on the shaft

Fig. 5 Load distribution change

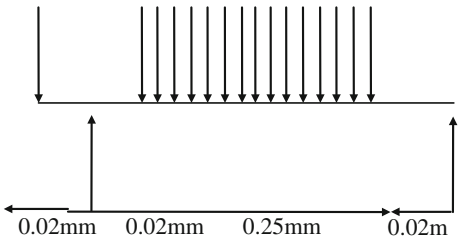
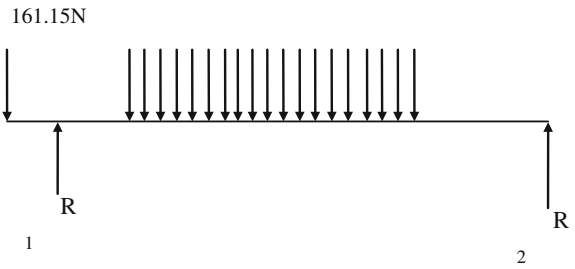


Fig. 6 Torque developed rotating shaft



For ease calculation the distribution local will be changed to a point load (Figs. 5, 6 and 7)

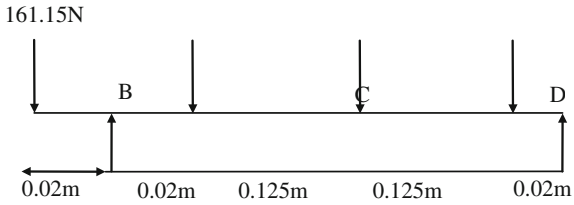


Fig. 7 Upward and downward balance of the force

Upward vertical force must balance downward vertical force

$$\begin{aligned} R_1 + R_2 &= 161.15 + 108 \\ R_1 + R_2 &= 269.15 \end{aligned} \tag{6}$$

Taking moment about D

$$\begin{aligned} &-(161.15 \times 0.31) + R1 \times 0.29 - 108 \times 0.125 = 0 \\ &= -46.86 + 0.29\,R1 - 13.5 = 0 \\ R_1 &= \frac{60.36}{0.29} = 208.13\,N \\ &0.29\,rl - 60.36 = 0 \end{aligned}$$

Substituting into

$$\begin{aligned} R_1 + R_2 &= 269.15 \\ 208.1 + R_2 &= 269.15 \\ R_2 &= 269.15 - 208.13 \\ R_2 &= 61.02\,N \end{aligned}$$

Representation of the shaft with all forces we have (Fig. 8)
Maximum bending moment, of shaft moment A to B

$$\begin{aligned} M_{A-B} &= -161.15 \times 0.02 \\ &= -3223\,N/M \end{aligned} \tag{7}$$

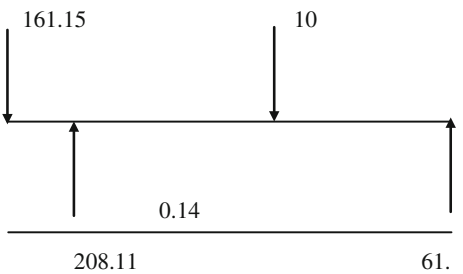


Fig. 8 Bending moment of the shaft at maximum

$$M_{B-c} = 161.15(0.02 + a) + 208.13(a)$$

where a = 0.145 M

$$\begin{aligned} M_{B-C} &= 161.15(0.02 = 0.145) + 208.13(0.145) \\ &= -28.30 + 30.179 \\ &= 1.979\,\text{N/M} \end{aligned}$$

$$\begin{aligned} M_{C-D} &= 161.15(0.165 + a\ ) + 208.13(0.145 + a\ ) - 108(a) \\ &= 49.957 + 60.358 - 15.66 \\ &= 5.26\,\text{N/M} \end{aligned}$$

The maximum bending is at moment C.D with value 5.26 N/M

4.5 Bending Moment Diagram

See Fig. 9.

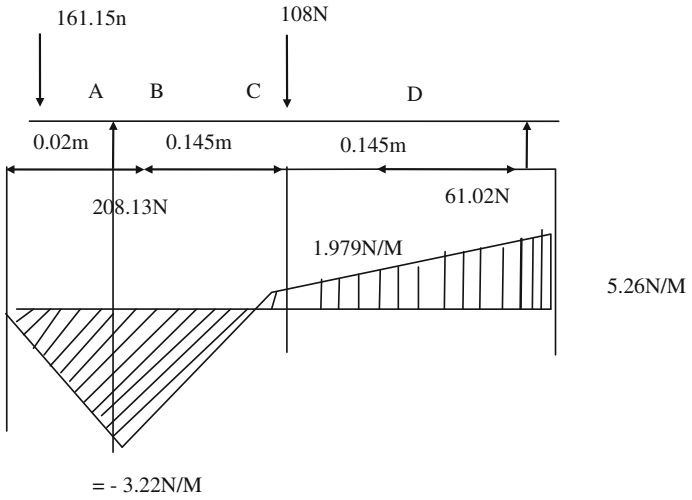


Fig. 9 Shaft bending moment

4.6 Shaft Diameter

Formula for shaft diameter

$$D = \left\{ \frac{32n}{\pi} (Km.Mmax)^2 + (Kt.T)^{2)1/2} \right\}^{1/3}$$

where

Km combined stock and fatigue for bending = 1.5
Kt combined stock and torque for tension = 1.0
Mmax maximum bending moment = 5.26 N/M

$$T = \text{torque} = 9.67\,\text{N/M}$$
$$N = \text{factor of softy} = 1.0$$
$$D = \left(\frac{32 \times 1}{3.142} (1.5 \times 5.26 \times 1000) \right.$$
$$D = (10.184(62252100 + 93508900)^{1/3}$$
$$D = \left(10.184(62252100 + 93508900)^{1/2}\right)^{1/3}$$
$$D = 10.1846(155761000)^{1/3}$$
$$D = (10.1846 \times 12480.42)^{1/3}$$
$$D = (127108.0855)^{1/3}$$
$$D = 50.28\,\text{mm}$$

4.7 Pressure Developed at the Conical Restriction

$$\begin{aligned} \text{Area of the cylinder section Ac} &= 2\pi RL \\ &= 2283.142 \times 0.09 \times 0.25 \\ &= 0.14139 \\ Ac &= 0.14\,m^2 \end{aligned} \tag{8}$$

$$\text{Area of cone inside Ak} = \frac{\theta \times 2\pi 1}{360} \tag{9}$$

$$\frac{\text{Sin}\theta}{2} = \frac{\text{Opp}}{\text{Hypo}}$$
$$\frac{\text{Sin}\theta}{2} = \frac{\text{r}}{\text{r}} = \frac{90}{250}$$
$$\frac{\text{Sin}\theta}{2} = 0.36$$
$$\frac{\theta}{2} = \sin^{-1} 0.36$$
$$\theta = 2 \times 21.1 = 42.2^{\circ}$$

$$\text{Ak} = \frac{42.2}{360} \times 2 \times 3.142 \times 0.25$$
$$= 0.184\,\text{m}^2$$

$$\text{Pressing area} = \text{Ac} - \text{Ak} \quad (10)$$
$$= 0.14 - 0.184$$
$$= -0.044\,\text{m}^2$$

Pressure developed in the barred

$$\text{P} = \frac{\text{force}}{\text{Effective area}} = \frac{235.14}{0.044} \quad (11)$$
$$= 5344.09\,\text{N/M}^2$$
$$= 5.34\,\text{KN/M}^2$$

4.8 Conveyor Delivery Velocity

$$\text{Conveyor's delivery min} = \text{rpm} \times \text{pitch} \quad (12)$$
$$\text{Pitch if design} = 100\,\text{mm}$$

$$\text{N}^2 = 447.9\,\text{rpm}$$
$$\text{Delivery rate} = 100 \times 447.9$$
$$= 44790\,\text{mm/min}$$

Therefore

$$\frac{\text{Velocity}}{60} = \frac{44790}{60}$$
$$= 746.5\,\text{mm/sec}$$
$$= 0.7465\,\text{m/sec}$$

4.9 Bearing Selection

Ball bearing is being used for this design because of the following.

- They can be able to withstand thrust action in the machine
- It serves as means of transferring loads between rotating area and stationary machine member.
- They can accommodate both radial and thrust load.
- They are easy to repair and assemble.

4.10 Gear Train Value

N_1 speed of gear (driver) in rpm
N_2 speed of gear (driver) in rpm
T_1 number of teeth on gear (driver)
T_2 number of teeth on gear (driver)

$$\text{SPEED RATION} = \frac{N_1}{N_2} = \frac{T_2}{T_1}$$

Recall: the ratio of the speed of the driver to the speed of the driver is known as train vale of the gear train.

$$\text{Train value} = \frac{N_2}{N_1} = \frac{T_1}{T_2}$$

where

$$N_1 = 1250\,\text{rpm}$$
$$N_2 = 447.9\,\text{rpm}$$
$$T_1 = 22\,\text{teeth}$$
$$T_2 = 65\,\text{teeth}$$

$$\text{Train value} = \frac{447.9 \times 65}{1250 \times 22}$$
$$= \frac{29113.5}{27500}$$
$$= 1.059$$

$$\text{Train value} = 1.06$$

Notice: Belt or Rope has a slipping factor on transmission of motion or power between two shafts. And effect of this slipping is due to velocity ratio of the system. Gear tooth helps in avoiding slipping on frictional wears. According to Allen et al. [1].

4.11 Testing and Analysis

Testing and Evolution were carried out at the Completion of the construction of the fruit juice extraction machine. The tests were carried out at the designed operation speed of 125 rpm.

However, the tests were carried out so as to determine the following.

- The rate of extraction and capacity of machine.
- To determine whether there is leakage in machine.
- To determine the hours or minute per hour the machine can operate.
- To determine the percentage and the efficiency of the machine.
- To determine number of hours, the operator or engineer can operate on it, in order to know the accurate time to dismantle the conveyor or conical restriction section (Table 3).

Thus the number of pineapple that can fill in the filter cylinder is twenty one (21) pineapples, output of about 14 Litres machine capacity = Output (in litres)

$$\text{Operating time capacity} = \text{output (in litres)}$$

Table 3 Determination of the machine capacity

Tests no.	Operating time (mins)	Name of the fruit	No. of fruit	Output (litres)
1	7	Pineapple	5	3
2	13	Pineapple	7	4.5
3	17	Pineapple	9	7

Operating time

$$\text{Total quantity of juice extracted} = 14.5\text{l}.$$

$$\text{No of tests} = 3$$

$$\text{Machine capacity} = \frac{14.5\text{l}}{2}$$

$$\text{EFFY} = \frac{\text{Max no. of juice extracted(l)}}{\text{Max. no. of pineapple}}$$

$$= 14.5/21 = 0.69 = 69\,\%$$

5 Conclusion

This research work has successfully presented a functional and highly efficient low cost fruit juice extracting machine by minimizing local technique of squeezing and sucking of fruits, hence improving the hygienic and health condition of individuals and maintain fluid balance in the body.

The machine can be used for fruit juice drink, soft drink, in restaurant, brewel and bakery industries. The machine is also design for home and industrial usage. The machine is versatility, durability, and the capacity were the main area of concentration.

Though after the design and fabrication of the machine, some lapse were notice the machine could not remove all fruits sediments after extraction and requires further filtration. Some of it's components were manufactured locally, it can be easily fabricated, reproduced and maintained without fear of spare part during maintenance.

Acknowledgment With due respect Sir, let me seize this opportunity to thank 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, I salute your efforts.

Recommendations

In the research work performance test and result analysis has been carried out, modification still needs to be done on the machine in order to increase machines efficiency. The modifications are as follows;

There are some sharp edges on the machine which can lead to injury during operation. Thus, it should be made smoother or the sharp edge of the machine should be curved to increase the safety of the machine operator or user.

The conveyor shaft should be made longer so that much of the extraction can be done before3 getting to the withdrawal end as some of the juice out with the roughages.

The waste collector should be made of stainless steel for good food handling because the chaff when dried properly can be preserve or used as flavoring in baking and beverage industries.

The slicing blade should be sharp and made of stainless steel in order to slice the fruit into smaller sizes before entering the conveyor for proper squeezing of fruit. The fruit should be heat treated to kill bacteria to avoid health hazard. Citrus containing PH level less than 4.0 can be pasteurized by heating the juice to 90 °C for a few seconds.

The outer and inner chamber should be more thinner in order to filter the juice and separate the shift from entering the juice collector through the help of the outer filter chamber.

References

1. Allen SH, Alfred RH, Herman GL (2002) Schaum's outline series of theory and problems of machine design, SI (Metric) edition chapter 8, 10. Published by Tata McGraw-Hil Publishing Company Ltd, New Delhi
2. Gbasouzor AI, Okonkwo AC (2012) Improved mechanized fruit juice extracting technology for sustainable economic development in Nigeria. In: Proceedings book of the world congress on engineering and computer science. Lecture notes in engineering and computer science, WCECS 2014. 22–24 Oct 2014, San Fransico, USA, pp 953–960
3. Jain RK (2007) Production technology. 16th edn. Published by Romesh Chander Khanna for Khanna Publishers, Delhi
4. Khumi RS, Gupta JK (2008) Theory of machines 14th revised edition. Eurasia Publishing House (PVT) Ltd, New Delhi

Efficient Operational Management of Enterprise File Server with User-Intended File Access Time

Toshiko Matsumoto and Takashi Onoyama

Abstract Toward efficient management of enterprise file server, listing potentially deletable files is expected to be effective. For this purpose, we propose a method of extracting frequent file access pattern to sort out files accesses by operating system or of file system, because such access patterns are supposed to be highly frequent. In our method, each access sequence of one user ID at short time intervals is processed in three steps: (1) path is encoded on the basis of relative position in folder tree structure, (2) accesses are clipped into independent sub-trees, and (3) statistical significance of number of observation is calculated according to occurrence probability. In this paper, we apply our method to file server access logs of a company and demonstrate that it can distinguish system-driven access accurately by utilizing our method to calculate user-intended access time, it can contribute to make convincing list of deletable files.

Keywords Access · File server · Folder tree · Frequent pattern · Operational management · Relative position

1 Introduction

Recently growing trends of unstructured files in enterprise organizations has received more and more attention, as digitalization of document increases volume of data stored in file servers [1]. In addition to de-duplication [2] and Information Lifecycle Management [3, 4], deleting unnecessary files is expected to be an effective and also radical way to reduce volume of file servers and to make operational management of file server more efficient. Although a file with old access time

T. Matsumoto (✉) · T. Onoyama
Hitachi Solutions, Ltd., 4-12-7, Higashishinagawa, Shinagawa-Ku, Tokyo 140-0002, Japan
e-mail: toshiko.matsumoto.jz@hitachi-solutions.com

T. Onoyama
e-mail: takashi.onoyama.js@hitachi-solutions.com

H.K. Kim et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-94-017-7236-5_33

is generally considered as deletable [5, 6], value of last access time is set not only by user-intended access but also by system-driven access: operating system performs crawling access against all objects (files and folders) in a subtree in order to execute a folder access, file system makes cache of a file and divides one access for a large file into multiple accesses, or an application program reads multiple setting files at a time in its start-up process [7]. Since such system-driven accesses are highly frequent, they may bury characteristics of user-intended accesses. Therefore, user-intended accesses are required to be distinguished from system-driven accesses and to be extracted for more convincing candidate list of deletable files. When access time is re-calculated from user-intended accesses, i.e. other than system-driven accesses, the access time accurately represents how recently the file is utilized. We proposed the "Relative-Position based Frequent Access Mining (RPFAM) method", which extract frequent access patterns from access log of enterprise file server [8]. In this paper, we demonstrate accuracy in extracting system-driven accesses and effectiveness in making list of potentially unnecessary files by applying RPFAM method to access log of enterprise file server.

2 Problems in Extracting Frequent Access Pattern

For convincingly listing deletable files, frequent access patterns are required to be extracted as supposed system-driven accesses, and accesses time is required to be re-calculated from other accesses, i.e. supposed user-intended accesses (Fig. 1) In addition to analysis on access pattern of each file [4, 9, 10], several methods have been proposed for mining frequent pattern in sales records or in web access log for each user [11–15]. However, there are three technical problems in applying these methods to extracting frequent pattern from access log of enterprise file servers.

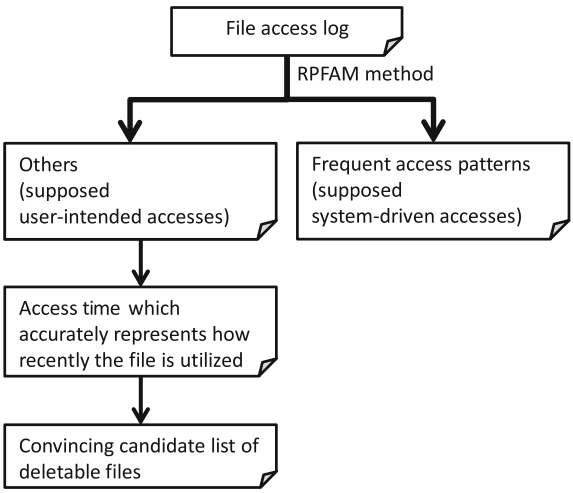


Fig. 1 Process overview of extraction of frequent access pattern toward candidate list of deletable files

First problem is about number of objects. Analysis of web access log deals with relatively small web sites, because it distinguishes each page as an independent object. As opposed to the case of web access log, analysis of file server access log is required to consider up to millions of objects or more. When each file or folder is analyzed as an independent object, number of observation of access could be too small to conclude meaningful results. Furthermore, common access characteristics could not be extracted from absolute file path when a folder subtree is assigned to a department and each user accesses to files in the subtree assigned to his/her department or business role. For example, suppose that a department provides several products, a subfolder is assigned for each one of products, and product folder contains marketing materials, i.e. catalogs, best practice sheet, comparison sheets and so on. When a sales person visits a customer, he/she access folders of products planed to be proposed. The accesses make operating system cache contents in the folders, and therefore sequentially read accesses would occur for files of the products. In this situation, access tendency "sequential access to multiple files in a folder" can not be extracted, because each product folder is treated as a distinctive object in absolute path.

Second problem is caused by parallel access. For analysis of web access log, we can utilize session ID to distinguish successive accesses. Since access log of file server only contains user ID (lacks session ID-like information), simultaneously multiple programs can mixed access sequences. Figure 2 shows an example where

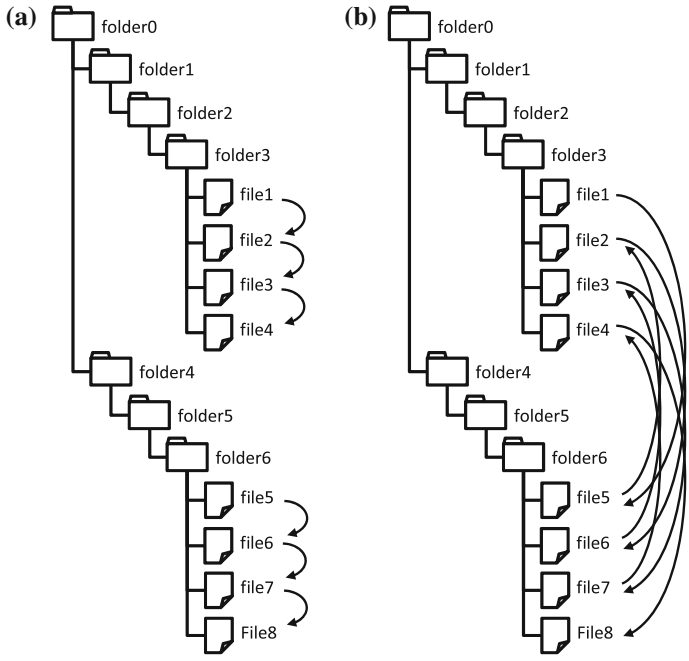


Fig. 2 Examples where two crawling-type accesses occur **a** separately and **b** simultaneously

two crawling access sequences are mixed. In case (b), the access log can not be considered as a crawling in a direct manner.

Third problem is about variation in length of access sequences. Pattern mining methods generally extracts patterns if their number of observation is larger than a threshold value "minimum support" [11–15]. Since number of observation decreases monotonously as length of pattern grows, threshold of number of observation suppress long pattern. Extraction method for long pattern is required, where file access characteristics is expected to be observed clearly.

3 Extraction Method for Frequent Access Pattern

In this section, we describe RPFAM method, which extracts frequent access pattern from access log of file server. We use the following notations.

$A = \{a_i\}_{i=1,\ldots,N}$	access log composed of ordered list of N accesses
$a_i = \langle p_i, u_i, t_i \rangle$	ith access composed of path, user, and timestamp
$p_i = \langle d_{i,0}, d_{i,1}, \ldots, d_{i,l_i} \rangle$	path of ith access a_i
$d_{i,j}$	jth-depth object (file or folder) of path of ith access: $d_{i,0}$ is the root folder for all i
u_i	user of ith access
t_i	timestamp of ith access
$t_{threshold}$	threshold value of time interval to decide two access are performed sequentially or not
$\subseteq$	inclusion relation of two paths: path $p_i = \langle d_{i,0}, d_{i,1}, \ldots, d_{i,l_i} \rangle$ is included in path $p_j = \langle d_{j,0}, d_{j,1}, \ldots, d_{j,l_j} \rangle$, i.e. $p_i \subseteq p_j$, when $d_{i,l} = d_{j,l}$ for all l of $l_i \leq l_j$ and $0 \leq l \leq l_i$

First, RPFAM method separates original sequence of accesses into sequence where user is identical and all time intervals are less than $t_{threshold}$. Next, it encodes path of each access in separated sequence based on relative position in folder tree structure, clips them into independent subtrees, and calculates occurrence probabilities of them. We denote sequence of encoded accesses as "access pattern." Finally RPFAM method outputs access patterns with number of observation larger than expected value based on occurrence probability. Three technical problems described in Sect. 2 are resolved by encoding based on relative position in folder tree structure, clipping accesses into independent subtrees, and calculating statistical significance of number of observation based on calculating occurrence probability. Detailed description is given for each step of RPFAM method in the following subsections.

Table 1 Example of relative position-based encoding

No.	Path	Encoded path
1	⟨folder1, folder3, file3⟩	⟨0, 0, 0⟩
2	⟨folder1, folder3, file2⟩	⟨0, 0, 1⟩
3	⟨folder2, folder4, folder0, file5⟩	⟨1, 0, 0, 0⟩
4	⟨folder1, folder3, file1⟩	⟨0, 0, 2⟩
5	⟨folder2, folder4, file6⟩	⟨1, 0, 1⟩

3.1 Encoding Paths Based on Relative Position in Folder Tree Structure

Sequence of accesses can be analyzed not based on absolute path but based on relative position of folder tree structure. Path of first access is encoded as $\langle 0, 0, \ldots, 0 \rangle$ and succeeding access is encoded by as follows.

Suppose that there are sequence of access $a_1, a_2, \ldots, a_n$, and $e_{i,l}$, encoded result of $d_{i,l}$, is given as:

$$e_{i,l} = 0, \text{ when } \langle d_{i,0}, d_{i,1}, \ldots, d_{i,l-1} \rangle \not\subset p_j,\ 1 \leq \forall j < i$$
$$e_{i,l} = e_{j,l}, \text{ when } 1 \leq \exists j < i \text{ s.t. } \langle d_{i,0}, d_{i,1}, \ldots, d_{i,l} \rangle \subseteq p_j$$
$$e_{i,l} = \max_{1 \leq j < i, < d_{i,0}, d_{i,1}, \ldots, d_{i,l-1} > \subset p_j} e_{j,l} + 1, \quad \text{when} \quad 1 \leq \exists j < i \quad \text{s.t.}$$
$$\langle d_{i,0}, d_{i,1}, \ldots, d_{i,l-1} \rangle \subset p_j \text{ and } \langle d_{i,0}, d_{i,1}, \ldots, d_{i,l} \rangle \not\subseteq p_k,\ 1 \leq \forall k < i$$

Table 1 shows example of encoded results. Files in folder "folder1/folder3" are encoded not in alphabetical order but in occurrence order. With this encoding, we can describe sequential access for files in a folder. Encoded result of third access represents that the access is for first file in different folder of first-depth, because first-depth folder object is encoded as "1" and the other folder file objects are encoded as "0."

3.2 Clipping Independent Subtrees

Sequence of accesses in Table 1 contains accesses for two independent subtrees as shown in Fig. 3. Independence is implied in encoding results: "1" and "0" is observed in first-depth and only "0" is observed in second-depth folders. For $a_1, a_2, \ldots, a_n$, RPFAM method examines whether $\exists l_1$ and $\exists l_2$ s.t. $l_2 > l_1$, $e_{i,l_1} > 0$, and $e_{i,l_2} = 0$, $1 \leq \forall i < n$. If there exists such l_1 and l_2, RPFAM method treats subtrees of depth of l_1 or more as independent for possible minimum value of l_1. This clipping step enables to treat each sequence of accesses are independent one when mixed sequence of accesses by multiple programs.

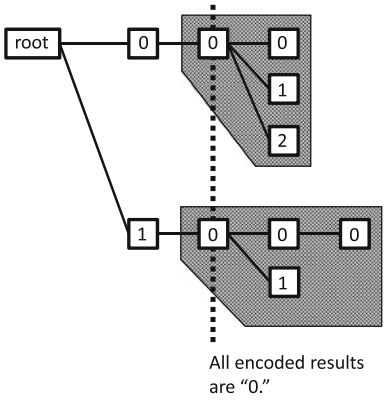


Fig. 3 Example of independent sub-tree in access sequence

3.3 Calculating Statistical Significance of Number of Observation for of Access Pattern Based on Occurrence Probability

We denote encoded result of ith access as $b_i = \langle q_i, u_i, t_i \rangle$. RPFAM method calculates occurrence probability of access pattern $b_1, b_2, \ldots, b_n$ on the basis of probability of going up/down in folder tree structure and distribution of number of objects in a folder. Suppose that there is a sequence of accesses $b_1, b_2, \ldots, b_n$ and path of target object b_i and b_{i+1} are $q_i = \langle e_{i,0}, e_{i,1}, \ldots, e_{i,l_i} \rangle$ and $q_{i+1} = \langle e_{i+1,0}, e_{i+1,1}, \ldots, e_{i+1,l_{i+1}} \rangle$, respectively $(1 \leq i < n)$. RPFAM method find l where $e_{i,0} = e_{i+1,0}$, $e_{i,1} = e_{i+1,1}, \ldots, e_{i,l} = e_{i+1,l}$, $e_{i,l+1} \neq e_{i+1,l+1}$, $1 \leq l < l_i$, and $1 \leq l < l_{i+1}$. Relative position of q_i to q_{i+1} can be described as "go up $(l_i - l + j)$ depth(s), go down $(l_{i+1} - l + j)$ depth(s), and select one object from objects in the subfolder for each step of going down" for $0 \leq \forall j < l$. We denote $upward(l)$ and $downward(l)$ as probability of going up l depth(s) and as going down l depth(s) for two neighboring accesses $(upward(0) + upward(1) + \cdots = 1$ and $downward(0) + downward(1) + \cdots = 1)$. We also define $sib(n)$ as probability of having n child objects $(sib(1) + sib(2) + \cdots = 1)$. With this definitions, probability of selecting a child object that encoded as can be expressed as following $child(m)$:

$child(m) = 1$, when no child object has been selected from the parent folder
$child(m) = \sum_{n=m}^{\infty} \frac{n-m+1}{n} sib(n)$, when one or more child object have been selected, but none of them is encoded as m
$child(m) = \sum_{n=m}^{\infty} \frac{1}{n} sib(n)$, when one or more child objects are selected and encoded as m

More specifically, when no child object has been selected yet from the object before, encoded result is always "0." When no encoded child object has been

selected from the object, the probability of selecting an object which is encoded as m can be described as probability of "the parent object has more than m child objects and a new child object is selected." On the other hand, when one or more m encoded object have been selected, the probability of selecting an object which is encoded as m can be described as probability of "the parent object has more than m child object and an existing child object is selected again."

With these probabilities, conditional probability of observing access b_{i+1} given access pattern $b_1, b_2, \ldots, b_i$ can be described as following $prob(b_{i+1}|b_1, b_2, \ldots, b_i)$:

$$\begin{aligned} prob(b_{i+1}|b_1, b_2, \ldots, b_i) = \sum_{j=0}^{l} (&upward(l_i - l + j) \\ &\cdot downward(l_{i+1} - l + j) \\ &\cdot \prod_{k=l-j+1}^{l_{i+1}} child(e_{i+1,k})) \end{aligned} \tag{1}$$

Probability of observing access pattern can be calculated by multiplying the conditional probabilities as follows:

$$prob(b_1, b_2, \ldots, b_n) = 1 \cdot \prod_{i=2}^{n} prob(b_i|b_1, \ldots, b_{i-1}) \tag{2}$$

Number of observation of focused access pattern accords with Poisson distribution with the probability $prob(b_1, \ldots, b_n)$ and number of accesses in access log. Since statistical significance of number of observation of access pattern can be calculated by comparing number of observation and the distribution, we can confirm statistical significance of number of observation independently of length of access pattern.

4 Experiment and Evaluation

In order to evaluate RPFAM method, we applied it to access log of enterprise file server. The access log contains 242,699 accesses of two hours. RPFAM method extracted 189,096 statistically significantly frequent access patterns at the significance level 0.05. 188,727 access patterns remain statistical significance after using Bonferroni adjustment [16]. Figure 4 shows scatter plot where x-axis represents observation number of access in access log, and y-axis represents occurrence probability of access pattern based on probability of going up/down in folder tree structure and distribution number of objects in a folder. If access randomly occurs, the x-values are expected to be proportional to y-values. Since data points in Fig. 3 deviate substantially from proportional relationship, access occurs in some tendency, not in random. Plots under dashed or solid line represent access patterns that are statistically significantly frequent with or without Bonferroni adjustment,

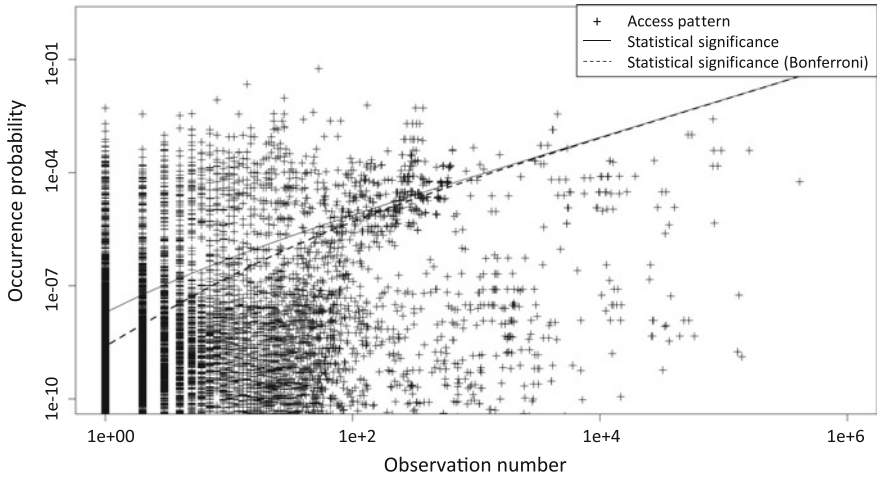


Fig. 4 Access patterns' observation numbers and occurrence probabilities based on folder tree structure

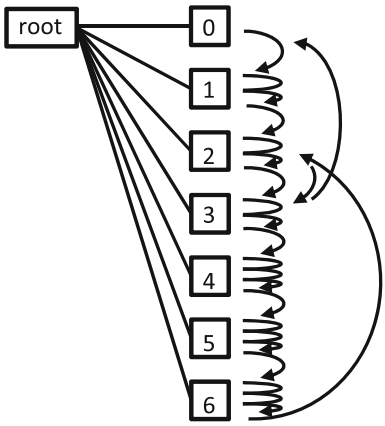


Fig. 5 An example of statistically significantly frequent access pattern

respectively. Figure 3 demonstrates that Bonferroni adjustment does not have significant effect to number of extracted access patterns when observation number is large enough.

Figures 5, 6, and 7 shows examples of especially significant access patterns. Access pattern of Fig. 5, 6, and 7 has observed 8 times (batch access cased by server program), 34 times (file system management process cased by administer account) and 6 times (accesses caused by backup program of end user) respectively.

To evaluate accuracy of RPFAM method, recall and precision are investigated as follows. Recall is calculated by comparing randomly selected system-driven access

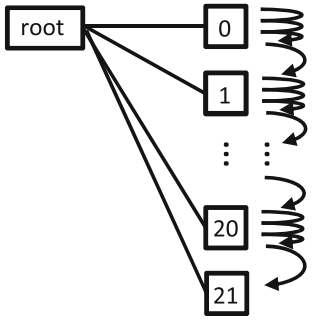


Fig. 6 Another example of statistically significantly frequent access pattern

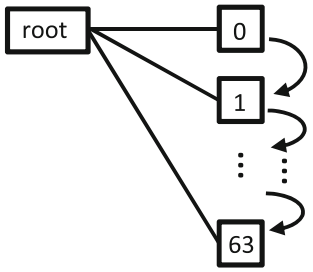


Fig. 7 Yet another example of statistically significantly frequent access pattern

sequences to statistically significantly frequent accesses patterns. Randomly selected 96 system-driven access sequences consists of 5759 accesses and 5755 accesses are covered by statistically significantly frequent access patterns. Since recall is calculated as 5755/5759 = 99.95 %, RPFAM method is unlikely to miss deletable files. With regard to precision of RPFAM method, 200 statistically significantly frequent access patterns are randomly selected after BonFerroni adjustment. Those access patterns contain 2802 accesses, and among them, one access was performed by a user. Since precision is calculated as (2802 − 1)/2802 = 99.96 %, user have to remove only a few files candidate deletable file list during manual confirmation process.

Table 2 Frequency of access patterns

	Without clipping independent sub-trees	With clipping independent sub-trees
Concentrated accesses	180,949 times (10 %)	397,954 times (6 %)
Crawling accesses	865,816 times (50 %)	4,446,145 times (71 %)
Non-typical accesses	696,280 times (40 %)	1,405,512 times (22 %)

Table 3 Types of access patterns

No.	Name	Description
1	Concentrated accesses	More than four accesses for one object
2	Crawling accesses	One-by-one accesses for more than four object
3	Non-typical accesses	Others

Now we evaluate RPFAM method in regards to three problems described in Sect. 2. First problem is about number of objects. For access sequences of length 1 (i.e. $a_1, \ldots, a_n$ of $n = 1$), number of unique path was 76,286 in the original access log. With the encoding step of RPFAM method, number of unique path was reduced to 1, because object in the first access is encoded as 0 by definition. For access sequences of length 2 (i.e. $a_1, \ldots, a_n$ of $n = 2$), number of unique pair of paths was 145,249 in the original access log. After the encoding step, the number of unique pairs was reduced to 245. These reductions demonstrate that number of unique access sequences can be dramatically suppressed by encoding path based on relative position of folder tree structure.

Second problem is caused by parallel accesses. Table 2 shows comparison result of number of observation of access patterns with or without clipping independent subtrees on the basis of types of access patterns described in Table 3. Although crawling accesses are frequently observed when system-driven access is performed, they look like non-typical access when multiple system process simultaneously access file server. Table 2 demonstrates that clipping independent subtrees decrease percentage of non-typical accesses and increase percentage of crawling accesses. This tendency indicates that many of crawling accesses look like non-typical accesses because multiple process simultaneously access to file server, and that they can be correctly as crawling accesses by clipping independent subtrees.

Third problem is about variation in length of access sequences. We compare access patterns ordered by statistically significance of number of observation and ordered by absolute number of observation. In Table 4, since value of correlation coefficient is 0.985, most of access patterns have similar order in both ways. However, 645 access patterns with top 10 % number of observation has bottom 10 % of statistical significance, their frequency is overestimated by simple absolute number of observation. We present two access patterns as examples of the gap between two ways. First example is an access pattern composed of single access.

Table 4 Comparison between access patterns evaluated by occurrence probability and evaluated by frequency occurrence

		Evaluated by frequency occurrence									
		~10 %	~20 %	~30 %	~40 %	~50 %	~60 %	~70 %	~80 %	~90 %	~100 %
Evaluated by occurrence probability	~10 %	38,633	44,954	146	0	0	0	0	0	0	0
	~20 %	37,010	33,008	13,716	0	0	0	0	0	0	0
	~30 %	7,446	5,590	69,693	1,005	0	0	0	0	0	0
	~40 %	0	0	0	82,729	1,005	0	0	0	0	0
	~50 %	0	0	0	0	82,729	1,005	0	0	0	0
	~60 %	0	0	0	0	0	82,729	1,005	0	0	0
	~70 %	0	0	0	0	0	0	82,729	1,005	0	0
	~80 %	0	0	0	0	0	0	0	82,729	1,005	0
	~90 %	0	0	0	0	0	0	0	0	82,729	1,005
	~100 %	645	182	179	0	0	0	0	0	0	82,729

Although number of observation is 243,591 and it looks large enough, occurrence probability is 1 and the number of observation has no statistical significance. Second example is a crawling access in Fig. 6. Although it is observed only 6 times, very small occurrence probability makes the number of observation be statistically significantly frequent. These examples indicate that calculating statistical significance based on occurrence probability is suitable to detect characteristics in system-driven accesses.

Above three evaluations demonstrates that RPFAM method can extract frequent accesses pattern from access log of file server and that RPFAM method is suitable for system-driven accesses. They are preferable features for distinguishing system-driven accesses from user-intended accesses.

5 Conclusion

We have proposed RPFAM method as a extraction method of frequent access pattern. RPFAM method encodes paths of sequence of accesses based on relative position in folder tree structure, clips them into independent subtrees, and calculates statistical significance of their number of observation by occurrence probability calculated with probability of going up/down layer(s) in folder tree structure and with distribution of number of child objects. These steps can cope with large number of access target objects, multiple simultaneous accesses, and various lengths of access sequences. In this paper, we demonstrate that RPFAM method can extract frequent pattern from access log of file server in high accuracy.

RPFAM method is expected to be effective for distinguish system-driven accesses from user-intended ones and for refining potentially deletable file list. When access time is re-calculated from user-intended accesses, we can convincingly construct candidate list of deletable files. Such a candidate list is expected to contribute for reduction of file server volume and to achieve efficient operational management of enterprise file servers. As a further improvement of RPFAM method, kind of file or connection origin of user can be utilized for more accurate extraction of system-driven accesses.

References

1. Gantz J, Reinsel D (2011) Extracting value from Chaos, IDC IVIEW. http://www.emc.com/collateral/analyst-reports/idc-extracting-value-from-chaos-ar.pdf
2. Meyer DT, Bolosky WJ (2011) A study of practical deduplication. In: Proceedings of 9th USENIX conference on file and storage technologies
3. Shah G, Voruganti K, Shivam P, Alvarez M, Shah G, Voruganti K (2006) ACE: classification for information lifecycle. Computer science IBM research report, RJ10372

4. Gibson T, Miller EL, Long DDE (1998) Long-term file activity and inter-reference patterns. In: Proceedings of 24th international conference on technology management and performance evaluation of enterprise-wide information systems
5. Tanaka T, Ueda R, Aizono T, Ushijima K, Naitoh I, Komoda N (2005) Proposal and evaluation of policy description for information lifecycle management, vol 1. In: Proceedings of the 2005 international conference on computational intelligence for modelling, control and automation, and international conference on intelligent agents, web technologies and internet commerce, pp 261–267
6. Ludescher C, Carroll T, Murphy J, Zarnstoff M (1991) File storage management for TFTR physics data, vol 2. In: Proceedings of 14th IEEE/NPSS symposium on fusion engineering, pp 856–859
7. Rao SNT, Prased EV, Venkateswarlu NB (2010) A critical performance study of memory mapping on multi-core processors: an experiment with k-means algorithm with large data mining data sets. Int J Comput Appl 1(9):1–9
8. Matsumoto T, Onoyama T (2014) Extraction method of frequent file access patterns based on relative position in folder tree. 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 1234–1234
9. Leung AW, Pasupathy S, Goodson G, Miller EL (2008) Measurement and analysis of large-scale network file system workloads. In: Proceedings of USENIX08
10. Roselli D, Lorch JR, Anderson TE (2000) A comparison of file system workloads. In: Proceedings of 2000 USENIX annual technical conference
11. Agrawal R, Srikant R (1994) Fast algorithm for mining association rules in large databases. Proc Int Conf Very Large Data Bases 1994:487–499
12. Agrawal R, Mannila H, Srikant R, Toivonen H, Verkamo AI (1996) Fast discovery of association rules. Adv Knowl Discov Data Min 307–328
13. Zaki MJ, Parthasarathy S, Ogihara M, Li W (1997) New algorithms for fast discovery of association rules. In: Proceedings of knowledge discovery in database 1997, pp 283–286
14. Bayardo RJ Jr (1998) Efficiently mining long patterns from databases. In: Proceedings of ACM special interest on managing data, pp 85–93
15. Pei J, Han J, Pinto B, Chen Q, Dayal U, Hsu M (2001) PrefixSpan: mining sequential patterns by prefix-projected growth. Proc IEEE Int Conf Data Eng 2001:215–224
16. Miller GR (1981) Simultaneous statistical inference, 2nd edn. Springer, New York

Uncertainty Characterization of Performance Measure: A Fuzzy Logic Approach

Sérgio Dinis Teixeira de Sousa, Eusébio Manuel Pinto Nunes and Isabel da Silva Lopes

Abstract The process of performance measurement encompasses the activities of designing, data collection and analysis. The lack of quality of performance measures (PMs) may influence decision-making. Since the process of performance measurement involves generally several actors, the decision-maker may not be aware of the level of uncertainty associated with performance measures. In this paper, fuzzy logic is used to represent the uncertainty generated in PMs during its design, use and analysis stages. The identification of uncertainty sources and the determination of an Uncertainty Index support actions to improve performance measures' quality. An application example is provided to show the usefulness of the proposed methodology.

Keywords Data quality · Decision-making · Fuzzy logic · Manufacturing · Performance measures · Uncertainty

1 Introduction

The monitoring of efficiency and effectiveness of processes or systems allows managers to control, decide, implement and observe the effects of their actions to understand if they are moving towards achieving their goals. The definition and selection of performance measures (PMs) have attracted the attention of several authors and different knowledge areas propose extensive lists of PMs. Models,

S.D. Teixeira de Sousa (✉) · E.M.P. Nunes · I. da Silva Lopes
Centro Algoritmi, University of Minho, Campus de Gualtar, 4710-057 Braga, Portugal
e-mail: sds@dps.uminho.pt

E.M.P. Nunes
e-mail: enunes@dps.uminho.pt

I. da Silva Lopes
e-mail: ilopes@dps.uminho.pt

H.K. Kim et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-94-017-7236-5_34

framework and standards are available in literature for the selection of an adequate set of PMs for a given process or business.

PMs can be considered a particular type of Data or Information and the literature refers some dimensions or attributes of Data/Information Quality as [1]: accuracy; completeness; timeliness; and consistency. This implicitly suggests that all data may lack some of these attributes.

Several other classifications for Data/Information Quality are proposed in literature [2]. For example, Galway and Hanks [3] classify data quality problems as operational, conceptual and organizational.

The quantification of uncertainty can contribute to highlight the lack of accuracy and precision of data or PMs. The decision-maker should know the existence and magnitude of uncertainty on PMs, once this uncertainty may lead to higher risk in decisions. The study of the sources of PMs' uncertainty can also contribute to identify adequate initiatives to improve the process of designing, using and analyzing the PMs. In this context, in order to get useful information about factors that induce uncertainty in PMs, adequate modeling tools can be applied, as is the case of fuzzy logic.

Fuzzy logic was introduced by Zadeh [4] as a mathematical technique to represent vagueness in everyday life. Fuzzy logic methodology can model complicated processes and deal with qualitative imprecise or vagueness knowledge and information [5]. It provides a tool for directly working with the linguistic terms used in the assessment of factors that contribute to uncertainty of PMs, and has many applications in Performance measurement field [6, 7]. When the available information from the process is qualitative, inexact, incomplete, imprecise, vague or uncertain, the notion of the membership function utilized by fuzzy theory is then adequate for depicting this knowledge.

To contribute to the field of Performance Measurement and to the field of Data Quality, this work studies the causes of uncertainty on the process of using PMs. Fuzzy numbers are used to develop a methodology that aims to obtain the value of an index for uncertainty evaluation of a given PM or (key) performance indicator.

The paper is organized as follows. Section 2 presents a literature review about performance measurement. In Sect. 3, the stages of the performance measurement process and the uncertainty sources that may affect it in each stage are defined. Section 4 presents the proposed methodology based on fuzzy logic to estimate PM uncertainty. Sections 5 and 6 present, respectively, an example of application of the methodology and the final conclusion of this work.

2 Performance Measurement

Performance measurement can be described as a set of steps involving firstly the design and implementation of the performance measure and, secondly, the use and review of performance measures. Concerning the first two steps, design and implementation, there are several works in literature that propose and define different performance measures in different knowledge area and that discuss the

implementation difficulties and critical success factors [8]. The "use" step involves a sequence of tasks that are repeated each time a new value of the performance measure is requested. These tasks aim to collect the necessary data and present results [9].

Juran and Godfrey [9] argue "the choice of what to measure and the analysis, synthesis, and presentation of the information are just as important as the act of measurement itself". These authors also argue that the measurement process belongs to a larger measurement system, which embraces the decisions that are made and the framework in which the process operates. This large system involves, typically, different actors.

A set of related PMs can be called a performance measurement system (PMS). Many PMSs are available to companies, such as the Balanced Scorecard, but many other activities related with assessing processes' performance or comparing the performance of similar subsystems, or generally doing benchmarking exercises, all rely on (key) PMs. To design a Performance Measurement System it is crucial to understand who will make the decisions (and how) and who will take actions [9], i.e. the purpose of each PM must be clear [10], and must promote the company's strategy [11]. Before determining what to measure and how to measure it, the overall framework in which the Performance Measurement System operates should be understood [9]. It can be concluded that the relevance of PMs is related to decisions they can support and that there are no bad PMs, only the bad use of them [12].

To increase quality of PMs some of its attributes or requirements are identified in the literature [11–13]: relevant; credible; precise; valid; reliable and frequent. Other authors refer some recommendations for both the performance measurement process and performance measures:

- data collection and methods for calculating the PMs must be clearly defined [14];
- presentation of PMs must be simple [15];
- PMs must be flexible [12], including being tied to desired results [16];
- more extensive use should be made of subjective data [11]; and
- ratio-based performance criteria are preferred to absolute numbers [14].

However, the designing of PMs may not comply with all of these recommendations and, even if they are all fulfilled at the design stage, during its implementation or use changes in the system on which the PMs are integrated may result in PMs that do not fulfill all the above mentioned requirements [17].

3 The Performance Measurement Process and Uncertainty Sources

The process usually referred in the performance measurement literature is divided into (a) design, (b) implementation, (c) use and (d) review [18], while in the field of quality management, according to Juran and Godfrey [9] those activities consist of: (i) understand the framework, (ii) plan the measurement, (iii) collect and store data,

(iv) analyze, synthesize, formulate recommendations, present results and recommendations, and (v) make decision and take action.

The process of performance measurement will be detailed in this work on three stages [19]: design (a; b; i; ii;), data collection and record (iii; c excluding analysis) and determination and analysis (iv; c only analysis). In terms of frequency, the first stage is the less frequent and the second one the most frequent, because each new analysis requires new data.

This work will analyze the uncertainty that could be introduced in the design and implementation stage of performance measurement process and subsequently in the use and analysis stages [19]. Finally, the overall uncertainty assessment will provide information to the performance measurement review process.

The decision/action process is out of the scope of this work but the information about the uncertainty of a PM may be relevant to ascertain the risk associated with a given decision/action based on a PM.

Risk is a possibility and consequence of a given event. Typically, PMs are used to verify if a goal is achieved or it is used to justify actions (to put a process under control, to allocate resources, etc.). If a PM value, including measurement errors and other uncertainties, is close to a target or limit, there is a possibility that its true value could lead to a different action.

3.1 Performance Measurement Design and Implementation

The process of performance measurement starts by PMs' Design. Despite the inexistence of a universal set of rules or model to develop an ideal PM or an ideal PMS, literature suggests principles, models, frameworks and attributes or requirements of good PMs.

Designing a PM consists of defining a set of attributes, such as [9] and [20–22]:

- name of PM;
- purpose;
- target;
- data source;
- PM owner;
- frequency of measurement;
- measurement method/equipment;
- formula;
- units of measurement;
- control/reaction limits;
- frequency/method of analysis;
- responsible of analysis;
- possible immediate actions;
- PM customer, among others.

Some of the above attributes may not be applicable. For example, if a PM's objective is to control a process, the definition of control/reaction limits or possible immediate actions will be important attributes, but if the PM objective is to ascertain if an organization goal, assessed by a (set of) PM is being achieved, those attribute may not be relevant.

Considering that each organization can influence the process of designing a PM, the result of this stage could be a perfect PM design or a design with many flaws or unwanted characteristics. Operational, organizational or financial constraints may also arise during implementation and will compromise the planning and specifications defined in the design phase. Therefore several factors can, ultimately, influence (induce errors of unknown magnitude) the values of the PM. It was decided, in his work to organize them in the well-known cause-and effect or Fishbone diagram, one the basic quality tools [9]. This diagram has the ability to represent graphically, involving a team in the identification of the controllable, causes and sub-causes of a given unwanted event, which in this case is: deficiencies at the design and implementation stage.

The controllable factors that are considered at this stage are (Fig. 1):

- People involved in the design and implementation stage (such as its experience and complacency);
- Environmental context where the measurement takes place (such as its complexity and predictability);
- Policy adopted by the organization (related to quality management and knowledge management);
- Human resources (HR) management and procedure used to design and implement the PM.

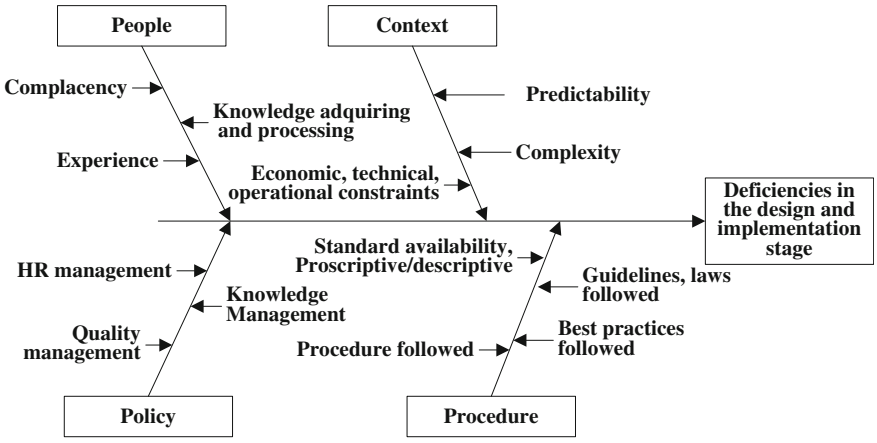


Fig. 1 Cause and effect diagram on deficiencies introduced by the design stage

3.2 Performance Measurement Collection and Record

The Measurement activity consists of obtaining data and can be performed in different ways, such as reading a value in a measurement device that may be installed in the production process or counting the number of occurrence of a predefined event.

To use the data collected in the previous activity to calculate a PM for a given period of time, data is registered in a computer or physical datasheet. The Data Record/Transmission activity may be made in different ways: data is automatically registered when the measurement is performed (automatic record system); or the data is recorded by someone who reads the value and writes it in a computer system or datasheet (handmade record).

In both of these activities, several factors can induce uncertainty. These factors are arranged in Fig. 2 according to four main groups:

- Equipment used to measure and to record data (accuracy, precision, proneness to error, etc.);
- Workplace environment (such as luminosity, tidiness and workplace organization);
- Operator (such as physical and mental fitness, and complacency);
- Method (existence of instructions for measuring and recording data and its clarity).

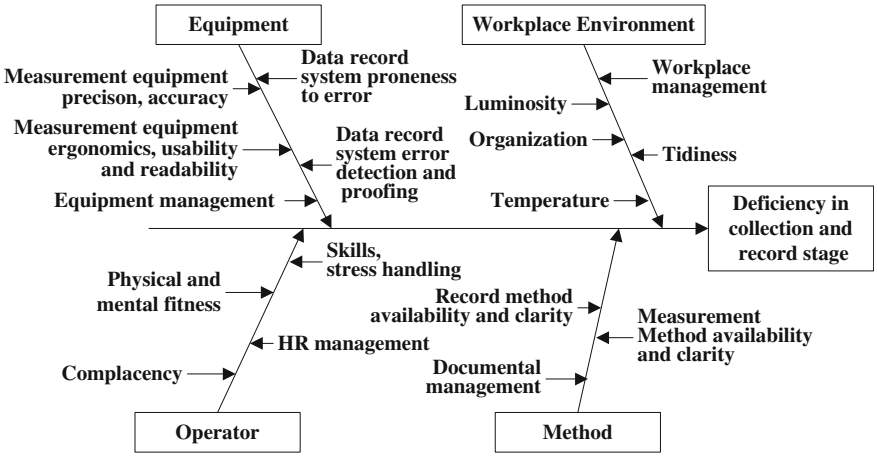


Fig. 2 Cause and effect diagram on the uncertainty introduced in the data collection and record stage

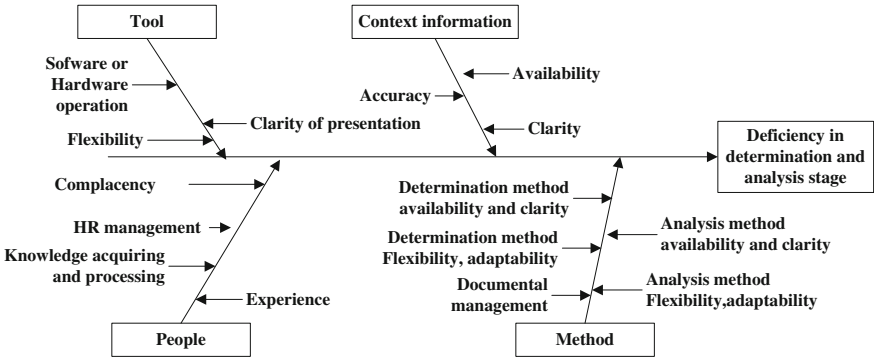


Fig. 3 Cause-and-effect diagram on the uncertainty introduced in the determination and analysis stage

3.3 Performance Measurement Determination and Analysis

The PM Determination activity consists of selecting recorded data for a specific period of time and applying a predefined expression for calculating the PM. This task may also be made automatically by a computer application or can be made manually.

Analysis consists of preparing data to make it useful to support decisions. It includes, managing missing or suspect PM values, trend analysis, summarizing data and compare with predefined values.

Similarly to the previous stages, PM determination and analysis may introduce errors in the PM value. The following factors are organized in four main groups (Fig. 3):

- Tool (clarity, flexibility and modus operandi of the tool used to determine and analyze the PM values);
- Context information (availability, clarity and accuracy of the context information used to interpret the PM values;
- People (such as experience and, knowledge acquiring and processing);
- Method (flexibility, adaptability, availability and clarity of both determination and analysis methods).

4 Methodology to Estimate PM Uncertainty Based on Fuzzy Logic

4.1 Input and Output Variables

Deficiencies generated in design, collection and record, and determination and analysis stages, introduce uncertainty in the measurement process that is reflected in

greater or lesser confidence in the PM value. Once the sources of uncertainty present in each of these stages of performance measurement are identified, the uncertainty assessment team evaluates the level of these deficiencies in each of these stages and their implications on the uncertainty of the PM.

Consider that the deficiencies in the design, collection and records, and analysis stages are represented by the variables $\tilde{x}_d$, $\tilde{x}_u$ and $\tilde{x}_a$, respectively, and that the uncertainty in the PM is represented by the variable y. In mathematical terms, one could express the relationship between input variables ($\tilde{x}_d$, $\tilde{x}_u$ and $\tilde{x}_a$) and output variable ($\tilde{y}$) by $\tilde{y} = f(\tilde{x}_d, \tilde{x}_u, \tilde{x}_a)$.

In many performances measurement processes there are frequently several factors that must be taken into account if we want to evaluate the deficiencies in each one of the three considered stages, and usually the implications of each one of these deficiencies are not well known. In addition, there is not an objective scale for measuring the resulting deficiencies, represented in this work by the input variables ($\tilde{x}_d$, $\tilde{x}_u$ and $\tilde{x}_a$), nor for the output variable $\tilde{y}$. Therefore, the most natural way to assess these variables by the team responsible for the PM uncertainty assessment is to use linguistic terms (words) of the natural language such as "very low", "often", "moderate", etc.

Concerning the function f that relates the input variables to the output variable, there is no known analytical relationship. However, the knowledge that the project team has about the implications of the deficiencies (identified at the stages level) causes on the PM uncertainty allows establishing a set of cause-consequence combinations of different levels of input variables to the output variable. In this context the general framework of fuzzy reasoning facilitates the handling of the PM uncertainty. Fuzzy logic and fuzzy systems are suitable for representing and employing knowledge that is imprecise, uncertain, or unreliable.

4.2 Fuzzy Logic Method

The general Fuzzy Logic method consists of four basic components, namely:

- Fuzzification (definition of the fuzzy set of the input and output variables);
- Fuzzy rule base (definition of the rules that correlate the input variables to the output ones);
- Fuzzy inference engine (aggregation of the contributions of the rules);
- Defuzzification of the results.

4.2.1 Fuzzyfication

It is the process of decomposing a system input variables into one or more fuzzy sets, thus producing a number of fuzzy perceptions of the input, and carrying out a mapping from real-domain variables.

In this work a fuzzy system with three input variables (intended to assess the deficiencies levels associated with design, collection and record and analysis stages) and an output variable (uncertainty in PM triggered by the input variables) is designed.

Each of the input variables is associated with five linguistic terms:

{Very Low, Low, Moderate, High, Very High}

For the output variable more discrimination was obtained by adding another four linguistic terms obtaining thereby nine levels:

{Exceptionally low, Extremely low, Very low, Low, Moderate, High, Very high, Extremely high, Exceptionally high}

Triangular membership functions are used to define the fuzzy set for each linguistic term. The membership functions are defined for the three inputs variables and the output variable. Graphical representation of $\tilde{x}_d$ is provided in Fig. 4. Similar graphs can be drawn for $\tilde{x}_u$ and $\tilde{x}_a$. Figure 5 represents the output variable. As

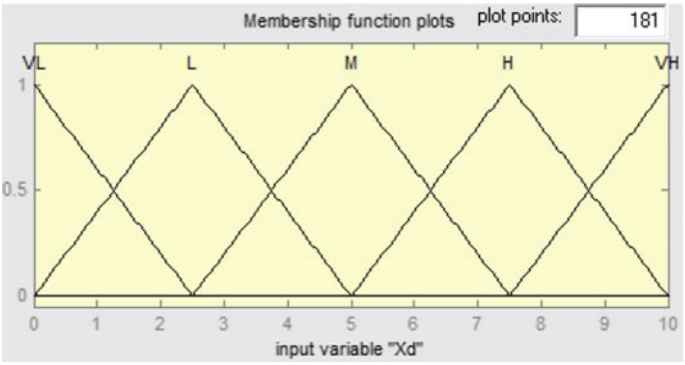


Fig. 4 Graphical representation (membership functions) of fuzzy variable $\tilde{x}_d$

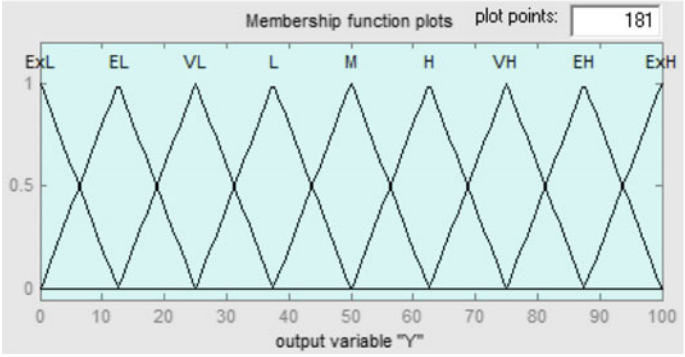


Fig. 5 Graphical representation (membership functions) of fuzzy variable $\tilde{y}$

shown in these figures, the input variables have their universes of discourse defined between 0 and 10, and the universe of discourse of the output variable is defined between 0 and 100. After determining the fuzzy membership functions of all variables, the knowledge about the relationship between input and output variables is mapped through fuzzy rules.

4.2.2 Fuzzy Rule Base

Fuzzy rules consist of a set of decisional type IF–THEN, this meaning that, the consequences occur only if the premises are real. The fuzzy rules represent the logical correlations between input and output variables and are of the following form:

$$R^{(k)} : \text{IF } \tilde{x}_1 \text{ is } \tilde{A}_1^k \text{ AND} \ldots \tilde{x}_n \text{ is } \tilde{A}_n^k, \text{ THEN } \tilde{y} \text{ is } \tilde{B}^k \tag{1}$$

where $\tilde{A}_i^k (i = 1, 2, \ldots, n)$ and $\tilde{B}^k$ are fuzzy sets, $\tilde{x} = (\tilde{x}_1, \tilde{x}_2, \tilde{x}_3)^T \in U$ and $\tilde{y} \in V$ are input and output linguistic variables, respectively, and k represents the number of the rules ($k = 1, 2, \ldots, S$).

The number of rules depends on the number of inputs and outputs and the desired behavior of the system. Once the rules have been established, such a system can be viewed as a non-linear mapping from inputs to outputs.

The rules are deduced from the knowledge and experience of the project team. In this case, from Eq. (1) we have:

$$\tilde{A}_i^k = (\text{VL}, \text{L}, \text{M}, \text{H}, \text{VH})$$

$$\tilde{B}^k = (\text{ExL}, \text{EL}, \text{VL}, \text{L}, \text{M}, \text{H}, \text{VH}, \text{EH}, \text{ExH})$$

$$\text{R}^{(1)} = \text{IF } \tilde{x}_d \text{ is EL and } \tilde{x}_u \text{ is EL and } \tilde{z}_a \text{ is EL, THEN } \tilde{y} \text{ is ExL}$$

As in this study the system has three inputs with five linguistic terms, there are 125 (5^3) possible fuzzy rules that can be defined.

Figure 6 presents the output as a function of two input variables considering the other variable equal to its average value. Similar graphs can be obtained using other combinations of input variables.

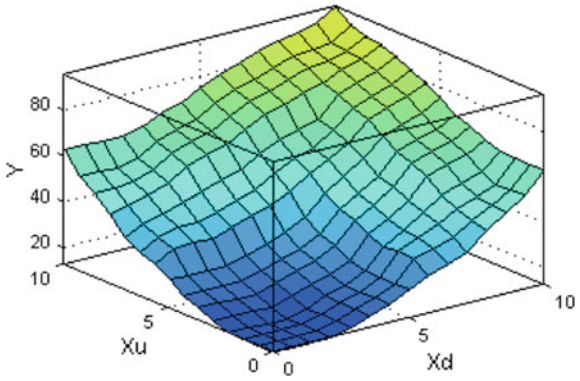


Fig. 6 Graphical representation of the relation between the input and output variables

4.2.3 Fuzzy Inference Engine

The fuzzy inference engine uses these fuzzy IF–THEN rules to determine a mapping from fuzzy sets in the input universe of discourse U ($U \in R^n$) to fuzzy sets in the output universe of discourse V ($V \in R$) based on fuzzy logic principles [23].

According to the truth degree of premises, each rule activates a portion of a specific output fuzzy set; therefore the result of fuzzy problem derives from the union of the several portions of areas activated at the same time. There are many fuzzy inference methods. This paper uses the min–max fuzzy inference method proposed by Mamdani [24]. The final output of a Mamdani system is one or more arbitrarily complex fuzzy sets which (usually) need to be defuzzified. It is not appropriate to present a full description of the functioning of fuzzy systems here; the interested reader is referred to [25] or [5].

4.2.4 Defuzzification of the Results

Defuzzification is the process that transforms the output fuzzy set to crisp output by applying specific defuzzification method. There are some methods of defuzzification, but the most common is the centroid method, this calculates the center of area of the fuzzy set and uses the value at which this occurs as the defuzzified output. In this work the centroid method was used to obtain a single crisp (real) value for the output variable.

5 Application Example

The application example concerns the manufacturing of Printed Circuit Boards (PCBs) to be used in car multimedia systems by a multinational company. In the production lines of PCBs, an automated optical inspection (AOI) system is used to

control and assess the quality of the reflow soldering process. PCBs are autonomously scanned by a camera to identify a variety of soldering defects such as open circuits or short circuits. These defects are measured by the volume of solder paste placed on a given PCB position and compared with pre-defined specifications.

To reduce defects detected at the final quality control test, aligned with the company continuous improvement culture, it was decided to use a PM to be calculated at the end of the reflow soldering process: number of soldering defects per million opportunities (DPMO).

A quality team was commissioned to define the PM and the methods of collection, record, determination and analysis. The same team implemented the PM. Since the equipment, AOI, is not able to measure all the positions of soldering deposition during the cycle time, only usual critical positions are analyzed. The PCB fixation mechanism and its position when optical inspection is made is a critical factor to the quality of the measurements.

In each PCB, AOI signals and registers the number of soldering defects in the critical inspected positions in a database. The number of inspected positions is also recorded for each PCB in the same database. Daily DPMO is calculated for each shift of the company production lines. The DPMO values are controlled daily by the line manager. Weekly, in the quality team meeting, the obtained values in each line and shift are compared to each other and with the established target, and possible tendencies are checked.

In one of these weekly meetings, the analysis of the PM uncertainty was undertaken following the methodology proposed in this paper and, starting with the three cause and effects diagrams proposed, the three input variables values were defined based on linguistic terms by consensus: $\tilde{x}_d$ = "low"; $\tilde{x}_u$ = "medium"; $\tilde{x}_a$ = "low". Since there was consensus, the modal value was taken as input. The proposed rules to express uncertainty in the PM based on these input variables was presented and agreed.

Figure 7 presents the results for the DPMO studied resulting in an uncertainty level of 12.5 (out of 100). In linguistic terms it belongs to the EL and VL membership functions. This could provide a basis to compare this PM with other PMs used by the organization and to support the reviewing of existing PMs to reduce overall PM uncertainty or to improve PM quality.

If there is no consensus on the level of variables, another value instead of the modal value can be taken. An example is provided in Fig. 8, where $\tilde{x}_d$ takes the value 3.5 considering that some of the participant attribute a "low" and other a "medium" level for this variable. In this case, the uncertainty level increases to 17.8.

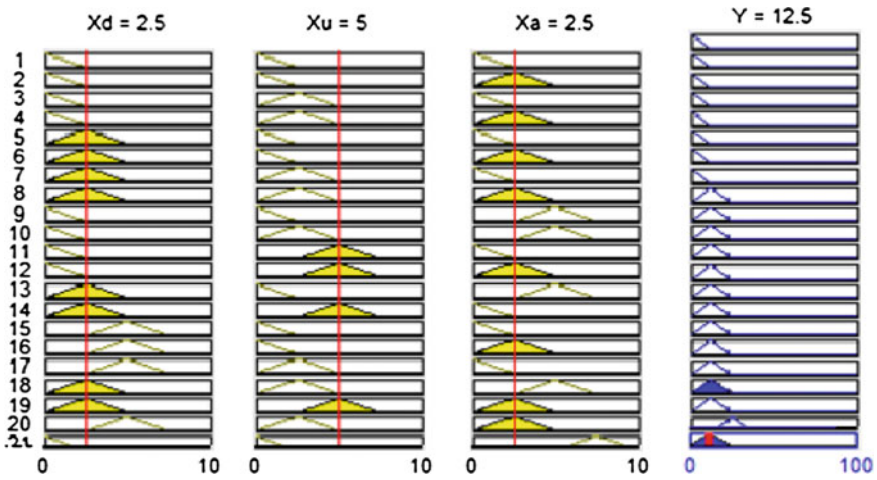


Fig. 7 Graphical representation of the output generation ($x_d = 2.5$; $x_u = 5$; $x_a = 2.5$)

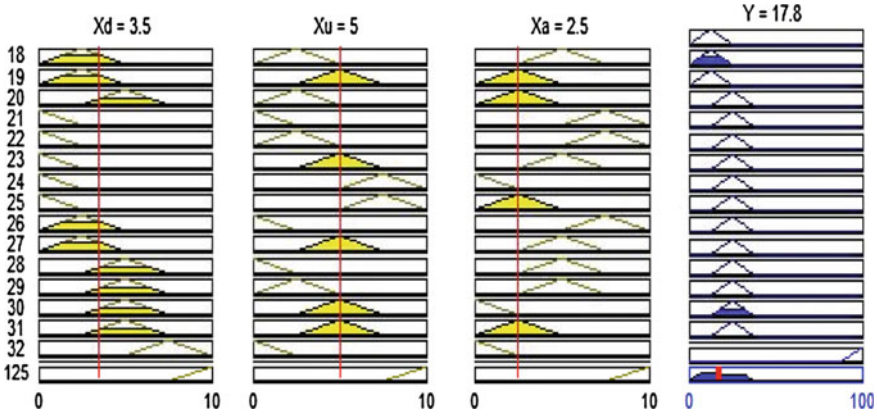


Fig. 8 Graphical representation of the output generation ($x_d = 3.5$; $x_u = 5$; $x_a = 2.5$)

6 Conclusion

In this work, the process of performance measurement was defined and analyzed to identify the sources or causes of uncertainty that may induce uncertainty in the PM. The causes can be assessed for any PM and some causes may be common for a given organization. In each particular case the presence and level of each source must be questioned to define its influence in the uncertainty of the considered PM.

Three cause-and-effect diagrams were used to graphically represent the uncertainty causes in each stage of the performance measurement process. The defined stages are: (1) design and implementation; (2) data collection and record; and

(3) analysis. Based on the defined diagrams, three fuzzy variables are defined. These input variables result in an output variable, also represented by a fuzzy number, which represents the level of uncertainty of a given PM.

The developed approach allows identifying the changes that can be introduced in the performance measurement process to obtain more trustable values for the key PMs used in decision-making.

This work is part of a project that aims to develop a framework to reduce the uncertainty of performance measurement systems. The knowledge of uncertainty associated with a PM also allows considering, in decision-making process, the risk due to lack of data quality.

Acknowledgment This work has been supported by FCT—Fundação para a Ciência e Tecnologia within the Project Scope: PEst-OE/EEI/UI0319/2014.

References

1. Batini C, Cappiello C, Francalanci C, Maurino A (2009) Methodologies for data quality assessment and improvement. J ACM Comput Surv 41(3):1–52
2. Lee YW, Strong DM, Kahn BK, Wang RY (2002) AIMQ: a methodology for information quality assessment. Inf Manag 40(2):133–146
3. Galway LA, Hanks CH (2011) Classifying data quality problems. IAIDQ's Inf Data Qual Newslett 7(4):1–3
4. Zadeh LA (1965) Fuzzy sets. Inf Control 8(3):338–353
5. Klir J, Yuan B (1995) Fuzzy sets and fuzzy logic: theory and applications. Prentice Hall, New Jersey
6. Yadav OP, Singh N, Chinnam RB, Goel PS (2003) A fuzzy logic based approach to reliability improvement estimation during product development. Reliab Eng Syst Saf 36(32):63–74
7. Kim BJ, Bishu R (2006) Uncertainty of human error and fuzzy approach to human reliability analysis. Int J Uncertainty Fuzziness Knowl-based Syst 14(1):111–129
8. Sousa SD, Nunes EP, Lopes IS (2014) Using fuzzy logic to characterize uncertainty during the design and use stages of performance measurement. In: Proceedings of the world congress on engineering and computer science. Lecture notes in engineering and computer science, San Francisco, pp 936–941, 22–24 Oct 2014, ISBN 9749881925374
9. Juran JM, Godfrey AB (1999) Juran's quality handbook, 5th edn. McGraw-Hill, USA
10. Basu R (2001) New criteria of performance management. Measuring Bus Excellence 5(4):7–12
11. Schalkwyk J (1998) Total quality management and the performance measurement barrier. TQM Mag 10(2):124–131
12. Macpherson M (2001) Performance measurement in not-for-profit and public-sector organizations. Measuring Bus Excellence 5(2):13–17
13. Ghalayini A, Noble J, Crowe T (1997) An integrated dynamic performance measurement system for improving manufacturing competitiveness. Int J Prod Econ 48:207–225
14. Globerson S (1985) Issues in developing a performance criteria system for an organisation. Int J Prod Res 23(4):639–646
15. Tenner A, DeToro I (1997) Process redesign. Addison-Wesley, Harlow
16. Franco M, Bourne M (2003) Factors that play a role in managing through measures. Manag Decis 41(8):698–710
17. Sousa SD, Nunes EP, Lopes IS (2012) Uncertainty components in performance measures. In: Gi-Chul Y et al (ed) IAENG transactions on engineering technologies—special issue of the world congress on engineering 2012. Springer, New York, pp 753–765

18. Braz R, Frutuoso G, Martins R (2011) Reviewing and improving performance measurement systems: an action research. Int J Prod Econ 133:751–760
19. Sousa SD, Aspinwall E (2010) Development of a performance measurement framework for SMEs. Total Qual Manage Bus Excellence 21(5):475–501
20. Lohman C, Fortuin L, Wouters M (2004) Designing a performance measurement system: a case study. Eur J Oper Res 156:267–286
21. Lima EP, Costa SE, Angelis JJ (2009) Strategic performance measurement systems: a discussion about their roles. Measuring Bus Excellence 13(3):39–48
22. Choong K (2013) Understanding the features of performance measurement system. Measuring Bus Excellence 17(4):102–121
23. Guimaraes ACF, Lapa CMF (2007) Fuzzy inference to risk assessment on nuclear engineering systems. J Appl Comput 7:17–28
24. Mamdani EH, Assilian S (1975) An experiment in linguistic synthesis with a fuzzy logic controller. Int J Man Mach Stud 7(1):1–13
25. Ross TJ (1995) Fuzzy logic with engineering applications, 1st edn. McGrawHill, Inc, New York

Survey on Maintenance Area of Companies of the Manaus Industrial Pole

Marcelo Albuquerque de Oliveira, Isabel da Silva Lopes and Danielle Lima de Figueiredo

Abstract The study presented in this paper aims to identify the practices of maintenance organization and management adopted by companies in the industrial pole of Manaus in Brazil in relation to their production equipment. A questionnaire was developed and sent to the company of the industrial pole and the received data was analyzed using descriptive statistics. This survey shows that companies have several opportunities to make improvements related to maintenance management, such as adopting preventive culture, using adequate CMMS to support maintenance activities with associated failure analysis tools, and adopting meaningful KPIs.

Keywords Maintenance strategy · Maintenance management · Maintenance performance · Maturity level · Questionnaire · Survey

1 Introduction

The importance of maintenance function and, therefore, of maintenance management has grown in recent years. Maintenance has achieved a significant role in organizations performance since it is directly responsible for the proper operation of the production process. Maintenance area needs to keep equipment in good working conditions and available to achieve high productivity level of quality products.

M.A. de Oliveira (✉) · I. da Silva Lopes · D.L. de Figueiredo
Centro Algoritimi, University of Minho, Campus de Gualtar, 4170-057 Braga, Portugal
e-mail: marcelo@maoconsultoria.com

I. da Silva Lopes
e-mail: ilopes@dps.uminho.pt

D.L. de Figueiredo
e-mail: danielle_lima@ig.com.br

H.K. Kim et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-94-017-7236-5_35

According to Kumar and Parida [1] maintenance is defined as the combination of all the technical and administrative actions, including supervision, intended to retain an item, or restore it to a state in which it can perform a required function. Maintenance management makes use of some tools and techniques to improve efficiency and minimize the impacts of unplanned stoppages looking for reducing costs.

The Industrial Pole of Manaus is one of the most modern in Latin America, bringing together major industries in the areas of electronics, motorcycles, optical products, computer products, chemical industry. Through a survey, data was collected about the maintenance practices adopted by the maintenance function of these companies in order to identify its current situation and ascertain potential improvements that this management function can bring. This will allow proceeding with a maturity model indicating the path to follow to achieve higher levels of performance in this area.

The survey aimed to comprise the management process of the overall equipment of companies that act at Manaus industrial pole concerning maintenance methodologies, maintenance department structure, KPI utilization, maintenance planning and staff development. The identification of real state of maintenance area leads us to a better planning of actions to implement the more appropriate management strategy and to propose suitable computerized tools, performance indicators, technical analysis and management tools and techniques, providing a set of potential improvements necessary for the successful evolution of the maintenance process and the resulting progress in achieving high level of performance.

This paper is organized as follow. Section 2 presents a literature review, describing maintenance management concepts. Sections 3 and 4 addresses data collection and analysis, evaluating the data obtained from organizations that act in different sectors. Overall findings of the survey are shared in Sect. 5. Section 6 presents the conclusion and future work.

2 Literature Review

2.1 Maintenance Management

Maintenance management establishes goals and objectives through standards and work procedures in order to obtain a better utilization of available resources, which are staff, equipment and materials. Effective maintenance extends equipment life, improves equipment availability and retains equipment in proper condition [2].

Cholasuke et al. [3] identified nine factors associated with high effective maintenance management: policy deployment and organization; maintenance approach (type of maintenance adopted); task planning and scheduling; financial aspect; continuous improvement; human resource management; contracting out maintenance; information management & CMMS; spare parts management.

Marquez and Gupta [4] propose a framework that comprises three pillars to support maintenance management, namely: the information technology pillar, the maintenance engineering and the organizational pillar.

According to Carnero and Noves [5], the complexity of modern industrial unit is increasing and maintenance management is now considered an important factor for improving the performance of the operation, the security, availability, service life and cost reduction and, to this end, the use of computerized systems is essential. These systems are known as computerized maintenance management systems (CMMS). Fernandez et al. [6] states that information is a significant factor for a successful management of maintenance and is the basis for computerized systems. Sherwin [7] explains that, currently, with the advance of computational systems, it is possible to determine a better way to manage the maintenance by optimizing their activities, integrating the maintenance function with the other activities through advanced IT systems (Information Technology), which became more necessary in the current days and, therefore, have become more economically feasible.

2.2 Maintenance Methodologies

Many tools available today have associated the word maintenance. It is important to note that these are not new types of maintenance but tools that allow the application of the main types of maintenance. Highlighted among them are: Maintenance Engineering; Lean Maintenance; Total Productive Maintenance; Reliability Centered Maintenance; Reliability Based Maintenance and Condition Based Maintenance.

Maintenance policies applied correctly aim at preventing and/or eliminating the occurrence of failures. Lack of fulfillment of what was previously defined as "proper performance" is defined as failure.

Alsyouf [8] sustains that proper maintenance practices can contribute to overall business performance through their impact on the quality, efficiency and effectiveness of companies operations.

2.3 Performance Measures

According to Muchiri et al. [9], indicators should support monitoring and control of performance, help the identification of performance and gaps, support learning and continuous improvement, support maintenance actions towards attainment of objectives and provide focus on maintenance resources to areas that impact manufacturing performance. They should be grouped them in maintenance results indicators and maintenance process indicators. The literature, in the context of maintenance, provides various expressions and terminology for performance

indicators, once they eventually adapt to the reality of companies. Campbell [10] proposes a classification for performance indicators as follow:

- Overall maintenance results;
- Maintenance productivity;
- Maintenance organization;
- Efficiency of maintenance work;
- Maintenance costs;
- Maintenance Quality;

Gulati [11] suggests that the first step in developing metrics is to involve people who are responsible for the measurements and ensure that the metric is specific, measurable, attainable, realistic and timely.

EN15341 Standard [12] highlights that maintenance performance is the result of complex activities, which can be evaluated by appropriate indicators to measure actual and expected results. Performance indicators are necessary to ensure stability and predictability of the maintenance process. This standard proposes three classes of indicators, namely: Economic indicators, technical indicators and organizational indicators.

In general, indicators are measures or numerical data set about processes that we want to control and improve. The most commonly used in maintenance are: Availability, Costs, Production losses due maintenance activities, Rework, Mean time between failures (MTBF), Mean time to repair (MTTR) and Overall equipment effectiveness (OEE) [13].

3 Data Collection

3.1 Questionnaire

To gather data about the organization and maintenance management practices adopted by Manaus industrial pole companies, a questionnaire was developed [14]. The questionnaire has five parts that cover the following topics:

- Section I—General Information: this section aims to identify the company profile, activity, number of employees, number of equipment under the responsibility of the maintenance area and origin of the company.
- Section II—Maintenance Management: identify the techniques, strategies and management tools used for maintenance.
- Section III—Maintenance Indicators: identify the degree of use of performance indicators in the management of maintenance.
- Section IV—Procedures and Maintenance Plans: identify the degree of organization and planning of maintenance in the company.
- Section V—Maintenance Staff: assess the degree of organization, development and training of the maintenance staff.

3.2 The Sample

The questionnaire was sent to the industrial plants of the industrial pole of Manaus, in Amazonas state (Brazil), to be filled in by the maintenance area. The industrial pole is organized in 19 sectors, including companies of the plastic industry, manufacturers of mobile phones, modems, set top boxes, televisions, laptops, audio, CD and DVD manufacturing, motorcycles, air conditioners, cameras, alarm and protection systems, naval industry, metallurgical industry and so on [14].

There are approximately 430 companies registered and operating in the industrial pole of Manaus, according to the official document provided by Suframa [15], which is the agency that manages the industrial pole of the region. The questionnaire was sent to all companies registered in the industrial pole and 71 companies from various sectors answered the questionnaire, resulting in a response rate of 16.5 %.

With respect to the origin of the surveyed companies, 44 % have national capital, i.e., they are local, while 56 % have international capital, i.e., are multinational. In relation to staff, 75 % of the companies have more than 100 employees. When referring the total number of equipment under maintenance area responsibility it was observed that about 70 % of them have more than 50 units to manage.

4 Data Analysis

4.1 Maintenance Management Function

Maintaining production equipment is regarded as a strategic factor by the organizations. Most respondents to the questionnaire (90.14 %) stated that the maintenance of production equipment is seen as a strategic factor within the organization.

Regarding outsourcing practices, 76.06 % of companies outsource services related to preventive and corrective activities covering 25 % or less of equipment, as shown by Fig. 1. Besides, concerning companies that outsource maintenance activities, 73.24 % of them affirm that machinery suppliers perform 40 % or less of services.

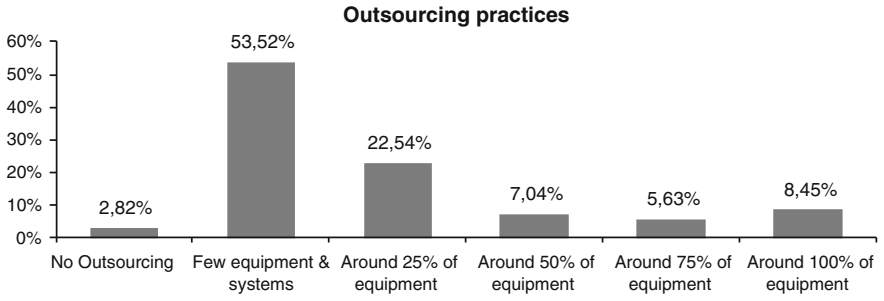


Fig. 1 Outsourcing practices

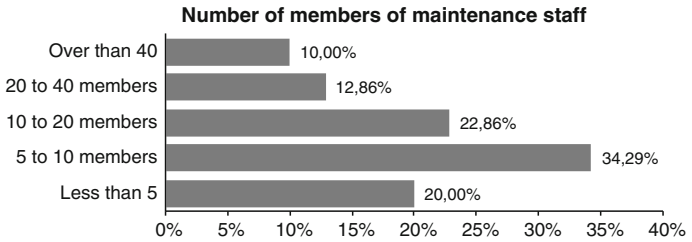


Fig. 2 Number of members of maintenance staff

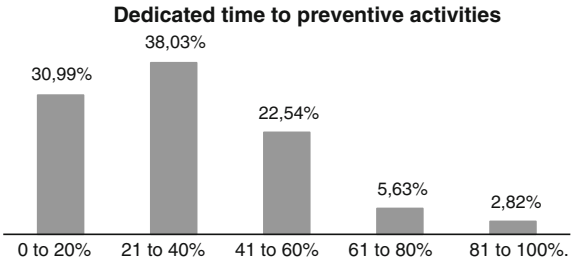


Fig. 3 Dedicated time to preventive activities

The maintenance staff is an important factor that demands strategies to improve competence level. Thus, it is necessary to manage activities and resources allowing the maintenance area to accomplish its function. Figure 2 presents results regarding the size of the maintenance team. It can be noticed that 57.15 % of the companies have between 5 and 20 employees working in the maintenance area and, it was observed that it is linked to the company size and the number of equipment, as it was expected. According to the data obtained in the survey 84.51 % of the companies adopt a preventive policy, 35.21 % assumes that they adopt predictive and 12.68 % detective approach. In terms of maintenance policies, it was noticed that around 15.49 % adopt only corrective policies. According to the answers, preventive activities stand out over the others, although most of the time the maintenance staff is dedicated to corrective. Figure 3 presents the percentage of time that companies dedicate to preventive maintenance activities. Although a large number of companies have indicated the practicing of preventive maintenance, about 69.02 % of companies spend only 40 % or less of their time on preventive activities.

60.56 % of the companies indicate that TPM is the model of maintenance management adopted by the company. However, the majority adopts a strategy focused on production rather than productivity and efficiency, which usually increases the probability of equipment breakdowns.

To analyze the techniques used to assist maintenance, a Likert scale with five levels regarding the frequency of use of such techniques was adopted: (1) Never, (2) Rarely, (3) Occasionally, (4) Often, (5) Very often.

Table 1 presents the average obtained for each technique. The average is not very high, since it is below 4 for all the techniques. Three techniques stand out

Table 1 Techniques utilization

Techniques	Average
5S	3.65
5 Whys	3.32
Cause and effect analysis	3.17
PDCA	2.92
Root cause analysis (RCA)	2.04
Failure mode and effect analysis (FMEA)	2.00
Reliability analysis	1.90
Hazard analysis	1.79
8D	1.76
Fault tree analysis (FTA)	1.59
Tree analysis	1.45
Reliability block diagram (RBD)	1.45
Other	0.34

namely Cause-and-effect analysis, 5S and 5 Whys, while the others are used rarely or never, fact that can be linked to a lack of training in such techniques.

Besides, Table 2 presents a list of main potential difficulties that were evaluated and, limited budget, insufficient parts in stock, insufficient number of technicians and non-compliance of the delivery date by parts suppliers were the main constraints appointed. The percentage of companies that have indicated having these constraints is above 40 %.

Table 2 Maintenance constraints

Constraints	Percentage (%)
Limited budget	66.67
Insufficient spare parts in stock	47.22
Insufficient number of technicians	43.06
Non-compliance of the delivery date by parts suppliers	43.06
Lack of training of technical team	34.72
Inappropriate information management system	30.56
Delay in subcontracted services	29.17
Non-compliance of the delivery date by material suppliers	29.17
Lack of time	27.78
Equipment/maintenance tools insufficient or inappropriate	26.39
Lack of administrative support	23.61
Low competence of technical team	23.61
Low motivation of the technical team	13.89
Other	12.50
Low educational level of the technical team	11.11

4.2 Maintenance Performance Measurements

Regarding performance measurement, it was observed that 78.87 % of companies states using some kind of KPI to evaluate its performance.

A Likert scale with five levels regarding the frequency of use of such KPIS was adopted. The average was calculated and presented in Table 3. As observed, the averages are not very high, since the values are below 4. Companies typically use more frequently economic indicators, availability and downtime to support maintenance staff in seeking improvements such as reducing the waiting time between calls, the time to perform an activity and the downtime due to equipment failures.

MTBF and MTTR are indicators that show how good is the maintenance plan to prevent stoppages just as MTTR to inform how prepared is the team to solve a problem, however they are weakly used.

A Likert scale with five levels regarding the degree of concordance [(1) Totally disagree, (2) Disagree, (3) Indifferent, (4) Agree, (5) Totally agree] about topics associated to downtime, repair time, waiting time and outsourcing was adopted. The average was calculated and presented in Table 4. The averages obtained are low, which means that the degree of concordance is low, except for the first statement. Company, in general, does not recognize that the downtime, the time to

Table 3 KPI adopted by companies

KPI	Average
Downtime	3.35
Availability	3.14
Economic	3.14
MTTF by machine	2.89
OEE	2.63
MTBF by machine	2.59
MTTF by line/area	2.56
Backlog	2.41
MTBF by line/area	2.08
MWT	1.94
Other	0.23

Table 4 Overall maintenance performance

Performance	Average
The downtime of equipment has been decreasing as a result of the effort in this area	3.68
Subcontracting occurs as a result of internal capacity lacking	3.16
The time resolution of faults (repair time) is considered high	2.97
The downtime due to equipment malfunction is considered high	2.97
The waiting time for starting a repair is considered high	2.72

repair, the waiting time to repair is high. However, the result does not seem to reflect the reality, it may be due to the fact that the questionnaire respondents are themselves responsible for the maintenance area and, naturally, there was a tendency to not expose the department itself. They also generally agree with the fact that the downtime has been decreasing as an effort in this area.

In addition, the average associated with the statement "Subcontracting occurs as a result of lacking of interval capacity" is medium, showing that other reasons origin the outsourcing of services, such as lack of competence, for instance.

4.3 Maintenance Planning and Scheduling

In general, all companies have any maintenance plan in which time intervals to execute the maintenance actions are established. The most common time planning frequency is monthly (46.48 %) and yearly (49.30 %) as shown by Fig. 4. Oddly the use of daily planning is low (5.63 %). Considering that many companies in the survey reported that they adopt TPM and, it was expected that daily activities planning would be held more frequently.

It was found that realization of some daily activities such as cleaning, inspection and minor repair by operator is not very frequent. The number shows that the frequency of analysis of failure data is more frequent, however, it is still low. Once again, it was expected that the frequency of cleaning and inspections and failure analysis would be more frequent due to TPM adoption.

A maintenance management system is an essential tool for all organizations, helping to improve maintenance department's efficiency and effectiveness. Figure 5 shows the support tools used by maintenance in its daily activities. 16.90 % of companies adopt a CMMS, 40.85 % make use of spreadsheets, 23.94 % combines CMMS and spreadsheet and 18.31 % use manual registration to organize maintenance activities. Considering those that use any CMMS, 48.28 % adopt standard CMMS available in the market and others make option to develop their own

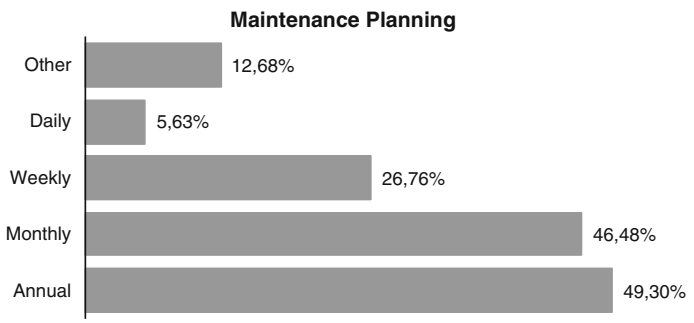


Fig. 4 Maintenance planning

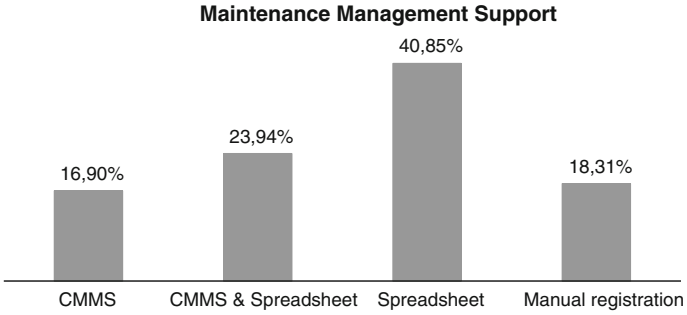


Fig. 5 Maintenance management support

software (17.24 %) or to subcontract a specialized company to develop the software (34.48 %).

Table 5 list some CMMS features. A Likert scale with five levels regarding the frequency of utilization of such features by maintenance staff was adopted. The averages presented on this table are referred only to companies that have these features in the maintenance software. Therefore, it allows to evaluate how useful is the software.

For most of the features the average is considered high (value above 4), which shows a great use of these features by the maintenance staff. The lowest values are registered for: planning to purchase spare parts and materials; registration of work instructions; data analysis for monitoring equipment condition and method to support failure analysis. It was also identify that, major companies applications has

Table 5 CMMS features

Features	Average
Generation and control of work orders	4.64
Historical about preventive maintenance works	4.64
Historical equipment repair	4.61
Equipment, parts and groups registration	4.54
Determination of performance indicators	4.30
Annual planning of preventive tasks	4.28
Monthly planning of preventive tasks	4.27
Maintenance budget	4.09
Allocation human resources to the maintenance	4.04
Inventory control of parts and materials	4.00
Weekly planning of preventive tasks	3.92
Historical about improvement works	3.85
Planning to purchase spare parts and materials	3.70
Registration of work instructions	3.68
Data analysis for monitoring equipment condition	3.44
Method to support failure analysis	3.21

only standards functions, such as equipment record, work orders control and generation, maintenance historical and maintenance activities planning.

4.4 Maintenance Staff

With respect to the organizational structure of the maintenance area, not all companies have a consolidated department, and the teams are usually subordinated to other areas, commonly engineering or production. In 47.89 % of cases, it was noticed that maintenance department has a formal structure, including management positions and other technical functions. Related to education level of maintenance staff, 64.29 % of the companies assumes that 20 % or less of their staff have high graduation which can be considered a low qualification level, i.e., the majority of people in this area has no upper level having only technical skills. Only 5.71 % of the companies state that around 81–100 % of their staff has high graduation, as shown by Fig. 6.

The survey pointed that 59.15 % of the companies recognized having a formal training plan for maintenance staff. However, it is not guaranteed that it will be accomplished due to many constraints.

Table 6 refers to trainings and aims to assess the type of training practiced. A Likert scale with five levels regarding the frequency of trainings was adopted. This table presents the averages obtained for each type of training. The average is higher regards training about equipment. Training in the production process and method of operation as well as in maintenance methodology, techniques and tools has an overall low frequency.

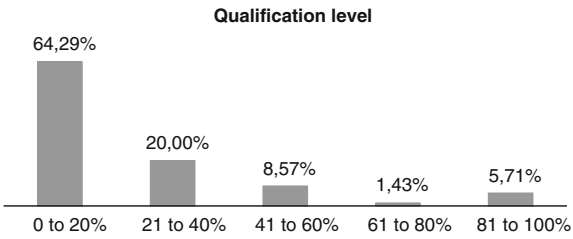


Fig. 6 Qualification level

Table 6 Frequency of trainings

Trainings	Average
Maintenance technicians receive training on equipment	3.54
Maintenance technicians receive training in the production process and method of operation	3.46
Maintenance technicians receive training in methodologies, technique and tools to help maintenance management	2.89

Lastly, the survey identifies that 56.34 % of the companies provide training sessions for 40 % or less of maintenance staff, while only 15.49 % of the companies provides training to around 81–100 % of their team. Therefore, the survey allows us to conclude that the investment in competencies improvements was low in the last year.

5 Survey Findings

In general, maintenance management of the companies of the Manaus industrial pole has many opportunities for improving, especially respecting to prevention and training, regardless of the segment of the company, its origin, its size, number of employees or maintenance team size.

Although most agree that the maintenance of production equipment is seen as a strategic factor within the organization, the survey data show that corrective policies is still predominant, attributed in many cases to the company's own strategy. It was observed a low level of training of the teams, reduced budget affecting the organization and planning of activities, lack of time to perform activities and even outsourced due to lack of own staff or lack of training of the technical team. Another important fact about this issue is that prevention programs are not adequate or sufficient to prevent or reduce the frequency of failures. This is strongly evidenced by the limited use of indicators for maintenance management and of tools, methodologies or techniques to support failure analysis. While companies affirm that activities aimed at reducing the downtime are performed, the percentage of time devoted to preventive activities is very low, according to the collected data. Aspects as limited budget, low number of technicians, parts shortages and reduced training were considered by companies as factors that strongly contribute to a maintenance performance under expectation. Even with respect to the planning of preventive activities, the majority performs annual and monthly planning. Evaluation/revision is relatively insufficient since for a high number of companies, it is made only using the directions offered by manufacturer's manual. Activities plan based on the machine's manual makes the maintenance plan outdated and barely able to act on faults over time.

For most companies, maintenance activities are managed and controlled without the use of CMMS, which shows a potential for growth in this area, since the majority adopts manual records or spreadsheets, not allowing better efficiency in the data analysis to improve planning and control activities. With respect to those who adopt some kind of CMMS, some had to adapt it according to their needs especially in solutions for reporting, allowing a better analysis of the results achieved and the consequent proposal to improve them. Moreover, most of those that adopt some sort of CMMS have acquired it on the market and these, in turn, are limited in features.

Regarding the use of techniques, the most elementary are used, particularly PDCA Cycle, 5S or 5 Whys, and there is a strong adherence to TPM though

without much practical result. The use of indicators for maintenance management was also considered as elementary, where a significant number ignores the specific terms, their composition and how to interpret them. In particular, the most used are Downtime, Economic, Availability, MTTR, MTBF, and OEE.

Indeed, the use of a suitable CMMS according to reality or maturity level of the organization would allow a better planning of activities, use of more appropriate management indicators as well as having added a support tool to failure analysis.

Regarding the structure of the maintenance department, few of them are autonomous, since, usually, maintenance is in charge of the engineering and/or production. In general, companies have a highly technical maintenance group without higher hierarchies. In addition, factors such as training and education level were considered points of attention and real opportunity for improvement.

Finally, the survey data show that companies of the industrial polo have a significant margin of costs reduction. It allowed the identification of improvement opportunities with various gains, such as: reducing losses in the production process, increasing operational efficiency and asset availability and achieving better control of activities.

6 Conclusion

The effectiveness of the maintenance function in an industrial unit depends on the equipment involved, the training of personnel, and mainly on the adopted strategy for maintenance management. In addition to modern equipment ownership, it is necessary to understand the concern about flaws, in its details, in order to attack not the consequences but the causes using the most appropriate tools and techniques. Maintenance area should advocate the existence of an effective planning and monitoring of activities using the most appropriate resources and applying the more advantageous tools and techniques.

Most companies have a basic level of maintenance management, which means that they have the opportunity to make improvements and to obtain important gains. They have also the capacity to become more competitive, increasing throughput and reducing losses.

This survey shows many opportunities to make improvements in the maintenance area in general, such as using CMMS, implementation of deep preventive culture, adopting meaningful KPIs and reinforce training plan to the maintenance team.

As future work, the authors aim to perform hypothesis tests with the answers of the questionnaire in order to validate some statements related to maintenance management.

Acknowledgment This work has been supported by FCT—Fundação para a Ciência e Tecnologia within the Project Scope: PEst-OE/EEI/UI0319/2014.

References

1. Parida A, Kumar U (2006) Maintenance performance measurement (MPM): issues and challenges. J Qual Maintenance Eng 12(3):239–251
2. Swanson L (2001) Linking maintenance strategies to performance. Int J Prod Econ 70(3):237–244
3. Cholasuke C, Bhardwa R, Antony J (2004) The status of maintenance management in UK manufacturing organizations: results from a pilot survey. J Qual Maintenance Eng 10(1):5–15
4. Marquez AC, Gupta JND (2004) Contemporary maintenance management: process, framework and supporting pillars. Int J Manage Sci 34:313–326
5. Carnero MC, Novés JL (2006) Selection of computerized maintenance management system by means of multicriteria methods. Prod Plann Control 17(4):335–354
6. Fernandez O, Labib AW, Walmisley R, Petty DJ (2003) A decision support maintenance management system: development and implementation. Int J Qual Reliab Manage 20(8):965–979
7. Sherwin DJ (2000) A review of overall models for maintenance management. J Qual Maintenance Eng 6:138–164
8. Alsyouf I (2009) Maintenance practices in Swedish industries: survey results. Int J Prod Econ 121:212–223
9. Muchiri P, Pintelon L, Gelders L, Martin H (2011) Development of maintenance function performance measurement framework and indicators. Int J Prod Econ 131:295–302
10. Campbell JD, Jardine AKS (2001) Maintenance excellence: optimizing equipment life-cycle decisions. Marcel Dekker, Ink, USA
11. Gulati R, Smith R (2009) Maintenance and reliability best practices. Industrial Press Inc., New York
12. EN15341:2007 Maintenance—maintenance key performance indicator. European Committee for Standardization (CEN), Brussels
13. Oliveira M, Lopes I, Figueiredo D (2012) Maintenance management based on organization maturity level. In: International conference on industrial engineering and operations management 2012, ICIEOM 2012, 9–11 February, Guimarães
14. Oliveira M, Lopes I, Figueiredo D (2014) Maintenance management practices of companies of the industrial pole of Manaus. In: Proceedings of the World congress on engineering and computer science 2014, WCECS 2014. Lecture Notes in Engineering and Computer Science, vol II, 22–24 October, San Francisco, pp 1016–1022
15. Suframa Profile [Online]. Available: http://www.suframa.gov.br/zfm_industria.cfm

Characterizations Severe Plastic Deformation of Copper Processed by Equal Channel Angular Pressing Technique

Sanusi Kazeem Oladele, Afolabi Ayo Samuel and Muzenda Edison

Abstract The equal channel angular pressing technique which is one of severe plastic deformation is now recognized for achieving very significant grain refinement of ultra-fine grained materials. This study reports the results of the mechanical tests and the microstructural analysis carried out on the specimens of ultra-fine grained copper processed by ECAP technique at room temperature using a die with a 126° between the die channels. Hardness test and tensile tests were conducted for samples cut out in two different directions to evaluate the mechanical properties. The microstructural characterization was carried out using optical electron microscope (OEM) and scanning electronic microscope (SEM). The results show ECAP technique introducing significant grain refinement and produced ultra-fine grains in copper and there is a potential for achieving high ductility and hardness properties in the copper alloy after processing.

Keywords Copper · ECAP · Mechanical properties · Microstructural evolution · Severe plastic deformation · Ultra-fine grained

1 Introduction

The manufacturing and processing of ultra-fine grain (UFG) materials have attracted growing scientific and industrial interests in the last decade as a result of the novel and attractive properties of these materials [1, 2]. These UFG materials

S.K. Oladele (✉) · M. Edison
Faculty of Engineering and the Built Environment, Department of Chemical Engineering, University of Johannesburg, Doornfontein Campus, Johannesburg 2028, South Africa
e-mail: Sansuik@gmail.com

A.A. Samuel
Department of Civil and Chemical Engineering, College of Science, Engineering and Technology, University of South Africa, Private Bag X6, Florida, Johannesburg 1710, South Africa

H.K. Kim et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-94-017-7236-5_36

exhibit a wide variety of unique properties that result from a large volume fraction of their grain and/or interphase boundaries [3–5]. These materials have mechanical properties that include extraordinarily high yield strength, high hardness, improved toughness and ductility with increasing strain rate [1, 6–8]. They also been found to exhibit marked different microstructures and mechanical behaviours from their conventional coarse grained polycrystalline counterparts, thus UFG materials have enhanced super-plasticity deformation at low and high strain rate [7]. Severe Plastic Deformation (SPD) is an effective approach to produce UFG materials and the technique has been use for the past decade [8–11]. The SPD method has provided new opportunities in the investigation of enhanced superplasticity in metallic materials. Any means of introducing large plastic strains in metals may lead to the reduction of the grain size but without any significant change in the overall dimensions of the materials [12]. The major SPD methods used are Equal Channel Angular Press (ECAP) and High Pressure Torsion (HPT) [13]. ECAP is a processing technique in which an intense plastic strain is imposed on a polycrystalline sample by pressing the sample through a special die in other to produce large fully dense samples containing an ultrafine grain size in the sub-micrometre or nanometre range [14, 15]. The idea of ECAP was the results of the early works of Segal and co-workers which showed the technique of pressing test samples through a die containing two channels, equal in cross section, intersecting at an angle of ϕ. [16–19]. As a result of the pressing, the sample underwent simple shear but it retained the same cross-sectional area so that it was possible to repeat the pressings for several cycles [20]. To accumulate very large strains, a sample can be forced to pass through the die several times and the strain path can be easily changed by turning the sample around its longitudinal axis between subsequent passes. Four standard routes are often used and they are A, BA, BC, and C [21–23]. A sample is rotated around its axis to an angle of 0°, 90°, and 180° for the routes A, B, and C, respectively. When using route BA, consecutive 90° rotations have opposite senses, while in route BC the sample is rotated in the same direction. ECAP is currently being widely investigated because of its potential to produce ultrafine grained microstructures in both pure metals and alloys.

The materials processed by ECAP also have the advantage of formation of an UFG structure with mainly high-angle grain boundaries, the absence of macroscopic damages and cracks in the samples, Microstructural homogeneity in the most volume of the samples, and formation of equiaxed grains. Research on the ECAP method of for UFG and nanocrystalline (NC) microstructure forming in soft metals, alloys and other materials has been carried out by many researchers [10, 15, 24]. The aim of this study is to investigate the mechanical properties and formation of structures results in copper produced by ECAP using six and twelve passes. The copper sample employed in this work was subjected to tensile test in perpendicular and parallel directions and hardness test to obtain information on the mechanical properties of UFG produced by this technique.

2 Materials and Methods

A sample of copper alloy obtained as extruded rods was used as starting materials in this study. This material was selected due to its good electrical and thermal conductivity, workability, corrosion resistance, and strength. The copper alloy also shows significant strain hardening, strain rate sensitivity and temperature dependence of plastic flow behaviour. The composition of the copper alloy used for the experiment is shown in Table 1. The ECAP die was constructed using K510 silver quenched and tempered steel. The die was a relatively simple design. It was machined as a two-piece split die, consisting of a highly polished smooth plate bolted to a second polished plate. The angle between the channels is 126° and the external curvature is 0°. The channel section is 14.5 mm in diameter. The die used has the same inlet and outlet channels with nearly identical dimensions. Since the cross-sectional dimensions of the sample remain unchanged on passage through the die, repetitive pressings was used to attain very high strains [11, 25]. Figure 1 shows the ECAP die used for the experiment. The specimens were deformed by passes through the die using routes Bc. Route Bc (samples rotation 180° after each pass) was chosen because of its tendency to develop a substructure after fewer ECAP passes than the other routes [15, 16, 24, 26]. Six and twelve passes of deformation were carried out at room temperature, with Molybdenum disulfide (MoS_2) used as lubricant in 63 tons pressing machine. The material was easily deformed at room temperature, because of its excellent ductility. Each specimen was removed from the die by pressing the next specimen into the die and final specimen inside the die was removed using a dummy specimen which then remains

Table 1 Chemical composition of copper alloy used

Elements	Cu	P	Sn	Ti	Cr	Mn	Fe	Si	Zn	Pb
wt%	90.77	0.61	1.20	0.30	0.49	0.57	0.77	0.08	1.56	2.93

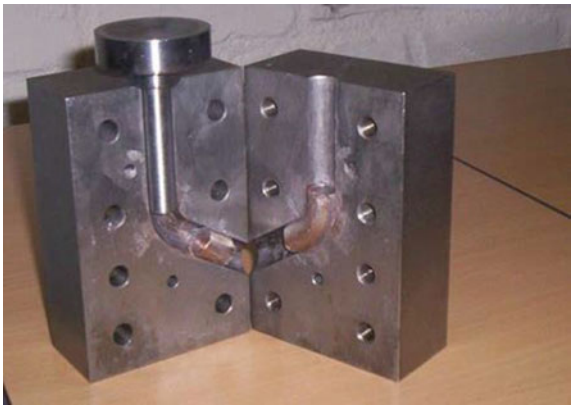


Fig. 1 The ECAP die for the experiment

within the die. For the tensile tests, the samples were cut in two directions parallel and perpendicular to the ECAP axis with gauge lengths of 3.75 mm and gauge diameter and thickness of 0.74 and 0.60 mm, respectively for tensile testing using an Instron testing machine with tensile velocity 0.00375 mm/s. The microstructures of the UFG and coarse-grained samples were investigated prior to and after the ECAP tests using optical electron microscopy (OEM) and scanning electron microscopy (SEM) was used in this study for measuring grain size, and to examine the crack morphology of ECAP materials to relate the fracture features to the microstructure and mechanical properties.

3 Results and Discussion

3.1 Tensile Test

The stress–strain curves of as-received, six passes, and twelve passes of tensile copper samples tested in parallel direction are shown in Fig. 2, while the stress–strain curves of the tensile samples tests in perpendicular directions are shown in Fig. 3. From the figures the mechanical strength increases with increasing deformation, the strength is higher in the material subjected to ECAP technique compared to as received samples both in two directions. The specimens subjected to six passes had better strength than the specimen subjected to twelve passes. The same results were obtained when the tensile properties were measured in perpendicular direction.

For the copper specimen processed at six passes, the tensile strength measured in perpendicular direction has an improved strength than in parallel direction as shown

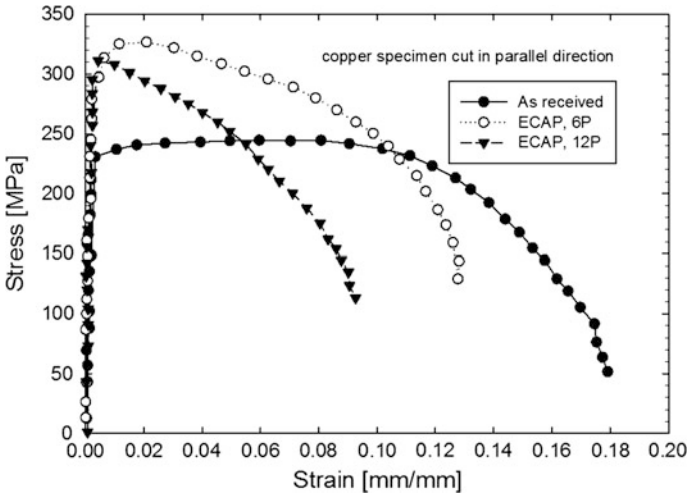


Fig. 2 Stress–strain curves for the copper alloy samples cut in parallel direction

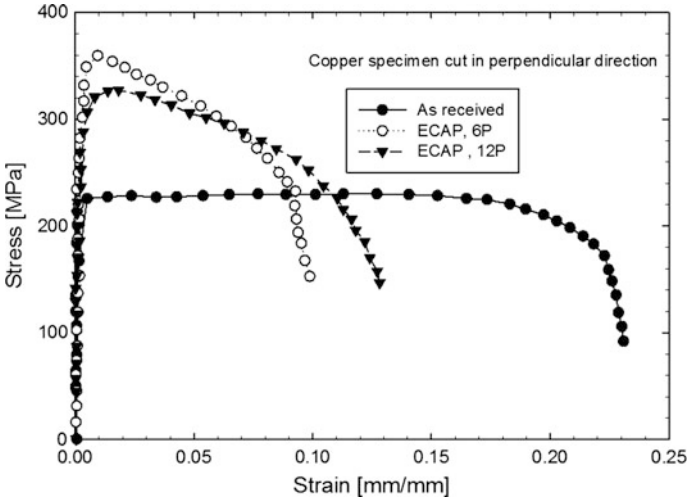


Fig. 3 Stress–strain curves for the copper alloy samples cut in perpendicular direction

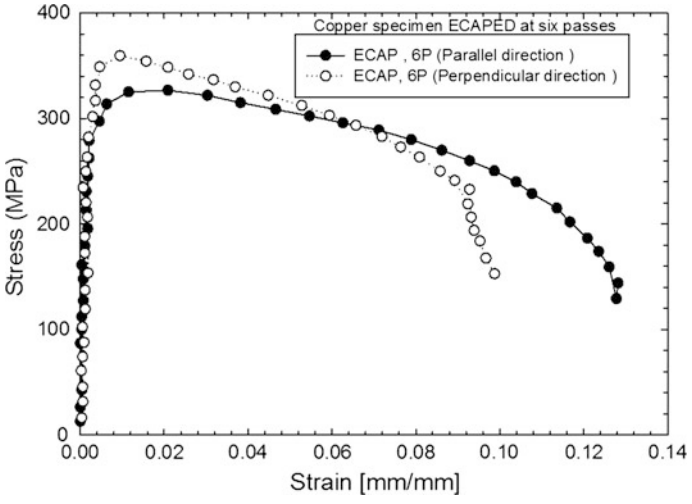


Fig. 4 Stress–strain curve for the copper alloy specimen at six passes and cut in parallel and perpendicular directions

in Fig. 4, the same observation is visible for the samples processed at twelve passes shown in Fig. 5. Based on the obtained results, one can conclude that the tested samples are characterized by significant differences of strength properties depending on the direction. Samples taken in parallel direction are characterized by clear yield and better mechanical strength related properties than samples taken at perpendicular direction.

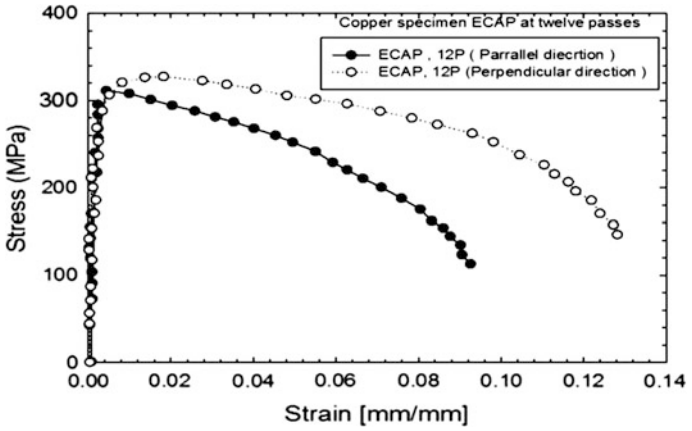


Fig. 5 Stress–strain curve for the copper alloy specimen at twelve passes and cut in parallel and perpendicular directions

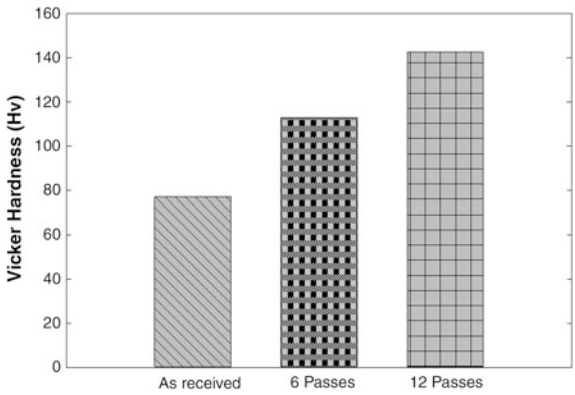


Fig. 6 The Vickers hardness results after a number of passes

3.2 Hardness Tests

Figure 6 shows the Vickers hardness results for as received copper after six and twelve passes. From the figure the Vickers hardness value for as received copper 77 Hv while for the copper process by ECAPECAP technique after six and twelve passes are 113 and 142.5 respectively. It can be deduced from the results that the hardness properties of copper was improved and also increased with the numbers of passes used.

3.3 Microstructural Evolution

Figure 7a shows the Optical microscope image of the as-received copper. While the Optical microscope image of the material copper alloy subjected to ECAP

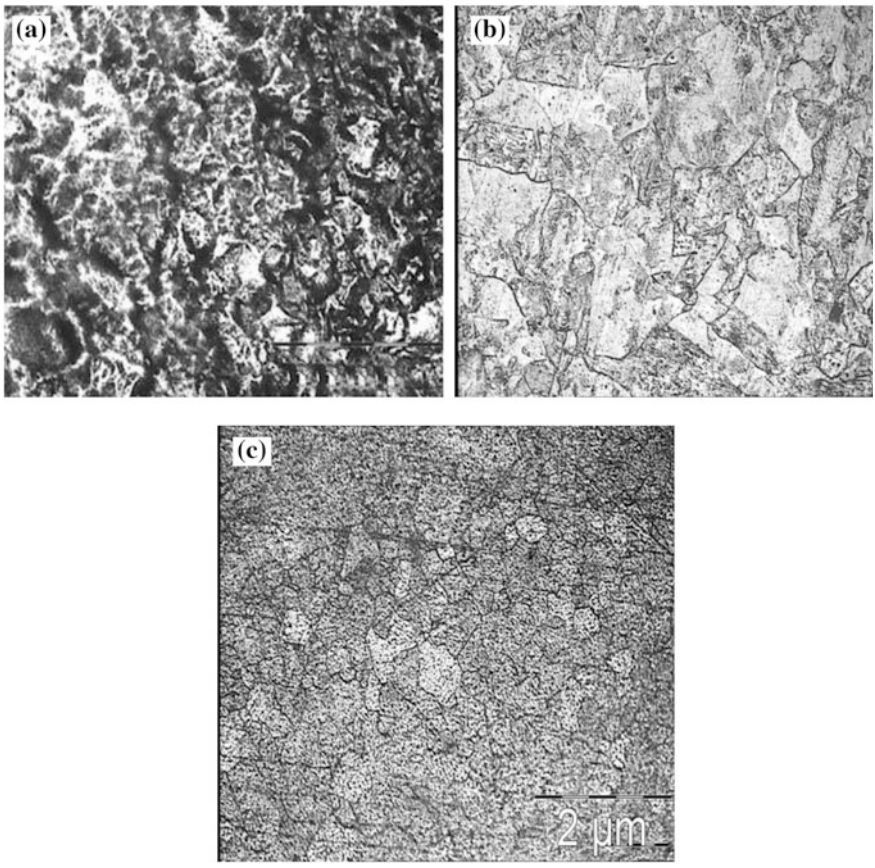


Fig. 7 **a** Optical microscope images of as-received copper. **b** Optical microscope images of the ECAP copper obtained at 6 passes. **c** Optical microscope images of the ECAP copper obtained at 12 passes

technique at six passes and twelve passes respectively are shown in Fig. 7b, c respectively. The microstructure of copper after the process is greatly reduced and presents relatively homogeneous grain size compare with the as received copper. The copper processed after six passes had an average grain size of about 0.75 μm and the copper processed after twelve passes had an average grain size of about 0.34 μm. The homogeneous microstructure and grain size refining are due to deformation process.

Figure 8c shows the fracture surfaces of copper when processed. A mixed morphology of shallow dimples and tearing ridges were formed on fracture surface. The size and depth of the dimple of the material depends on plastic deformation capability and the better the plastic deformation ability of the metal, the more prone the necking is and the greater the size of the micropores.

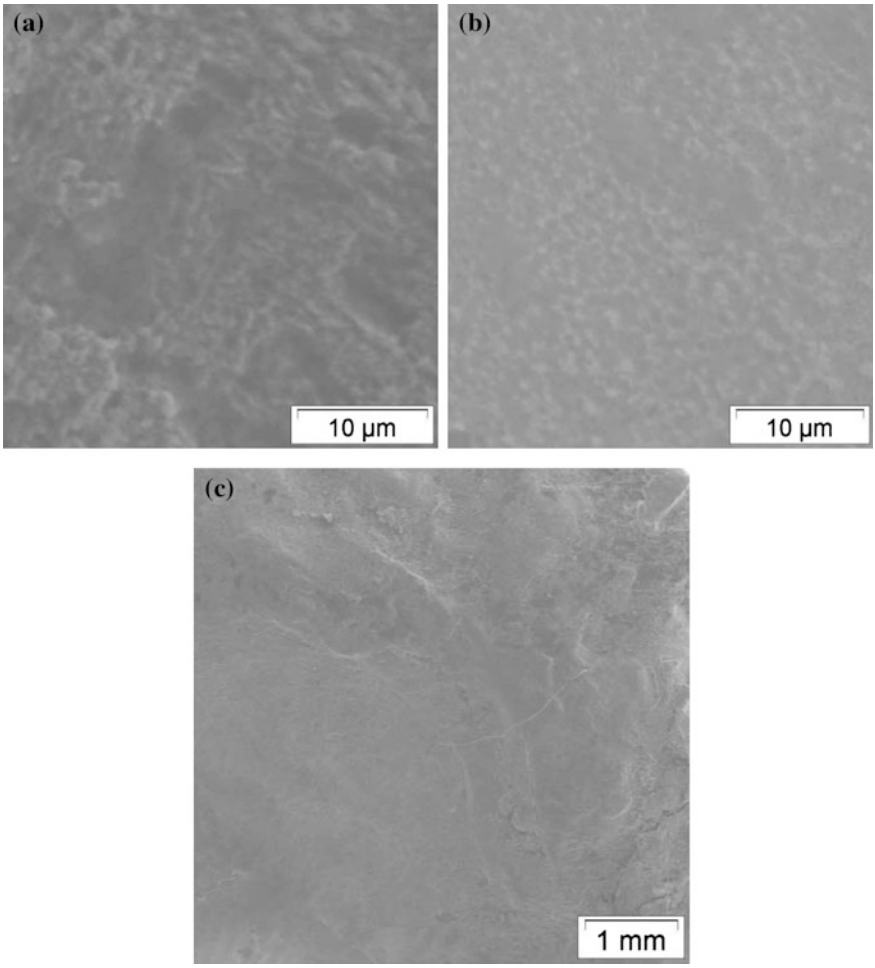


Fig. 8 **a** SEM image of the ECAP copper obtained at six passes. **b** SEM image of the ECAP copper obtained at twelve passes. **c** SEM image of Fracture surface of processed UFG copper

4 Conclusions

In summary, copper alloy was satisfactorily processed by ECAP technique at room temperature using B_c route and the following conclusions were obtained.

1. Processing by ECAP technique introducing significant grain refinement and produced ultrafine grains in copper.
2. The microstructure of copper after the process was greatly reduced and presented a relatively homogeneous grain size compared with the as received copper.

3. These processed copper samples exhibited improved mechanical properties.
4. The tested samples are characterized by significant differences of strength properties depending on the direction.
5. A mixed morphology of shallow dimples and tearing ridges were formed on fracture surface of the sample due to plastic deformation capability.

Acknowledgments The authors gratefully acknowledge the financial supports of the National Research Foundation (NRF) and Faculty of Engineering and the Built Environment, Department of Chemical Engineering University of Johannesburg, South Africa.

References

1. Gleiter H (2001) Nanostructured materials: basic concepts and microstructure. Acta Mater 48:1–29
2. Sanusi KO, Afolabi AS, Muzenda E (2014) Microstructure and mechanical properties of ultra-fine grained copper processed by equal channel angular pressing technique. In: Proceedings of the World congress on engineering and computer science 2014,WCECS 2014,22–24 October, San Francisco, pp 1049–1053
3. Gleiter H (2000) Nanostructured materials: basic concepts and microstructure. Acta Metall 48:1–29
4. Furukawa M, Horita Z, Nemoto M, Langdon T (2002) The use of severe plastic deformation for microstructural control. Mater Sci Eng A 324:82–89
5. Kwon Y-J, Shigematsu I, Saito N (2003) Production of ultra-fine grained aluminum alloy using friction stir process. Mater Trans 44(7):1343–1350
6. Valiev R (2003) Paradoxes of severe plastic deformation. Adv Eng Mater 5(5):296–300
7. Kumar K, Swygenhoven H, Suresh S (2003) Mechanical behaviour of nanocrystalline metals and alloys. Acta Mater 51:5743–5774
8. Langdon T (2011) Processing by severe plastic deformation: historical developments and current impact. Mater Sci Forum 9(4):667–669
9. El-Rayesa MM, El-Danafa EA (2012) The influence of multi-pass friction stir processing on the microstructural and mechanical properties of Aluminum Alloy 6082. Mater Process Technol 212:1157–1168
10. Berbon P, Tsenev N, Valiev R, Furukuwa M, Horita Z, Nemoto M (1998) Fabrication of bulk ultrafine-grained materials through intense plastic straining. Metall Materials Trans 29:2237–2243
11. Horita Z, Furukawa M, Nemoto M, Langdon TG (2000) Development of fine grained structures using severe plastic deformation. Mater Sci Technol 16(11–12):1239–1245
12. Valiev R (2003) Paradoxes of severe plastic deformation. Adv Eng Mater 5(5):296–300
13. Kurzydlowski K (2004) Microstructural refinement and properties of metals processed by severe plastic deformation. Bulleting of the polish academy of sciences and technical sciences, vol 52, no. N4
14. Segal V, Reznikov V, Drobyshevskiy A, Kopylov V (1981) Metally. Russ Metall 1(115):99
15. Iwahashi Y, Horita Z, Nemoto M, Langdon TG (1998) The process of grain refinement in equal-channel angular pressing. Acta Mater 46(9):3317–3331
16. Valiev R, Langdon T (2006) Development in used of ECAP processing for grain refinement. Rev Adv Mater Sci 13:15–26
17. Segal V (1999) Equal channel angular extrusion: from macromechanics to structure formation. Mater Sci Eng A 271:322–333

18. Semiatin S, DeLo D (2000) Equal channel angular extrusion of difficult-to-work alloys. Mater Des 21:311–322
19. Kommel L, Hussainova I, Volobueva O (2007) Microstructure and properties development of copper during severe plastic deformation. Mater Des 28:2121–2128
20. Iwahashi Y, Wang J, Horita Z, Nemoto M, Langdon TG (1996) Principle of equal-channel angular pressing for the processing of ultra-fine grained materials. Scripta Mater 35(2):143–146
21. Valiev R, Islamgaliev R, Alexandrov I (2000) Bulk nanostructured materials from severe plastic deformation. Prog Mat Sci 45(2):103–189
22. Zhu Y, Langdon T (2004) Fundamentals of nanostructured materials by severe plastic deformation. J Met 56(10):58–63
23. Segal VM (1995) Material processing by simple shear. Mater Sci Eng A 197:157–164
24. Furukawa M, Iwahashi Y, Horita Z, Nemoto M, Langdon T (1998) The shearing characteristics associated with equal-channel angular pressing. Mater Sci Eng A 257:328–332
25. Horita J (2005) Production of ultrafine-grianed structures using equal-channel angular pressing. J Jpn Weld Soc 74:88
26. Valiev RZ, Langdon TG (2006) Development in used of ECAP processing for grain refinement. Rev Adv Mater Sci 13:15–26

The Effects of Microstructural Evolution and Mechanical Behaviour of Unalloyed Medium Carbon Steel (EN8 Steel) After Subsequent Heat Treatment

Kazeem Oladele Sanusi, Cullen Mayuni Moleejane, Olukayode Lawrence Ayodele and Graeme John Oliver

Abstract The effect of microstructural evolution on mechanical properties of unalloyed medium carbon (EN8) steel with emphasis on the effects of grain size within solid phase mixtures and the mechanical response of the material is investigated and reported. Specimens with a range of microstructures (grain size and phase) were prepared by heat treatment. The microstructures were carefully characterized using both optical electronic microscope (OEM) and scanning electron microscope (SEM) and the mechanical properties were studied using tensile and hardness tests. The results indicated that structural parameter that directly controls the yield strength, the ultimate tensile strength and elongation at failure have some influence (direct or indirect) on the stress flow and formability of the material.

Keywords Grain sizes · Hardness test · Heat treatment · Mechanical properties · Medium carbon steel · Microstructures · Solid phase mixtures · Tensile test

K.O. Sanusi (✉)
Faculty of Engineering and the Built Environment, Department of Mechanical Engineering Science, University of Johannesburg, Johannesburg 2006, Gauteng, South Africa
e-mail: sanusik@gmail.com

C.M. Moleejane · O.L. Ayodele · G.J. Oliver
Faculty of Engineering, Department of Mechanical Engineering, Cape Peninsula University of Technology, P.O. Box 1906 Bellville Campus, Cape Town, South Africa
e-mail: moleejc@eskom.co.za

O.L. Ayodele
e-mail: AyodeleO@cput.ac.za

G.J. Oliver
e-mail: oliverg@cput.ac.za

H.K. Kim et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-94-017-7236-5_37

1 Introduction

There are numerous metallurgical variables (composition and process parameters) that influence the physical and mechanical properties of materials and the prediction of mechanical properties of materials is more complicated because of different mechanisms that come in play such as chemical composition, the evolution of deformation microstructure, amount of deformation, heat treatment profile, and average grain size distribution. Many of the important mechanical properties of steel, including yield strength and hardness, the ductile-brittle transition temperature and susceptibility to environmental embrittlement can be improved considerably by refining the grain size [1–4]. With the knowledge of dynamical interplay between deformation and grain microstructures size, it is possible to predict the behavior of steel when subjected to various working conditions [5, 6]. Numerous aspects of microstructures and their effects on mechanical behaviour of metals have been studied, including grain size [1, 7], grain boundary [8–10], crystal structure [11], crystal orientation [12]. The production of different Materials with the controlled manipulation of their microstructure in order to improve properties is an active field of study in materials science [1, 13, 14]. In general, as the average grain size decreases, the metal becomes more resistant to plastic flow (yield strength increase) and as the grain size increases, the opposite effect on strength occurs (yield strength decreases) [9, 13, 14]. EN8 steel is an unalloyed medium carbon steel with good tensile strength. This steel is suitable for the manufacture of parts such as general purpose axles and shafts, gears, bolts and studs. The materials is usually recommended by the suppliers to carried out heat treatment after initial stock removal to achieved better mechanical properties toward the core. The aim of this study is to investigate the microstructure of EN8 steel by the size and morphologies of constituent phases and or the grains after subsequent heat treatment, also using a tensile and hardness tests to study the effects of heat treatment on the mechanical property of this material From this study, the influence of grain size on mechanical behaviour of steel based on heat treatment history can be understood.

2 Materials and Methods

The test specimen used in this investigation was unalloyed carbon steel designated as (EN8), the steel was received as rolling bar. The chemical composition was determined and is given in Table 1. Tensile specimens were machined in standard dimensions and prepared using heat treatment experiments. They were divided into four sets for tensile tests. These sets were shown in Table 2. Scanning electron microscopy (SEM) and optical electronic microscope (OEM) were used for phase identification and microstructure characterization of these samples. The test specimen was examined for the following features: The size, shape and type of grain present. The aim of the heat treatment is to obtain a range of microstructures (grain

Table 1 Chemical composition of EN8

Element	C	Si	P	Mn	S	Fe
Composition (wt%)	0.036	0.02	0.025	0.54	0.05	Balance

Table 2 Heat treatment regime used produces various microstructural features

Specimen	Heat treatment
A	Austenised at 950 °C, held for 180 min and cooled in furnace.
B	Austenised at 914 °C, held for 10 min, cooled in furnace to 680 °C
C	Austenised at 914 °C, held for 10 min, cooled in furnace
D	Austenised at 914 °C held for 3 min cooled in furnace to 715 °C and quenched in oil

size and phase). The mechanical properties were determined through tensile test and hardness test.

3 Results and Discussion

Figure 1a show the SEM and Fig. 1b OEM image of as-received EN8 steel used for this experiment, while Fig. 2 show the SEM images of the specimens A, B, C and D prepared and Fig. 3 shows the OEM images of the specimens A, B, C and D prepared according the heat treatment regime described in Table 2. It can be observed from these Figures that the heat treatment regime employed produced five different microstructures. The images show both grains and Solid phase mixtures present in the specimens. Another interesting feature that can be observed with

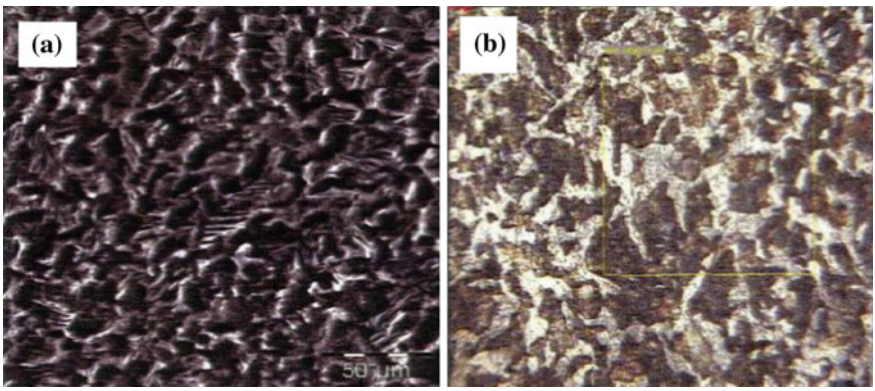


Fig. 1 **a** SEM and **b** OEM images of the as received EN8 steel samples of the as received EN8 steel samples

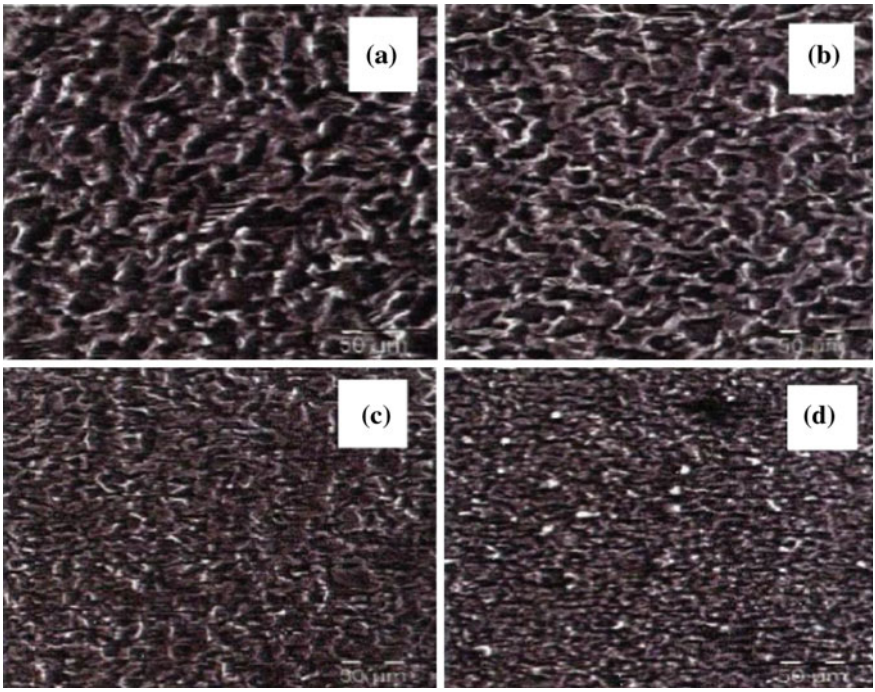


Fig. 2 SEM images of samples prepared according the heat treatment regime described

SEM and OEM is the individual grain boundaries, which can be distinguished quite easily at higher magnifications (see Fig. 4).

It can be observed that the images shown in Figs. 2 and 3 are populated by different phases. The heat treatment processes employed produced multiphase microstructures with different morphologies. The microstructures of specimens A and B (Figs. 2a and 3a) are mainly composed of ferrite and pearlite. However, a slight change in morphology and marked coarsening of microstructure are observed in specimen B. Micrograph B, presented in Fig. 2c, has a similar microstructure to micrograph D, basically a ferritic matrix with second phase islands. However, the heat treatment temperature and soaking/holding time provided two interesting effects: heating steel at 914 °C and reducing holding time at this temperature resulted in formation of small grains in micrograph D (Fig. 2d). On the other hand, increasing holding time at this temperature produced relatively large grains in micrograph C. The fast cooling rates in heat treatments C and D favoured the formation of the phases of martensite and bainite, in comparison with heat treatment B. whereas heat treatment B did not favour the formation of bainite and martensite. The microstructure obtained in heat treatment C presented a similar fourth phase volume fraction in comparison to the microstructure obtained in treatment D. This is suspected to be martensite.

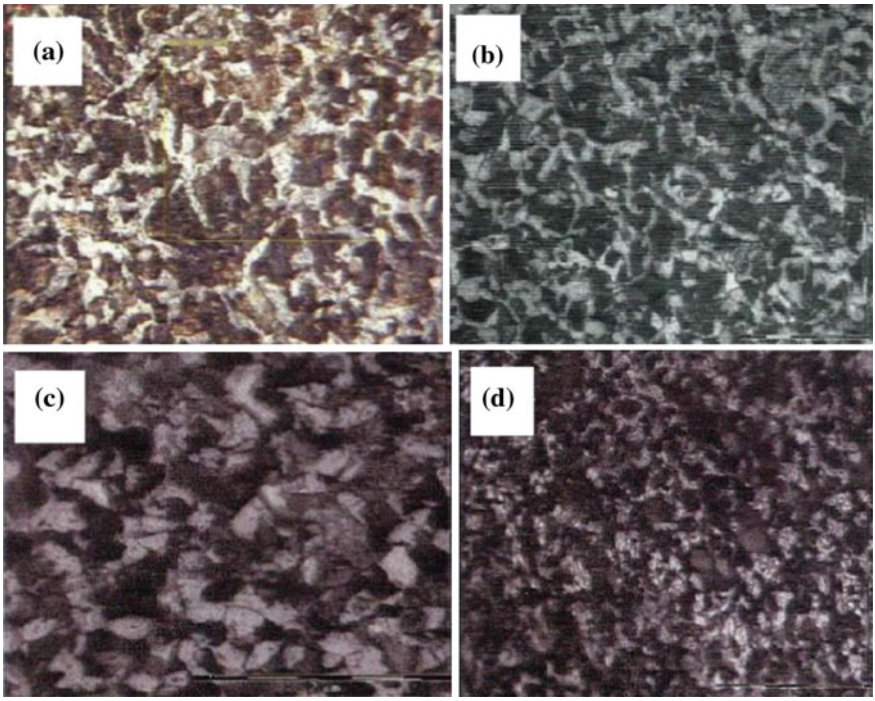


Fig. 3 OEM images of samples prepared according the heat treatment regime described

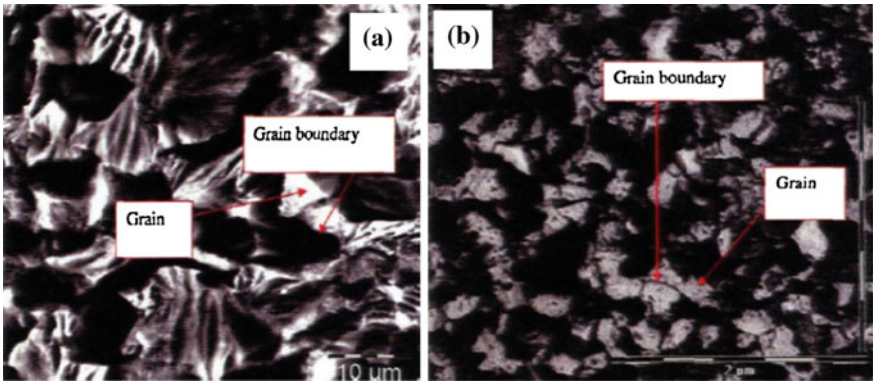


Fig. 4 **a** SEM micrography showing individual grains and grain boundaries. **b** OEM micrography showing individual grains and the grain boundaries obtained in the heat treatment B. Ferrite (*clear*), martensite and bainite (*dark*)

Table 3 shows the volume fraction of the steel phases obtained from statistical analysis of OEM for all the microstructures obtained by the heat treatments A, B, C and D investigated in this study and Table 4 shows the average grain size and

Table 3 Phase volume obtained from OEM analysis

Specimen	Phase 1	Phase 2	Phase 3	Phase 4
	% volume	% volume	% volume	% volume
A	21.05	74.94	3.91	0.13
B	17.18	75.95	6.71	0.16
C	16.09	73.56	7.00	3.35
D	23.04	61.62	8.38	6.96

Table 4 Grain size and Brinell hardness values of EN8 studied

Specimen	Average grain size (μm)	Hardness (Brinell)
A	123.46	164
B	156.25	147
C	46.3	194
D	39.6	206

hardness test results of the specimen. It can be observed that an increase in time at an inter-critical temperature increased the grain size. Heat treatment of specimen D produced microstructure with small grains, while heat treatment of specimen B promoted grain growth. From micrograph C results, it can be observed that the soaking time above critical temperature had a major influence on the microstructure. Heat treatment either at or above the critical temperature caused metallurgical changes and soaking time favoured the growth of grains. The distribution of the phases and grains also showed a continuous change of the microstructure with increasing temperature and soaking time. The Grain size values of EN8 studied is shown in Fig. 5. As reported by Tamehiro and Nakasugi [15] grains generally grow, coalesce and coarsen with increasing temperature and soaking time, Brunig [16] reported that nucleation of micro-void in ductile material is generally governed by large strains of the material. They initiate at discontinuities soon after the onset of plastic yielding and grow due to plastic deformation [17] of the surrounding matrix

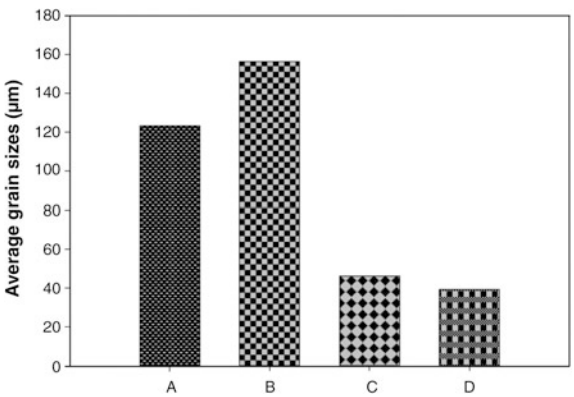


Fig. 5 Grain size values of EN8 studied

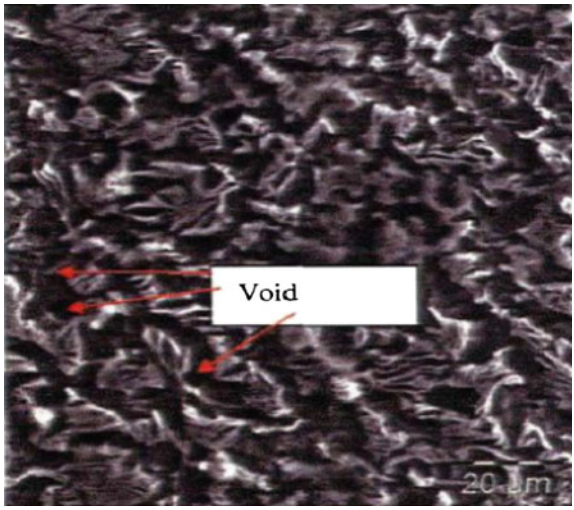


Fig. 6 SEM micrography showing micro-void nucleation

material [18]. Upon further plastic deformation of the material the micro-voids tend to grow and possibly change shape by means of deviatoric and volumetric strain [19–21] which may result in the formation of micro-cracks as micro-voids start to combine micro-void nucleation at the areas of large strains (at the necked area) was distinctly evident (see Fig. 6). Figure 7 also shows the micrographs of EN8 showing variation in grain size where specimen A with larger grain sizes and Specimen D with smaller grain sizes.

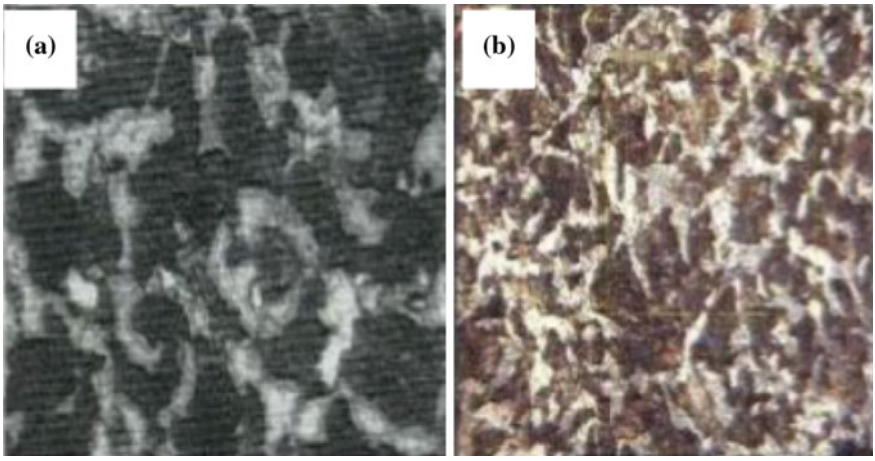


Fig. 7 Micrographs of EN8 showing variation in grain size **a** specimen A with larger grain sizes. **b** Specimen D with smaller grain sizes

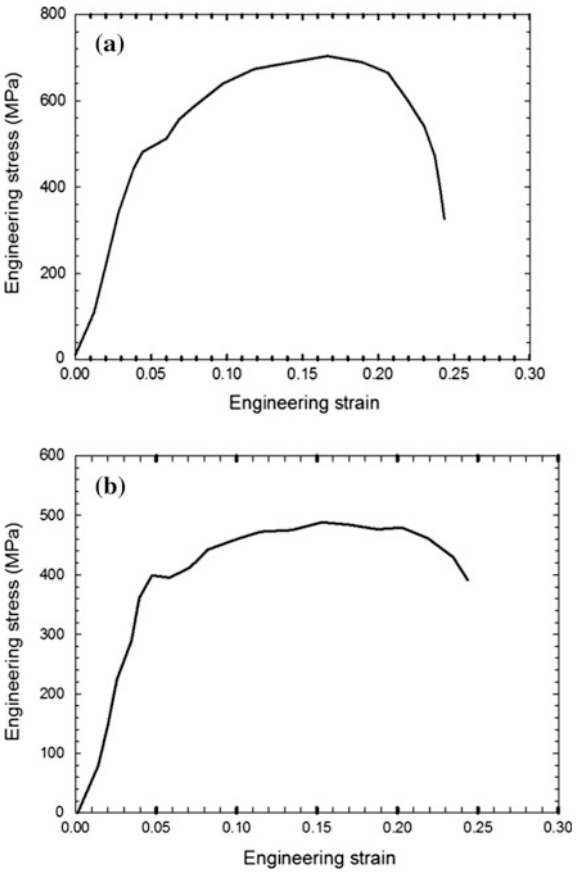


Fig. 8 **a** Stress versus strain curve of EN8 austenised at 950 °C, held for 180 min and cooled in furnace. **b** Stress versus strain curve of EN8 austenised at 914 °C, held for 10 min, cooled in furnace to 680 °C

Figure 8a shows the stress versus strain curves of EN8 austenised at 950 °C, held for 180 min and cooled in furnace. The yield strength (YS), ultimate tensile strength (UTS) and the corresponding elongation of the EN8 austenised at 950 °C, held for 180 min and cooled in furnace were 475.22 MPa, 694.68 MPa and 13 %, respectively, Fig. 8b show the result of EN8 steel austenised at 914 °C, held for 10 min, cooled in furnace to 680 °C. The YS, UTS and the corresponding elongation were 397.86 MPa, 492.59 MPa and 16.9 %, respectively.

The reduction of the strength values observed in specimen B can be attributed, among other factors, to coarse grains. The reduction in the yield strength and tensile strength values in specimen B was worsened by small volume fraction of the harder phase (martensite). Specimen B exhibits relatively lower strength and higher ductility as compared to specimen A. This can be attributed to relatively large grains of specimen as there is very little difference in the volume fraction of the fourth phase between specimens A and B.

Figure 9a shows the stress versus strain curve of En8 austenised at 914 °C, held for 10 min, cooled in furnace. The YS, UTS and the corresponding elongation of

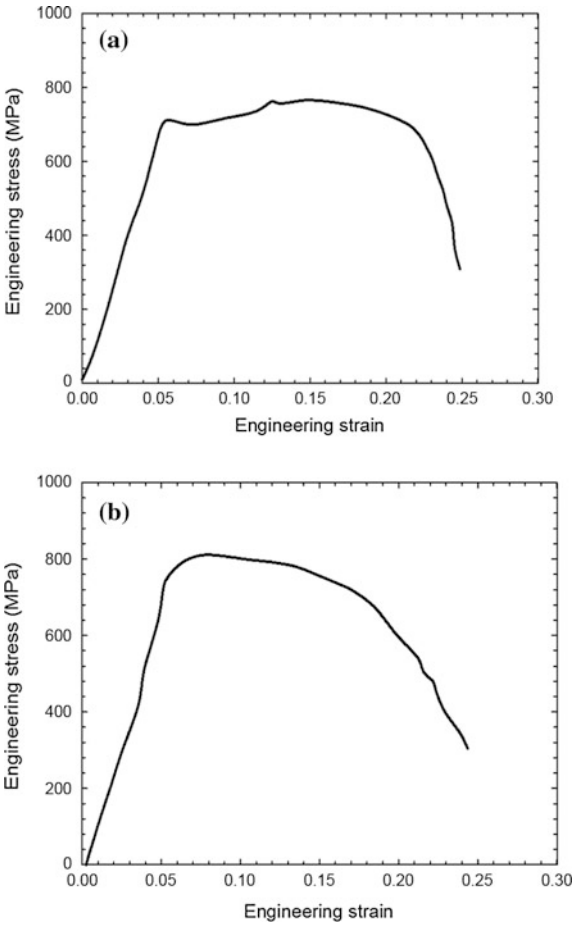


Fig. 9 **a** Stress versus strain curve of EN8 austenised at 914 °C, held for 10 min, cooled in furnace. **b** Stress versus strain curve of EN8 austenised at 914 °C held for 3 min cooled in furnace to 715 °C and quenched in oil

materials were 762.26 MPa, 762.57 MPa and 11.8 %, respectively. Figure 9b shows the stress versus strain curve of EN8 austenised at 914 °C held for 3 min cooled in furnace to 715 °C and quenched in oil. The YS, UTS and the corresponding elongation were 789.41 MPa, 781.41 MPa and 9.4 %, respectively. The yield strength of the specimen B was measured to be 391 MPa, which is a significant decrease from that of specimen A (475 MPa). As the holding time was reduced, the yield strength increased significantly whereas elongation at failure decreased. Further decrease in the holding time at critical temperature temperatures resulted in increase in the yield strength, values and reduction in elongation at failure (Ef) values. The yield strength of specimen D (781 MPa) is found to be equivalent to over 8 % enhancement compared to that of the specimen C. This is in agreement to the findings reported in [11]. They also reported that a number of factors such as heat treatment, alloying content and impurities affect the relationship between grain size and mechanical properties of steels. It is therefore imperative to control the heat

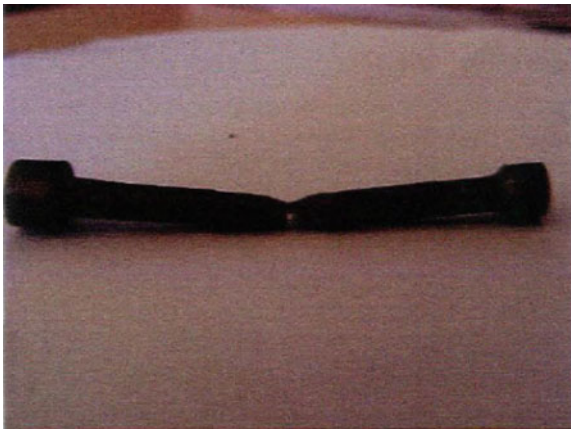


Fig. 10 Fracture specimen showing a cone and cup fracture mode

treatment process and keep impurities in the steel to minimum acceptable levels. In this study, the heat treatment regime was designed in such a way that formation of martensite and bainite phases was limited to a bare minimum level. Figure 10 shows the fracture specimen showing a cone and cup fracture mode.

The results of hardness test carried out on the steel samples were shown in Table 4 and Fig. 11. It is observed that increase in austenising temperature soaking time and slow cooling rate in heat treatment B induced a marked decrease in hardness in specimen B. This reduction in hardness may be attributed to two microstructural aspects: grain coarsening and reduction in fourth phase observed in specimen B. The maximum hardness values were obtained in heat treatments C and D. This confirmed the suspicion of the presence of harder phases (martensite and bainite) in specimens C and D as alluded to above. However, from the analysis of the phase volume fractions, grain size and hardness values of specimens A, B, C and D; it is observed that grain size has a major influence on variation in the hardness values observed.

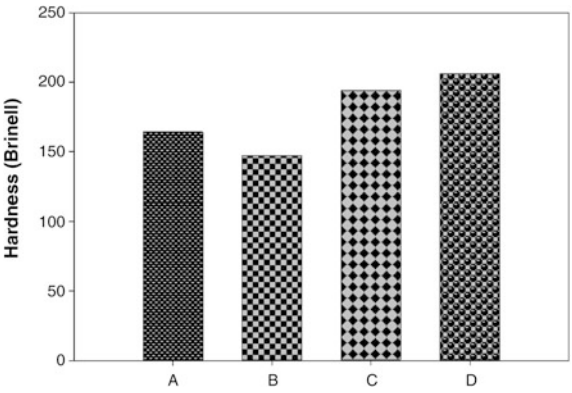


Fig. 11 Brinell hardness values of EN8 studied

3.1 Effects of Phase Volume Fraction on Yield Strength

It has been observed that the yield stress and work hardening of steel depend on individual phases present in the material. Considering the variation of yield strength with grain size and phase volume fractions [22]. Based on the experimental results, the effects of microstructural features such as grain size and phase volume fraction and mechanical properties (Young's modulus, the yield strength, the ultimate tensile strength, and elongation at failure) on the deformation behaviour of the steel studied can be predicted. Since any structural parameter that directly controls the yield strength, the ultimate tensile strength and elongation at failure, should have some influence (direct or indirect) on the stress flow and formability of the material.

4 Conclusions

1. In the study, the plastic deformation behavior and experimental methods used to quantify grain size and mechanical properties were described and following the analysis of the results obtained, the following conclusions were drawn;
2. Grain size and its distribution have a significant impact on the yielding characteristics of steel. It was observed that when a structure is composed of multiphase with different grain size distribution, the maximum plastic deformation will occur within the "weaker" phase. While the "stronger" phase will experience higher von Mises stress.
3. The annealing temperature and holding time have great influence on the grain size and phase distribution. The increase in annealing temperature and soaking time promotes the coarsening of grains and even an increase in formation of second phases (bainite and martensite) upon cooling.
4. Since the yield strength, increased linearly with decreasing grain size in conventional materials and, elongation at failure decreasing with decreasing grain size, coarse grains obtained in heat treatment B would be beneficial for manufacturing processed that depend on material flow (for example cold forming). Heat treatment C and D would be ideal for applications where high strength and high hardness values are required.

Acknowledgements The authors gratefully acknowledge the financial supports of Faculty of Engineering, Department of Mechanical Engineering, Cape Peninsula University of Technology, Cape Town and Faculty of Engineering and the Built Environment, Department of Chemical Engineering University of Johannesburg, South Africa.

References

1. Morris Jr J (2001) The influence of grain size on the mechanical properties of steel. eScholarship Repository, University of Califonia, pp 1–8
2. Moleejane CM, Sanusi KO, Ayodele OL, Oliver GJ (2014) Microstructural features and Mechanical behaviour of Unalloyed Medium Carbon Steel (EN8 Steel) after subsequent heat treatment. In: Proceedings of the world congress on engineering and computer science 2014, WCECS 2014, San Francisco, pp 1034–1039, 22–24 Oct 2014
3. Hall E (1951) The deformation and ageing of mild steel: III discussion of results. In: Proceedings of the physical society of London, pp 747–753
4. Petch N (1953) The cleavage strength of Polycrystals. Journal of the Iron and Steel Institute, pp 25–28
5. Armstrong R (1970) The influence of polycrystal grain size on several mechanical properties of materials. Metall Mater Trans B 1(5):1169–1176
6. Agrawal B (1988) Introduction to engineering materials. Tata McGraw-Hill, New Delhi
7. Zhao M, Qiu T, Nagai K, Yang K (2006) Grain growth and Hall-Petch relation in dual-sized ferrite/cementite steel with nano-sized cementite particles in a heterogenous and dense distribution. Scr Mater 54:1193–1197
8. Ueji R, Tsuji N, Minamino Y, Koizumi Y (2002) Ultra grain refinement of plain low carbon steel by cold rolling and annealing of martensite. Acta Mater 50:4177–4189
9. Chen S, Gan D (1986) Effects of grain boundary carbides on the tensile and impact properties of type 316 stainless steel. Mater Sci Eng 84:65–76
10. Muszka K, Majta J, Bienias L (2006) Effects of grain refinement on mechanical properties of microalloyed steel. Metall Foundary Eng 32(2):87–97
11. Liu K, Shan Y, Yang Z, Liang J, Lu L, Yang K (2006) Effect of heat treatment on prior grain size and mechanical property of a maraging stainless steel. J Mater Sci Technol 22(6):769–774
12. Liu Y, Wanga P, Lia J, Lua C, Quekb K, Liuc G (2003) Parametric study of a sprorocket system during heat-treatment process. Finite Elem Anal Des 40:25–40
13. Sen I, Tamirisakandala S, Miracle B, Ramamurty U (2007) Microstructural effects on the mechanical behaviour of B-modified H-6Al-4V alloys. Acta Materialia 55:4983–4993
14. Rack H (1978) Age hardening-grain size relationships in 18Ni maraging steels. Mater Sci Eng 34:263–270
15. Tamehiro H, Nakasugi H (1985) Austenite grain size of Tltaniun-microalloyed, continuously cast steel slabs. T Iron Steel I Jpn 25(4):311–317
16. Brunig N (2003) An anisotropic ductile damage model based on irreversible thermodynamics. Int J Plast 19:1679–1713
17. Barlat F, Aretz H, Yoon J, Karabin ME, Brem J, Dick R (2005) Linear transformation-based anisotropic yield functions. Int J Plast 21:1009–1039
18. Bonora N (1997) A nonlinear CDM model for ductile failure. Eng Fract Mech 58(1/2):11–28
19. Lassance D, Scheyvaerts F, Pardoen T (2006) Growth and coalescence of penny-shaped voids in metallic alloys. Eng Fract Mech 73:1009–1034
20. Voyiadjis G, Abed F (2006) A coupled temperature and strain rate dependent yield function for dynamic deformations of bee metals. Int J Plast 22:1398–1431
21. Mediavijla J, Peerlings R, Geers M (2006) Prediction of phase transformation during laser surface hardening of AISI 4140 including the effects of inhomogeneous austenite formation. Mater Sci Eng A 435–436:547–555
22. Mahdi L, Zhang L (2000) A Numerical Algorithm for the full coupling of mechanical deformation and phase transformation in surface grinding. Comput Mech 26:148–156

An Overview on Friction Stir Spot Welding of Dissimilar Materials

Mukuna P. Mubiayi and Esther T. Akinlabi

Abstract Understanding the fundamental process mechanisms of any manufacturing process is vital for its long-term development. Friction Stir Welding (FSW) process was invented and experimentally proven by The Welding Institute (TWI) in 1991 for joining Aluminium alloys. Friction Stir Spot Welding (FSSW) is a variant of the FSW which is found to be environmental friendly and an efficient process. FSSW technique has been gaining ground when compared to resistance spot welding (RSW) and could be used in various industries including, automobiles, ship building, aerospace, electrical and construction. FSSW has been successfully used to join several materials used in the above mentioned industries. In this review, FSSW studies are briefly summarised in terms of the evolving microstructure and mechanical properties between aluminium alloys and other materials such as copper, steel and magnesium.

Keywords Aluminium · Copper · Friction stir spot welding · Magnesium · Microstructure · Steel

1 Introduction

The current work is an extension of our previous work [1]. Friction Stir Spot Welding (FSSW) is a variant of Friction Stir Welding (FSW) process for spot welding applications. A non-consumable rotating tool is plunged into the workpieces to be joined. Upon reaching the selected plunge depth, the rotating tool is

M.P. Mubiayi (✉) · E.T. Akinlabi
Department of Mechanical Engineering Science, University of Johannesburg, Kingsway Campus, Corner Kingsway and University Road, Auckland Park, P.O. Box 524, Johannesburg, South Africa
e-mail: patrickmubiayi@gmail.com

E.T. Akinlabi
e-mail: etakinlabi@uj.ac.za

H.K. Kim et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-94-017-7236-5_38

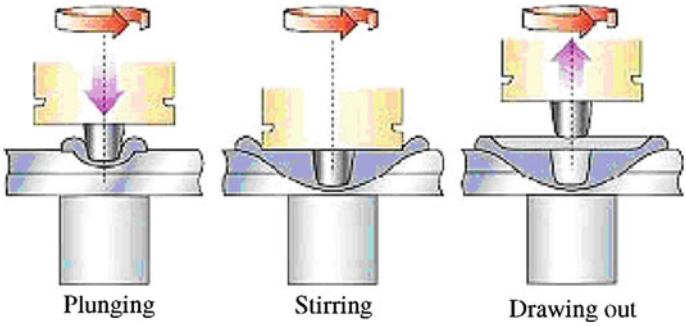


Fig. 1 Schematic illustration of friction stir spot welding process [3]

held in that position for a pre-determined time sometimes referred to as dwell period. Subsequently, the rotating tool is retracted from the welded joint leaving behind a friction stir spot weld. During FSSW, tool penetration and the dwell period basically determine the heat generation, material plasticisation around the pin, weld geometry and therefore the mechanical properties of the welded joint [2]. A schematic illustration of the FSSW process is shown in Fig. 1.

FSSW process uses a tool, similar to the FSW tool [4]. The shoulder generates bulk of the frictional or deformational heat whereas; the pin assists in material flow between the work pieces [2]. Besides the tool, the other parameters involved in FSSW are, the tool rotation speed; tool plunge depth and the dwell period. These parameters determine the strength and the surface finish of the welded joints [2].

A nomenclature is required to accurately describe the different microstructural regions present after FSSW. The cross section of the spot weld shows the five characteristics including the Parent Material (**PM**), the Heat Affected Zone (**HAZ**), Thermomechanically Affected Zone (**TMAZ**), The Stir Zone (**SZ**) and the Hook as shown in Fig. 2.

The Parent Material (**PM**) is the material that is remote from the welded region that has not been deformed; however it may have experienced thermal cycling from the weld. This is not affected by the heat in terms of the microstructure or the mechanical properties.

The Heat Affected Zone (**HAZ**) is the region which lies closer to the weld-center and has experienced a thermal cycle during welding which has modified the microstructure and/or the mechanical property, there is no plastic deformation in

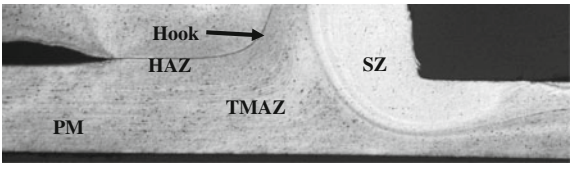


Fig. 2 Cross-sectional appearance of a typical friction stir spot weld [2]

this region. Whereas, the Thermomechanically Affected Zone (**TMAZ**) is found in the region where the tool has plastically deformed the material. In some materials, it is possible to obtain significant plastic strain without recrystallization in this region. There is a distinct boundary between the recrystallized zone and the TMAZ.

The Stir Zone (**SZ**) is the fully recrystallized region that is, in the immediate vicinity of the tool pin. The grains within the stir zone are roughly equiaxed and often an order of magnitude smaller than the grains in the parent material. Whereas, the Hook is a characteristic feature of Friction Stir Spot Welds in lap configuration where there is a formation of a geometrical defect originating at the interface of the two welded sheets [5].

There are many published reviews on Friction Stir Welding and processing [6–12], but so far there is no detailed review on Friction Stir Spot of similar and dissimilar materials. This review paper is focused on showing the current status of FSSW between similar and dissimilar materials and suggestions to fill the gaps to expand FSSW industrially.

2 Current status of Friction Stir Spot Welding (FSSW) of similar and dissimilar materials

2.1 FSSW between Aluminium Alloys

A number of studies have been conducted on Friction Stir Spot Welding between Aluminium alloys over the years. Uematsu et al. [13], joined T4 treated 6061 using a double-acting tool consisting of outer flat shoulder and inner retractable probe, which could re-fill probe hole. The microstructures of the weld zone were classified into MZ (mixed zone) and SZ, where fine equiaxed grains were observed due to dynamic recrystallisation during FSSW process. They further found that the tensile strength of the joint was improved by a re-filling process because the effective cross sectional area of the nugget was increased [12].

Merzoug et al. [3], conducted experiments on AA6060-T5 using a tool steel of the type X210 CR 12 and the rotational speed of the tool ranged from 1000 to 2000 rpm. The tensile tests made it possible to establish that the sample produced at 1000 rpm and 16 mm/min has a good quality of welding, which has 5 kN to 16 mm/min and 1000 rpm compared to 1.98 kN for 25 mm/min and 2000 rpm. The microhardness approached the maximum value as they moved away from the nugget zone.

Zhang et al. [14], spot welded AA 5052-H112 of 1 mm thickness. They concluded that softening occurs in the welds. A minimum hardness of 19.2 HV, which equals to 45.7 % of that of the PM, was measured in the HAZ. In addition, hardness in the TMAZ and SZ improved due to the recrystallisation which makes the hardness distribution exhibit a W-shaped appearance 13. The joints strength

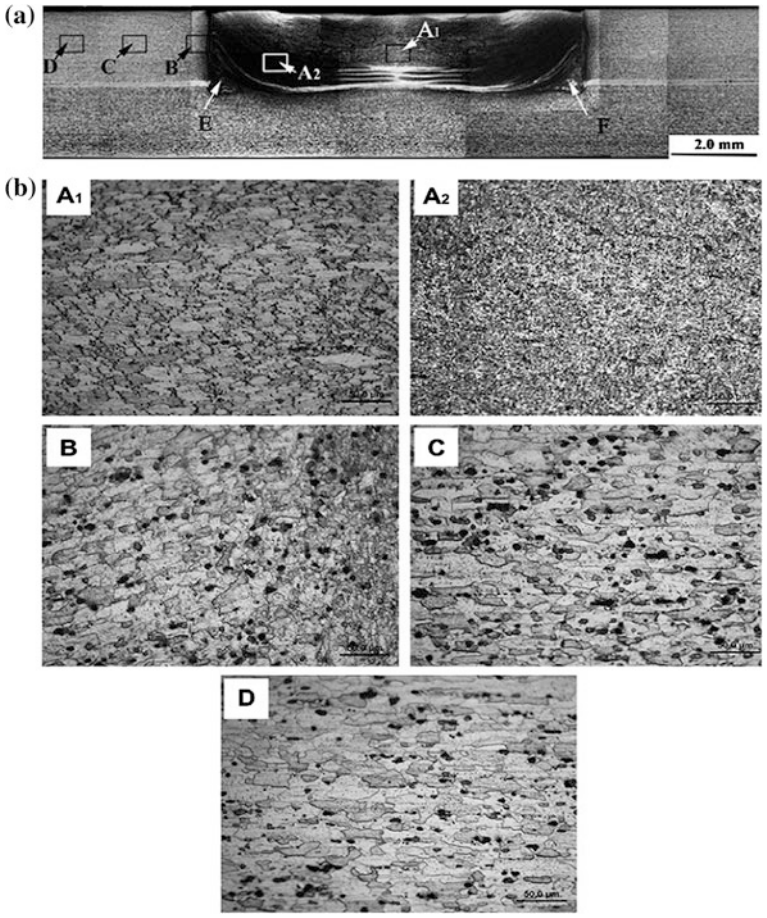


Fig. 3 Microstructures on longitudinal section of RFSSW joint made at welding condition of the rotational speed of 1200 rpm and dwell time of 4 s: **a** cross section of weld zone, **b** magnified views of the regions A1–D marked in (**a**), respectively [15]

decreases with increasing tool rotational speed, while it is almost independent of the given tool dwell times [14].

Shen et al. [15] used AA 7075-T6 plates of 2 mm thickness, the rotational speeds and the dwell time were varied, which were 1500, 1750 and 2000 rpm, and 3, 4 and 5 s, respectively. They investigated the microstructure and the mechanical properties of the refilled FSSW of AA7075-T6. The keyhole of the weld was refilled successfully, the microstructure of the weld exhibits variations in the grain sizes in the width and the thickness directions as depicted in Fig. 3 [15].

Additionally, they observed, defects associated to the material flow, such as hook, voids, bonding ligament and incomplete refill [15].The hardness profile of the weld exhibited a W-shaped appearance in the macroscopic level. They attributed

the change in the hardness to the comprehensive effects of several factors, in which the precipitation state plays a decisive role.

Shen et al. [16], joined 6061-T4 aluminium alloy sheets with 2 mm thickness using high-speed steel tool (JIS,SKD61),whose shoulder diameter is 10 mm with a concave profile. A preferable appearance of the joint was obtained at higher rotational speed and longer duration time. The microstructures of the weld was divided into four regions, BM, HAZ, TMAZ and SZ, there exist dynamic recrystallisation and dissolution of precipitates in the weld. The hook geometries vary significantly depending on rotational speed and dwell time. The formation of hook was attributed to the insufficient pressure vertical to the tool, and the amount of material extruded upward and the effective weld width increases with the increase of the rotational speed and dwell time [16]. Furthermore, the Vickers hardness profile of the sheets showed a W-shaped or an upside down V-shaped appearance. The minimum hardness reaches 46.7 HV in the periphery of the HAZ and TMAZ and different variation of Vickers hardness in each region of the weld was attributed to the comprehensive effects of the strain-hardening, the dissolution of strengthening phase and the variation in the grain sizes. The tensile/shear strength increases with the increasing rotational speed at a given duration time. Though, under a given rotational speed, differences in tensile/shear strength among three dwell times are rather small. The tool rotational speed plays a determinant role in determining the tensile/shear strength [16].

Tozaki et al. [17], joined AA6061-T4 sheets with 2 mm thickness using different probes lengths of 3.7, 3.1 and 2.4 mm with a shoulder diameter of 10 mm. The probes were made of high-speed steel (Japanese Industrial Standard (JIS), SKD61) and had a standard metric M3.5 left-hand thread. A constant tool plunge rate of 20 mm/min and a shoulder plunge depth of 0.2 mm below the upper plate surface were applied. Furthermore, the tool rotational speeds and the tool holding times were also varied, which were 2000, 2500 and 3000 rpm, and 0.2, 1 and 3 s, respectively. They observed that, the microstructures of the welds varied significantly depending on the probe length, tool rotational speed and tool holding time and the tensile shear strength increased with increasing probe length [17].

Badarinarayan et al. [5] joined annealed AA 5083 sheets with two different thicknesses of 1.64 and 1.24 mm. The tool shoulder diameter was 12 mm with a concave profile, and the pin length was 1.6 mm. The two different pin geometries are conventional cylindrical and triangular pin. They concluded that the tool pin geometry significantly affects the hook. In the FSSW-C (cylindrical pin) weld, the hook runs gradually upward and then bypasses the stir zone and points downward towards the weld bottom. Whereas, in the FSSW-T (triangular pin) weld, the hook is directed upward towards the stir zone and ends with a very short plateau.

Wang and Lee [18], spot welded AA6061-T6 with a thickness of 1 mm. They found in their experimental results that under lap-shear loading conditions, the failure is initiated near the SZ in the middle part of the nugget and the failure propagates along the circumference of the nugget to final fracture. The location of the initial necking/shear failure is near the possible original notch tip and the failures of the Friction Stir Spot Welds were fractured through the TMAZ near the

weld nuggets 18. Furthermore, the hardness initially decreases upon approaching the boundary between the base metal and the HAZ, then drops sharply to a minimum in the TMAZ. After passing the TMAZ, the hardness gradually increases up to the SZ hardness [18].

Buffa et al. [19], used AA6082-T6 aluminium alloy, 1.5 mm in thickness. They used an H13 tool steel quenched at 1020 °C, characterized by 52 HRC hardness. The shoulder was 15 mm in diameter and a 40° conical pin was adopted, with a major diameter of 7 mm and a minor diameter of 2.2 mm; the pin height was 2.6 mm. They used a variation of FSSW process and successfully produced the welds.

Wang et al. [20] joined commercially pure AA1050-H18 sheets with a 300 μm thickness. The experiments results suggest that under lap-shear loading conditions, the failure is initiated near the SZ in the middle part of the nugget and the failure propagates along the circumference of the nugget to final fracture. The location of the initial necking/shear failure is near the possible original notch tip and the failures of the Friction Stir Spot microwelds were fractured through the TMAZ near the weld nuggets.

Yuan et al. [21], spot welded 1 mm thick AA6016-T4 sheets using two tools. A CP tool which is a conventional tool with a center pin, has a concave shoulder with 10 mm diameter and a 1.5 mm long step spiral pin with root diameter of 4.5 mm and tip diameter of 3 mm. The OC tool is the off-center feature tool with the same concave shoulder shape and diameter, and three off-center 0.8 mm long hemispherical pin features. Both tools were machined from Densimet tungsten alloy. Results indicated that the tool rotation speed and the plunge depth profoundly influenced the lap-shear separation loads [21].

Furthermore, both tools exhibited maximum weld separation load, about 3.3 kN at 0.2 mm shoulder penetration depth; different tool rotation speeds, 1500 rpm for the CP tool and 2500 rpm for the OC tool [21].

Jeon et al. [22], used Friction Stir Spot Welding process to join, 3 mm thick 5052-H32 and 6061-T6 aluminium alloy sheets. The z-force and torque histories as a function of the tool displacement vary significantly during the FSSW process. The force and torque histories during the FSSW process can be distinguished by different stages based on the contact phenomena between the tool and joined sheets. The shapes of the z-force histories are somewhat different for the selected material combinations, while the torque histories have quite similar shapes. The differences in the z-force histories for the different material combinations may be explained based on the different mechanical behaviours of the aluminium alloys at various elevated temperatures [22].

Thoppul and Gibson [23], used AA6111-T4 to produced spot welds. From the microstructural studies, it is clear that increasing the processing time increases both the tool depth of penetration and the bonding area between the lap joints.

Su et al. [24], investigated the Friction Stir Spot Welding of 5754 and Al 6111 sheets using a tool having a smooth pin with or without a dwell period and spot welds were made using a threaded tool without the application of a dwell period. They did not observe dissimilar intermixing in the spot welds made using a tool

with a smooth pin with or without the application of a dwell period. They further proposed that dissimilar intermixing during the dwell period in spot welding results from the incorporation of upper (Al 5754) and lower (Al 6111) sheet materials at the top of the thread on the rotating pin [24].

Babu et al. [25], welded 3 mm thick AA2014-T4 and T6 conditions with and without alclad layers to investigate the effects of tool geometry and welding process parameters on joint formation. A good correlation between process parameters, bond width, hook height, joint strength, and fracture mode was observed. They further found that the presence of the Alclad layers and the base metal temper condition have no major effect on the joint formation and joint strength [25].

Pathak et al. [26], joined AA5754 sheets using tools with circular and tapered pin considering different tool rotational speeds, plunge depths, and dwell times. Symmetric temperature profiles have been observed near the sheet-tool interface during spot welding using tools with circular and tapered pin at different rotational speeds. The peak temperature increases with increase in tool rotational speed and dwell time. Tool geometry also affects the temperature distribution, as under similar condition, tool with circular pin generated more heat than tool with tapered pin. The lap shear test with welded samples shows influence of tool rotational speed, plunge depth, and dwell time. The common observation for both the tools is that lap shear load increases with the increase in the said parameters [26].

2.2 FSSW between Aluminium and Magnesium

FSSW process has been successfully used to Friction Stir Spot Weld aluminium to magnesium used especially in the automotive and the aerospace industry.

Suhuddin et al. [27] successfully joined Al alloy AA5754 to Mg alloy AZ31. Their microstructure analyses showed that the grain structure development in the stir zone was affected by grain boundary diffusion, interfacial diffusion and dynamic recrystallisation, which resulted in fine equiaxed grains of $Al_{12}Mg_{17}$ in the weld center. Whereas the hardness profile of the Mg/Mg similar weld exhibited a W-shaped appearance, the lower hardness values appeared in the TMAZ and HAZ of both Mg/Mg and Al/Al similar welds. In the Al/Mg dissimilar weld, a characteristic interfacial layer consisting of intermetallic compounds (IMC) $Al_{12}Mg_{17}$ and Al_3Mg_2 was observed. Both the Mg/Mg and Al/Al similar welds had significantly higher lap shear strength, failure energy and fatigue life than the Al/Mg dissimilar weld. While the Al/Al weld displayed a slightly lower lap shear strength than the Mg/Mg weld, the Al/Al weld had higher failure energy and fatigue life [27].

Chowdhury et al. [28], used FSSW process to spot weld commercial AZ31B-H24 Mg and AA5754 with a thickness of 2 mm. They used a tool made from H13 tool steel which had a diameter of 13 mm for the scrolled shoulder and 5 mm for the left-hand threaded pin. A pin length of 2.8 mm, tool rotational rate of 2000 rpm, tool plunge rate of 3 mm/s, tool removal rate of 15 mm/s, shoulder plunge depth of 0.2 mm and dwell time of 2 s was used. There was a presence of

intermetallic compounds ($Al_{12}Mg_{17}$ and Al_3Mg_2). The microhardness profile of the Mg/Mg weld exhibited a W-shaped appearance, where the hardness gradually increased towards the keyhole direction [28].

Chowdhury et al. [29] conducted a study on FSSW of Commercial AZ31B-H24 Mg and AA5754-O Al alloy sheets with a thickness of 2 mm were selected for FSSW. They observed a distinctive interfacial layer consisting of $Al_{12}Mg_{17}$ and Al_3Mg_2 intermetallic compounds in the Friction Stir Spot Welded dissimilar Al/Mg and Mg/Al adhesive joints. Furthermore, they stated that in comparison with the Al/Mg weld without adhesive, the extent of forming the intermetallic compounds decreased in the dissimilar adhesive joints. They also observed a much higher hardness with values in between HV90 and 125 in the stir zone of Al/Mg and Mg/Al adhesive welds due to the presence of intermetallic compound layer [29]. It was also observed that both Mg/Al and Al/Mg adhesive welds had significantly higher lap shear strength and failure energy than the Al/Mg dissimilar weld without adhesive [29].

Choi et al. [30], Friction Stir Spot joined 6K21 Al alloy and AZ31 Mg alloy with a tool made of general tool steel (SKD11) and composed of a shank, a shoulder, and a pin. The shoulder diameter, pin diameter, pin height and weld tilt angle of the tool were 13.5, 9.5, 0.5 and 0 mm, respectively. The obtained results demonstrated the formation of IMCs in the interface between the Al and Mg alloy joints. These IMCs were revealed as Al_3Mg_2, formed on the Al substrate, and $Al_{12}Mg_{17}$, formed on the Mg substrate. In addition, the thickness of the intermetallic compounds layer increases with increasing tool rotation speed and duration time, and has a significant effect on the strengths of the joints. Heavy thicknesses of intermetallic compounds layer seriously deteriorates the mechanical properties of the joints. The maximum tensile shear fracture load of the Al and Mg alloy joint was about 1.6 kN, however, the load value decreased with increasing of tool rotation speed and duration time, owing to the cracks in the IMCs [30].

2.3 FSSW between Aluminium and Steel

Chen et al. [31], welded 1 mm thick 6111-T4 Al and DC04 low carbon steel sheet. The tool had an 11 mm diameter steel shoulder, with a scroll profile to improve the flow of material, and a tapered 3 mm diameter WC 1 mm long probe. The radius of the probes orbital path was 2.5 mm which produced a swept area of 8 mm diameter on the steel surface. They produced high quality friction spot welds between thin Al and steel automotive sheet within a weld time of one second which is a desired target time by industries.

Sun et al. [32], used a concave-shaped shoulder geometry tool with a diameter of 12 mm and a probe with a diameter of 4 mm to FSSW a 1 mm thick commercial 6061 Al alloy and a mild steel. They observed no obvious intermetallic compound layer along the Al/Fe interface after producing the welds. Furthermore, they observed that the shear tensile failure load can reach a maximum value of 3607 N.

The pin length has little effect on the weld properties, which indicates that the tool life can be significantly extended by this new spot welding technique [32].

Bozzi et al. [33], joined AA 6016 (1, 2 mm thick) to a galvanized IF-steel sheet (2.0 mm thick) using a tool machined into tungsten rhenium alloy (W25Re). The intermetallic compounds layer thicknesses increases with the rotational speed and the penetration depth. They also noticed that the intermetallic compounds seems to be necessary to improve the weld strength, but if the IMC layer is too thick cracks initiate and propagate easily through the brittle IMC tangles [33].

Figner et al. [34], Friction Stir Spot Welded HX340 LAD sheets of steel of 1 mm thickness and aluminium AA5754-H111 of 2 mm thickness. They observed by using proper selection of spindle speed and dwell time, the strength of the spot weld can be improved significantly. Thus, a maximum load in the shear tension test of 8.4 kN per spot can be achieved while by increasing the dwell time, the amount of intermetallic phases (IMP) increases and breaks off, causing a drop in the strength [34]. More research needs to be conducted to optimise the process in order to use FSSW between Aluminium and Steel industrially.

2.4 FSSW between Aluminium Alloys and Copper

Efforts have been made to produce Friction Stir Spot Welds between Aluminium and copper. This section summarise studies conducted and published. Özdemir et al. [35], Heideman et al. [36] and Shiraly et al. [37] have successfully Friction Stir Spot Welded a 3 mm thick AA1050 to pure copper, 1.5 mm thick AA 6061-T6 to oxygen free pure copper and 500 µm thick 1050 Aluminium foils to 100 µm thick Copper foils respectively. It was noticed while conducting an investigation on the existing literature on FSSW between Aluminium and Copper only few results are available, therefore, it is of importance that more research has to be conducted to optimised the process to enable it to be used as an alternative to riveting and Resistance Spot Welding.

Özdemir et al. [35] produced Friction Stir Spot welds using three different plunge depths namely 2.8, 4 and 5 mm, using a tool with a shoulder diameter of 20 mm and a pin with a diameter of 5 mm. Furthermore, the spot welds were produced using 1600 rpm rotation speed with 10 s hold time [35]. They produced spot welds with no macroscopic defects and the grains on the copper side close to the Al/Cu interface were finer than those of copper base metal. The difference in the grain sizes was attributed to the effect of the rotating pin which deformed the grains close to the Al/Cu interface and the recrystallization of grains in the stir zone of the copper metal due to heat input [35]. Furthermore, the EDS analyses conducted revealed the formation of hard and brittle intermetallic compounds AlCu, Al_2Cu and Al_4Cu_9 formed at the interface [35]. The tensile shear test results showed that 2.8 mm plunge depth produced poor results whereas 4 mm plunge depth showed the highest values of shear tensile test compare to the 5 mm, it was suspected to be due to the penetration of Cu into Al in a more diffused way [35]. Özdemir et al.

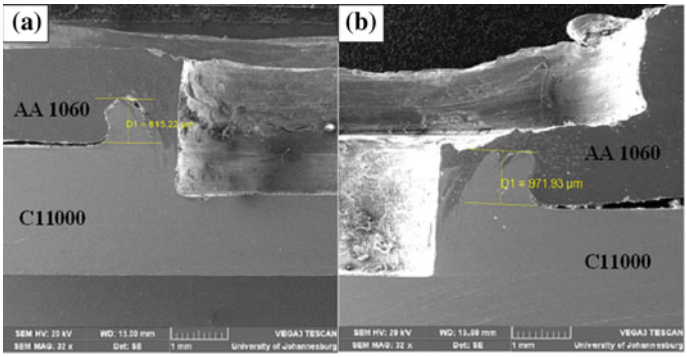


Fig. 4 Copper ring size of weld produced using flat pin (4 mm length) and flat shoulder (15 mm diameter) at welding conditions of 800 rpm and 0.5 mm (**a**) and 1.0 mm (**b**) shoulder plunge depth

[35], also indicated that the hardness increases at the bottom region of the pin hole (in the Cu material) due to heat input introduced by the rotating pin. Furthermore, they stated that as the plunge depth increases, the grain size decreases, which caused higher hardness at the Cu side for the 5 mm plunge depth and due to more diffuse and selective penetration of Cu into Al for 5 mm plunge depth, higher hardness values were obtained on the Al side [35].

On the other hand, Heideman et al. [36] conducted metallurgical analysis on AA 6061-T6 to oxygen free Cu using Friction Stir Spot Welding process. The tool used was a threaded pin design using a prehardned H13 tool steel with a shoulder of 10 mm, pin diameter of 4 mm and the thread pitch of 0.7 mm. Two different plunge depths were used: 0 and 0.13 mm and two different weld times of 3 and 6 s [36]. They used rotation speeds varying from 1000 to 2000 rpm. Furthermore, they indicated that, the plunge depth, rotation speed and tool length were the primary factors affecting the strength of the welds. The presence of an intermetallic interface was not observed in the strong welds, they were only in the form of small particles that do not connect along the bond line to become most detrimental to the weld quality [36]. Heideman et al. [36] also showed a vertical cross section of the spot weld with a Cu ring appearing on each side of the keyhole. Cu rings images with their measurements using different process parameters are exhibited in Fig. 4.

Most recently, Shiraly et al. [37] performed FSSW of Al/Cu composite produced by accumulative roll-bonding process using a triangular pin with no features.

They found that the weld made at lower tool rotation rate was not bonded; this was due to no intermixing between the upper and lower sheets. Furthermore, the maximum shear failure load increased with the increasing tool rotation rate, which can be attributed to the increasing area and effective length of Stir Zone (SZ). The experimental interpretations showed the presence of the intermetallic compounds (Al_2Cu and $AlCu_3$) in the Stir Zone (SZ). The presence of the intermetallic compounds and the material crushing increased the hardness in the Stir Zone.

3 Conclusion

A literature review has been conducted on the FSSW process of dissimilar materials. It shows that significant research and development of the FSSW process has been achieved worldwide and this process has established itself as a viable joining option for the automotive and aerospace industries. However, more researches need to be further conducted to fully understand and optimise the process. It was also noticed that not much importance has been shown on producing FSS welds between aluminium and copper which could be an alternative solution to riveting and Resistance Spot Welding (RSW) since spot welding between aluminium and copper could be useful in making electrical connections and components. Although, the ability of FSSW to join lightweight, high strength aluminium alloys to other materials such as magnesium, copper and steel is desirable, extending this process into high melting temperature materials has proven challenging due to tool cost and tool wear rates. It is expected that if the process is applied efficiently, it can be a technical and economical process compared to the traditional welding processes.

Acknowledgments The financial support of the University of Johannesburg and the Tertiary Education Support Programme fund of ESKOM are acknowledged.

References

1. Mubiayi MP, Akinlabi ET (2014) Friction stir spot welding of dissimilar materials: an overview. Proceedings of the world congress on engineering and computer science 2014, WCECS 2014, 22–24 October 2014. Lecture notes in engineering and computer science. San Francisco, USA, pp 1089–1094
2. Badarinarayan H (2009) Fundamentals of friction stir spot welding. Ph.D. thesis, Missouri University of Science and Technology, USA
3. Merzoug M, Mazari M, Berrahal L, Imad A (2010) Parametric studies of the process of friction spot stir welding of aluminium 6060-T5 alloys. Mater Des 31:3023–3028
4. Timothy JM (2008) Friction stir welding of commercially available superplastic aluminium. Ph.D. thesis, Department of Engineering and Design, Brunel University, Brunel
5. Badarinarayan H, Yang Q, Zhu S (2009) Effect of tool geometry on static strength of friction stir spot-welded aluminum alloy. Int J Mach Tools Manuf 49(2):142–148
6. Ma ZY (2008) Friction stir processing technology: a review. Metall Mater Trans A 39A: 642–658
7. DebRoy T, Bhadeshia HKDH (2010) Friction stir welding of dissimilar alloys—a perspective. Sci Technol Weld Joining 15(4):266–270
8. Sivashanmugam M, Ravikumar S, Kumar T, Seshagiri Rao V, Muruganandam D (2010) A review on friction stir welding for aluminium alloys, 978-1-4244-9082-0/10/$26.00 ©2010 IEEE, 216–221
9. Rai R, De A, Bhadeshia HKDH, DebRoy T (2011) Review: friction stir welding tools. Sci Technol Weld Joining 16(4):325–342
10. Mubiayi MP, Akinlabi ET (2013) Friction stir welding of dissimilar materials: an overview. Paper presented at ICAMAME 2013: International Conference on Aerospace, Mechanical, Automotive and Materials Engineering (WASET), Johannesburg, South Africa, 29–30 Apr 2013

11. Mubiayi MP, Akinlabi ET (2013) Friction stir welding of dissimilar materials between aluminium alloys and copper—an overview. Lecture notes in engineering and computer science: proceedings of the world congress on engineering 2013, WCE2013, London, UK, pp. 1990–1996, 3–5 July 2013
12. Kumar A, Jadoun RS (2014) Friction stir welding of dissimilar materials/alloys: a review. Int J Mech Eng Rob Res 1(1):106–113, ISSN: 2278-0149, Special issue
13. Uematsu Y, Tokaji K, Tozaki Y, Kurita T, Murata S (2008) Effect of re-filling probe hole on tensile failure and fatigue behaviour of friction stir spot welded joints in Al–Mg–Si alloy. Int J Fatigue 30:1956–1966
14. Zhang Z, Yang X, Zhang J, Zhou G, Xiaodong X, Zou B (2011) Effect of welding parameters on microstructure and mechanical properties of friction stir spot welded 5052 aluminum alloy. Mater Des 32:4461–4470
15. Shen Z, Yang X, Zhang Z, Cui L, Li T (2013) Microstructure and failure mechanisms of refill friction stir spot welded 7075-T6 aluminum alloy joints. Mater Des 44:476–486
16. Shen Z, Yang X, Zhang Z, Cui L, Yin Y (2013) Mechanical properties and failure mechanisms of friction stir spot welds of AA 6061-T4 sheets. Mater Des 49:181–191
17. Tozaki Y, Uematsu Y, Tokaji K (2007) Effect of tool geometry on microstructure and static strength in friction stir spot welded aluminium alloys. Int J Mach Tools Manuf 47:2230–2236
18. Wang DA, Lee SC (2007) Microstructures and failure mechanisms of friction stir spot welds of aluminum 6061-T6 sheets. J Mater Process Technol 186:291–297
19. Buffa G, Fratini L, Piacentini M (2008) On the influence of tool path in friction stir spot welding of aluminum alloys. J Mater Process Technol 208:309–317
20. Wang DA, Chao CW, Lin PC, Uan JY (2010) Mechanical characterization of friction stir spot microwelds. J Mater Process Technol 210:1942–1948
21. Yuan W, Mishra RS, Webb S, Chen YL, Carlson B, Herling DR, Grant GJ (2011) Effect of tool design and process parameters on properties of Al alloy 6016 friction stir spot welds. J Mater Process Technol 211:972–977
22. Chi-Sung J, Sung-Tae H, Yong-Jai K, Hoon-Hwe C, Heung Nam H (2012) Material properties of friction stir spot welded joints of dissimilar aluminum alloys. Trans Nonferrous Met Soc China 22:p605–p613
23. Thoppul SD, Gibson RF (2009) Mechanical characterization of spot friction stir welded joints in aluminum alloys by combined experimental/numerical approaches part I: Micromechanical studies. Mater Charact 60:1342–1351
24. Su P, Gerlich A, North TH, Bendzsak GJ (2007) Intermixing in dissimilar friction stir spot welds. Metall Mater Trans A 38A:584
25. Babu S, Sankar VS, Janaki Ram GD, Venkitakrishnan PV, Madhusudhan Reddy G, Prasad Rao K (2013) Microstructures and mechanical properties of friction stir spot welded aluminum alloy AA2014. J Mater Eng Perform 22(1):71
26. Pathak N, Bandyopadhyay K, Sarangi M, Panda SK (2013) Microstructure and mechanical performance of friction stir spot-welded aluminum-5754 sheets. J Mater Eng Perform 22 (1):132
27. Suhuddin UFH, Fischer V, dos Santos JF (2013) The thermal cycle during the dissimilar friction spot welding of aluminum and magnesium alloy. Scripta Mater 68:87–90
28. Chowdhury SH, Chen DL, Bhole SD, Cao X, Wanjara P (2012) Lap shear strength and fatigue life of friction stir spot welded AZ31 magnesium and 5754 aluminum alloys. Mater Sci Eng A 556:500–509
29. Chowdhury SH, Chen DL, Bhole SD, Cao X, Wanjara P (2013) Lap shear strength and fatigue behaviour of friction stir spot welded dissimilar magnesium-to-aluminum joints with adhesive. Mater Sci Eng A 562:53–60
30. Choi H, Ahn BW, Lee CY, Yeon YM, Song K, Jung SB (2011) Formation of intermetallic compounds in Al and Mg alloy interface during friction stir spot welding. Intermetallics 19:125–130

31. Chen YC, Gholinia A, Prangnell PB (2012) Interface structure and bonding in abrasion circle friction stir spot welding: a novel approach for rapid welding aluminium alloy to steel automotive sheet. Mater Chem Phys 134:459–463
32. Sun YF, Fujii H, Takaki N, Okitsu Y (2013) Microstructure and mechanical properties of dissimilar Al alloy/steel joints prepared by a flat spot friction stir welding technique. Mater Des 47:350–357
33. Bozzi S, Helbert-Etter AL, Baudin T, Criqui B, Kerbiguet JG (2010) Intermetallic compounds in Al 6016/IF-steel friction stir spot welds. Mater Sci Eng A 527:4505–4509
34. Figner G, Vallant R, Weinberger T, Schrottner H, Pasic H, Enzinger N (2009) Friction stir spot welds between aluminium and steel automotive sheets: influence of welding parameters on mechanical properties and microstructure. Weld World 53(1/2):R13–R23
35. Özdemir U, Sayer S, Yeni C, Bornova-Izmir (2012) Effect of pin penetration depth on the mechanical properties of friction stir spot welded aluminum and copper. Mater Test IN Joining Technol 54(4):233–239
36. Heideman R, Johnson C, Kou S (2010) Metallurgical analysis of Al/Cu friction stir spot welding. Sci Technol Weld Joining 15(7):597–604
37. Shiraly M, Shamanian M, Toroghinejad MR, Jazani MA (2014) Effect of tool rotation rate on microstructure and mechanical behavior of friction stir spot-welded Al/Cu composite. J Mater Eng Perform 23(2):413–420

ANFIS Modeling for Higher Machining Performance of Aluminium Tempered Grade 6061 Using Novel SIO_2 Nanolubrication

Mohd Sayuti Ab Karim and Ahmed Aly Diaa Mohammed Sarhan

Abstract Aluminum Al6061-T6 is a common alloy which is used for many purposes due to its superior mechanical properties such as hardness and weldability. Implementation of CNC milling machine in processing Al6061-T6 would be a good process especially in producing varieties shape of products to adapt with different applications. However, the demand for high quality focuses attention on product quality, especially the roughness of the machined surface, because of its effect on product appearance, function, and reliability. Introducing correct lubrication in the machining zone could improve the product quality. Due to complexity and uncertainty of the machining processes, soft computing techniques are being preferred for predicting the performance of the machining processes and. In this chapter, a new application of ANFIS to predict the performance of machining AL-6061-T6 using SiO_2 nanolubricant is presented. The parameters of SiO_2 nanolubrication include SiO_2 concentration, nozzle angle and air carrier pressure are investigated for the lowest cutting force, cutting temperature and surface roughness with the 96.195, 98.27 and 91.37 % accuracy obtained between experimental and numerical measurement, respectively.

Keywords Al6061-T6 alloy · ANFIS modeling · Cutting force · Cutting temperature · End milling · SiO_2 nanolubrication · Surface roughness

M.S.A. Karim · A.A.D.M. Sarhan (✉)
Centre of Advanced Manufacturing and Material Processing,
Department of Engineering Design and Manufacturing, Engineering Faculty,
University of Malaya, 50603 Kuala Lumpur, Malaysia
e-mail: ah_sarhan@um.edu.my

M.S.A. Karim
e-mail: mdsayuti@um.edu.my

H.K. Kim et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-94-017-7236-5_39

1 Introduction

Aluminum has many benefits over other materials, including a high strength to weight ratio, corrosion resistance, formability, and price. Aluminum AL6061-T6 is an alloy which contains magnesium and silicon as major alloying elements. It has been a common alloy which is used for many purposes since it has the superior mechanical properties such as hardness and good weldability [1]. The capability of the CNC milling machine to make complicated special products would be a noteworthy advantage for Aluminum Al6061-T6. However, the demand for high quality focuses attention on the surface condition and the quality of the product, especially the roughness of the machined surface, because of its effects on product appearance, function and reliability [1–3].

The tribological characteristic of machining process can be improved by introducing lubrication in the machining zone [4, 5]. Correct application of lubricants has been proven to greatly reduce friction in the tool chip interface, this results in improving the surface quality. Although the significance of lubrication in machining is widely recognized, the usage of conventional flooding application in machining processes has become a huge liability. Not only does the Environmental Protection Agency (EPA) regulate the disposal of such mixtures, but many countries and localities also have classified them as hazardous wastes as they contain environmentally harmful or potentially damaging chemical constituents [6, 7]. Beside that economically, the cost related to the lubrication and cutting fluid is 17 % of total production cost which is normally higher than that of cutting tool equipments which incurs only 7.5 % of total cost. Consequently, eliminating the use of lubricants, if possible, can be a significant economic incentive [8–10].

At present, many efforts are being undertaken to develop advanced machining processes using less lubrications [11]. Promising alternatives to conventional flood coolant applications are the minimum quantity lubrication (known as MQL) [12]. Klocke and Eisennblatter [13] state that MQL is referred to the use of lubrication of only a minute amount-typically of a flow rate of 50–500 ml/h which is about three to four orders of magnitude lower than the amount commonly used in flood cooling condition. In addition, the dry chips can be recycled without incurring large cleaning expenses making the application of nanolubrication a plausible solution [8, 9, 14].

Nowadays, many nanolubricant has been identified by the advancement in modern technology which makes possible to sustain and provide lubricity over wide range of temperature [15–17]. Nanolubricant is a kind of new engineering material consisting of nanometer-sized particles dispersed in base oil. It would be an effective method to be used in reducing friction between two contact surfaces and depends on the working conditions. Lubricants are expected to withstand the high machining temperatures, non-toxic, easy to be applied and effective in term of cost [18]. The effectiveness of the lubrication depends on the morphology, crystal structure of solid lubricants, the way of particle introduced to the tool-workpiece interface and quantity [12, 19].

In addition, the productivity in the machining industry could increase through cost reduction by abandonment of the cutting fluid, saving the environment and at the same time improve the machining performance. Physical analysis of nanolubricant [20] showed the nanoparticle dispersed can easily penetrate into the rubbing surfaces and have large effect of elastohydrodynamic lubrication. Moreover, thermal conductivity of nanolubricant increases linearly with the concentration, which performs as hydrodynamic interaction to enhance thermal transport capability [21].

Many types of nanoparticle have been used as a lubricant by researchers in order to investigate its effects on the machining performance. It is well documented that silicon dioxide (SiO_2) nanoparticle is a hard and brittle material and cheap and available in market. This nanoparticle has very good mechanical properties especially in term of hardness (Vickers hardness—1000 kgf mm^{-2}) and in very small size range from 5 nm up to 100 nm.

In line with the previous research work as reviewed above, the investigation of optimum SiO_2 lubrication parameters in milling of Al6061-T6 is carried out to investigate the effective improvement of the machined surface quality. These parameters include nanolubricant concentration, nozzle angle and air carrier pressure. Due to complexity and uncertainty of the machining processes, of late, soft computing techniques are being preferred to physics-based models for predicting the performance of the machining processes and optimizing them [22].

Soft computing techniques are useful when exact mathematical information is not available and these differ from conventional computing in that it is tolerant of imprecision, uncertainty, partial truth, approximation, and met heuristics. ANFIS is one of the soft computing techniques that play a significant role in input-output matrix relationship modeling. It is used when subjective knowledge and suggestion by the expert are significant in defining objective function and decision variables. ANFIS is preferred to predicting machining performance based on the input variables due to nonlinear condition in machining process [22, 23].

Following the literature above, for predicting of the cutting force, cutting temperature and surface roughness, this study has been conducted using ANFIS modeling by anticipating nanolubrication concentration, air pressure and nozzle orientation as lubrication parameters.

2 Design of Experiment

The most important stage in the designing of experiment lies in the selection of lubrication parameters and identifying the experimental array. In this experiment with three parameters and four levels each, the factors design used is a $L_{12}(4^3)$ experimental array. This array is chosen due to its capability to check the interactions among parameters. The parameters and levels are assigned as in Table 1 for twelve experiments as shown in Table 2.

Table 1 The lubrication parameters in twelve experimental condition levels

Lubrication parameters		Level (i)			
		1	2	3	4
A	Nanoparticle concentration (wt%)	0	0.2	0.5	1.0
B	Air pressure (bar)	1	2	3	4
C	Nozzle orientation (°)	15	30	45	60

Table 2 The orthogonal array of $L_{12}(4)^3$

Exp.	Parameters combinations		
	A	B	C
1	1	1	1
2	1	2	2
3	1	3	3
4	1	4	4
5	2	1	2
6	2	2	1
7	2	3	4
8	2	4	3
9	3	1	3
10	3	2	4
11	3	3	1
12	3	4	2

3 Experimental Set up and Procedure

The second step is to run the experiments based on the selected experimental array. The twelve experiments were carried out in a random sequence to eliminate any other invisible factors, which might also contribute to the cutting force, cutting temperature and surface roughness. The experimental set-up is shown in Fig. 1. The machine used in this study is a vertical-type machining center (Cincinnati Milacron Saber TNC750 VMC). The cutting process of a rectangular workpiece of Al-6061-T680 × 50 × 25 mm^3 is selected as a case study. The cutting tool used is high speed steel (HSS) with 2 flute and 10 mm diameter to represent the most common tool selection in milling industry suitable for slot milling process. The tool moves to cut a stroke of 200 mm.

The cutting speed, feed and depth of cut used are 5000 min^{-1}, 100 mm/min and 5 mm, respectively and they are selected based on the tool manufacturer's recommendations. The cutting forces were measured using a Kistler three-axis dynamometer (type 9255B). The measured cutting force signals (X, Y, and Z directions) were captured and filtered with low path filters (10 Hz cut off frequency) while, the cutting temperature is measured by using the thermocouple (K-Type Testo 925), and each test measurement was repeated three times in order to reduce

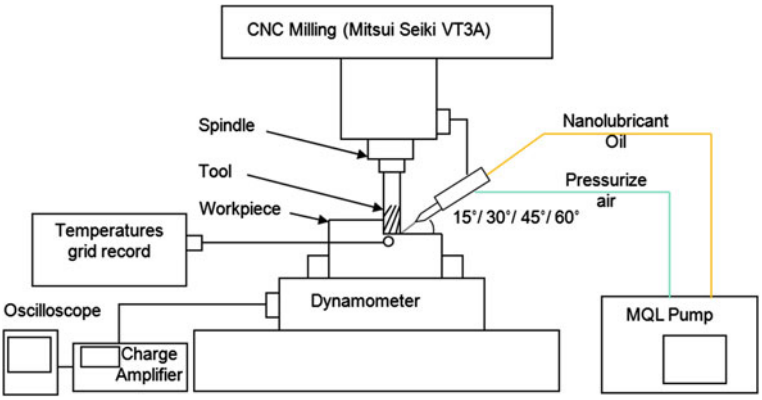


Fig. 1 The experimental set-up

abrupt readings error. The thermocouple has been installed under the machining surface and the measured temperature reflects the amount of heat dissipated in the workpiece. For every machining run, the temperature has been measured at every 2 min while the machined surface roughness has been measured using surface meter (MarSurf PS1 Perthometer) at 700 µm cut off distance.

The nanoparticle-oil is prepared by adding SiO_2 nanoparticles with an average size of 5–15 nm and different concentrations; 0, 02, 0.5 and 1 wt% to the mineral oil followed by sonification (240 W, 40 kHz, 500 W) for 48 h in order to suspend the particle homogeneously in the mixture. In this research work, the mineral oil used is Shell Dromus BL lubricant oil. To deliver the oil to the tool chip interface area, the MQL system is used. The experimentation is carried out using a thin-pulsed jet nozzle that is developed in laboratory and controlled by a variable speed control drive. The nozzle has been equipped with additional air nozzle to accelerate the lubricant into the cutting zone and to reduce the oil consumption up to 25 %. The flexible design allows the injection nozzle to be located at any desired position without interfering with the tool or workpiece during the machining process. The diameter of the nozzle orifices is 1 mm and the MQL oil pressure is set to be 20 MPa with delivery rate of 2 ml/min.

4 Experimental Results

The experimental tests are carried out using the proposed experimental set-up. Figures 2 and 3 show the variety of cutting force at different concentration of nanoparticle and example of surface roughness at 0.8 MPa air pressure, 20,000 min^{-1} spindle speed, 0.25 mm/min feed rate, and 1 mm axial depth of cut, respectively.

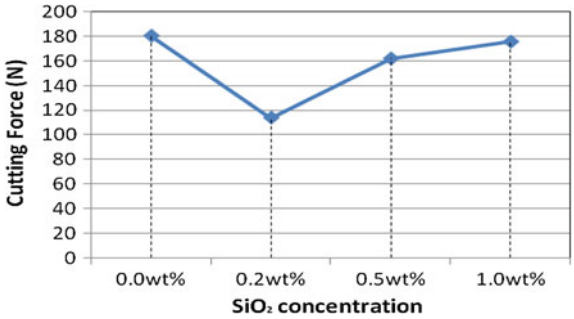


Fig. 2 Variation of cutting force at different concentration of nanoparticle

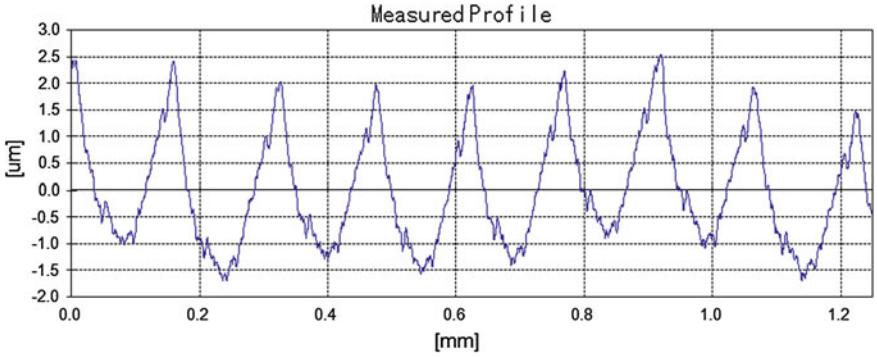


Fig. 3 An example of surface roughness (nanoparticle concentration: 0.0 wt%, air pressure: 1 bar and nozzle angle 15°)

5 ANFIS Modeling

The measured cutting forces, surface roughness and cutting temperature were used as the training data set to build the ANFIS model. Five network layers were used by ANFIS to perform the following fuzzy inference steps as shown in Fig. 4: Layer 1—input fuzzification, Layer 2—fuzzy set database construction, Layer 3—fuzzy rule base construction, Layer 4—decision making, and Layer 5—output de-fuzzification [24, 25].

Layer 1: the output of the node is the degree to which the given input satisfies the linguistic label associated with this node. Gaussian membership functions are chosen to represent the linguistic terms because the relationship between the cutting parameters and roughness is not linear.

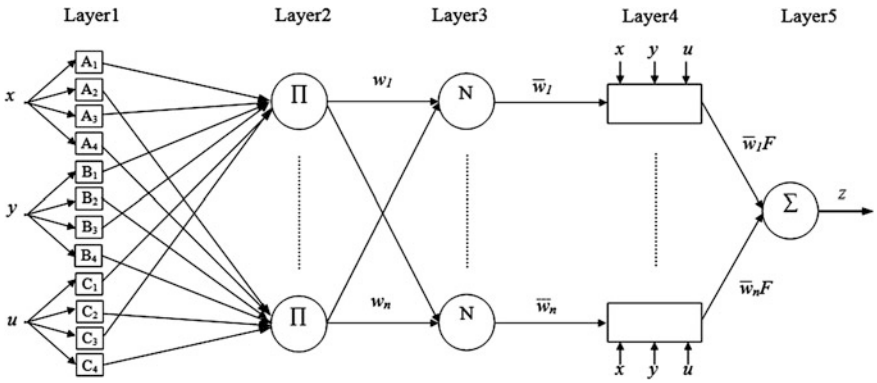


Fig. 4 ANFIS architecture for Sugeno fuzzy model

$$\text{First parameter membership function: } A_i(x) = \exp\left[-0.5((x - a_{i1})/b_{i1})^2\right] \quad (1)$$

$$\text{Second parameter membership function: } B_i(y) = \exp\left[-0.5((y - a_{i2})/b_{i2})^2\right] \quad (2)$$

$$\text{Third parameter membership function: } C_i(u) = \exp\left[-0.5((u - a_{i3})/b_{i3})^2\right] \quad (3)$$

where a_{i1} to a_{i4}, b_{i1} to b_{i4} are the parameter sets.

As the values of these parameters change, the Gaussian functions vary accordingly, thus exhibiting various forms of membership functions on linguistic labels A_i, B_i, and C_i. The parameters in this layer are referred to as principle parameters.

Layer 2: each node computes the firing strength of the associated rule. The nodes of this layer are called rule nodes. The outputs of the top and bottom neurons are as follows:

$$\textit{Top neuron} \quad w_1 = A_1(x) \times B_1(y) \times C_1(u) \quad (4)$$

$$\textit{Second neuron} \quad w_2 = A_1(x) \times B_1(y) \times C_2(u) \quad (5)$$

$$\textit{Bottom neuron} \quad w_n = A_4(x) \times B_4(y) \times C_4(u) \quad (6)$$

Layer 3: every node in this layer is labelled with N to indicate the normalization of the firing levels. The output of the top and bottom neurons is normalized as follows:

$$\textit{Top neuron} \quad \bar{w}_1 = w_1/(w_1 + w_2 + \cdots + w_n) \quad (7)$$

$$\textit{Second neuron} \quad \bar{w}_2 = w_2/(w_1 + w_2 + \cdots + w_n) \quad (8)$$

$$\textit{Bottom neuron} \quad \bar{w}_n = w_n/(w_1 + w_2 + \cdots + w_n) \tag{9}$$

Layer 4: the output of the top and bottom neurons is the product of the normalized firing level and the individual rule output of the first rule and second rule respectively.

$$\textit{Top neuron} \quad \bar{w}_1 z_1 = \bar{w}_1(a_1 x + b_1 y + c_1 u) \tag{10}$$

$$\text{Second neuron} \quad \bar{w}_2 z_2 = \bar{w}_2(a_2 x + b_2 y + c_2 u) \tag{11}$$

$$\textit{Bottom neuron} \quad \bar{w}_n z_n = \bar{w}_n(a_n x + b_n y + c_n u) \tag{12}$$

Layer 5: the single node in this layer computes the overall system output as the sum of all incoming signals, i.e.

$$z = \bar{w}_1 z_1 + \bar{w}_2 z_2 + \cdots + \bar{w}_n z_n \tag{13}$$

If a crisp training set $\{(x^k, y^k, u^k)\, k = 1, \ldots, k\}$ is given, then the parameters of the hybrid neural net (which determine the shape of the membership functions of the premises) can be learned by descent-type means.

The error function for pattern k is given by:

$$E_k = \left(o^k - z^k\right)^2 \tag{14}$$

where o^k is the desired output and z^k is the output computed by the hybrid neural net [21].

6 ANFIS Prediction Model Results

Figure 5a, b are examples to show the relation between input parameters change and cutting force of a predicted by ANFIS model. As can be seen in Fig. 5a, the cutting force is minimum at 0.2 %wt of SiO_2 concentration, the cutting forces increases with the increasing of the SiO_2 concentration. While from Fig. 5b, it is clearly seen that nozzle angle is less significant to change the cutting force.

Figure 6a, b shows the predicted cutting temperature by ANFIS model in relation to lubrication parameters where the temperature is significantly increased with the increasing of both, the air pressure and SiO_2 concentration parameters. However, the lowest cutting force value can be obtained at the lowest value of air pressure (1 bar) and lowest value of SiO_2 concentration parameters (0 %wt, pure oil). From Fig. 6b, it appears that the lowest nozzle angle (15°) and lowest SiO_2 concentration (0 %wt, pure oil) will produce the lowest cutting temperature.

Figure 7a, b shows the predicted surface roughness by ANFIS model in relation to lubrication parameters in machining of Al6061-T6. As can be seen in Fig. 7a, the

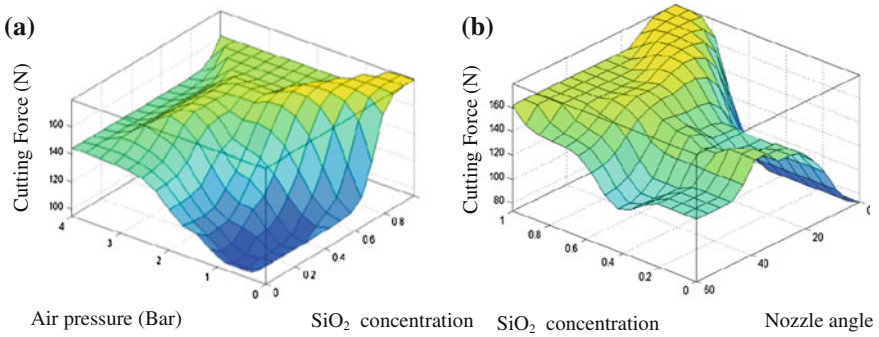


Fig. 5 The predicted cutting force by ANFIS in relation to lubrication parameters

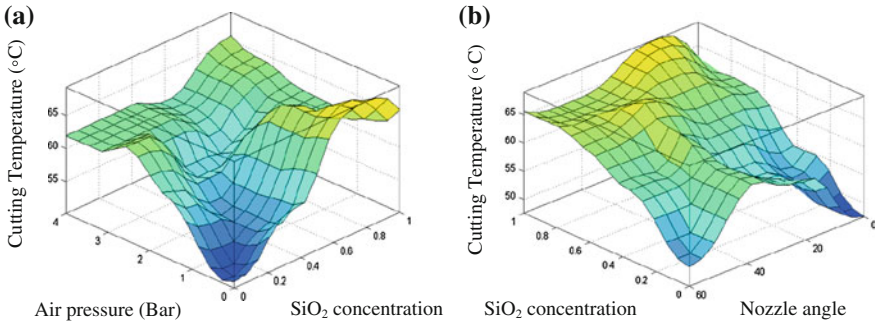


Fig. 6 The predicted cutting temperature by ANFIS model in relation to lubrication parameters

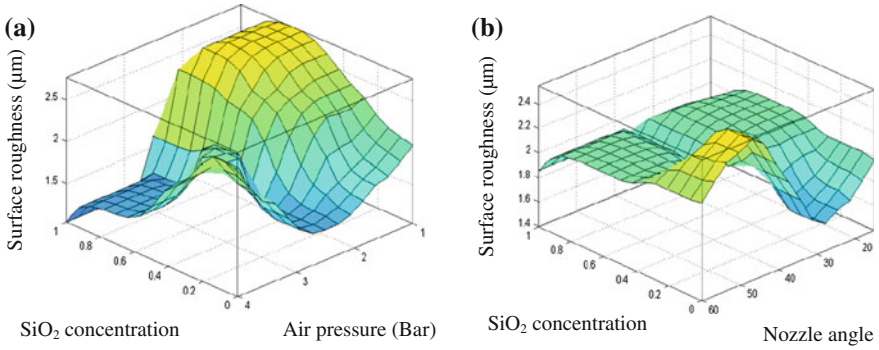


Fig. 7 The predicted surface roughness by ANFIS model in relation to lubrication parameters

surface roughness is significantly decreased with the increase of the both air pressure and SiO_2 concentration parameter. The best surface quality can be obtained at (3 bar) air pressure value and (4 %wt) SiO_2 concentration. However,

From Fig. 7b, it is clearly seen that the best surface quality can be obtained at (30°) nozzle angle.

7 ANFIS Model Accuracy and Error

To investigate the ANFIS model accuracy and error, other new four experimental tests from separated experiment were carried out while the proposed ANFIS model is used to predict the cutting force, cutting temperature and surface roughness at the same conditions.

Table 3 is presenting the parameters input for accuracy and error of the ANFIS model prediction. The individual error percentage is obtained by dividing the absolute difference of the predicted and measured values by the measured value as shown in Eq. (15) where (e_i) is individual error; (R_m) is measured value and (R_p) is predicted value [17].

$$e_i = \left(\frac{|R_m - R_p|}{R_m} \right) \times 100\ \% \tag{15}$$

Meanwhile, accuracy is calculated to measure the closeness of the predicted value to the measured value. The model accuracy is the average of individual accuracy as shown in Eq. (16) where A is the model accuracy and N is the total number of data set tested.

$$A = \frac{1}{N} \sum_{i=1}^{N} \left(1 - \left[\frac{|R_m - R_p|}{R_m} \right] \right) \times 100\ \% \tag{16}$$

The error for data set result was calculated and the model accuracy for ANFIS model was determined. The measured and predicted results of cutting force, cutting temperature and surface roughness are shown in Fig. 8a–c, respectively.

For the cutting force, cutting temperature and surface roughness the highest percentage of error for ANFIS model prediction are 8.97, 3.5 and 16.42 %, respectively. The low level of errors shows that the ANFIS predicted model results were very close with actual experimental values.

Table 3 The parameters input for accuracy and error of the ANFIS model prediction

No of exp.	Parameters (inputs)		
	A	B	C
1	4	3	2
2	3	4	1
3	4	4	1
4	4	2	3

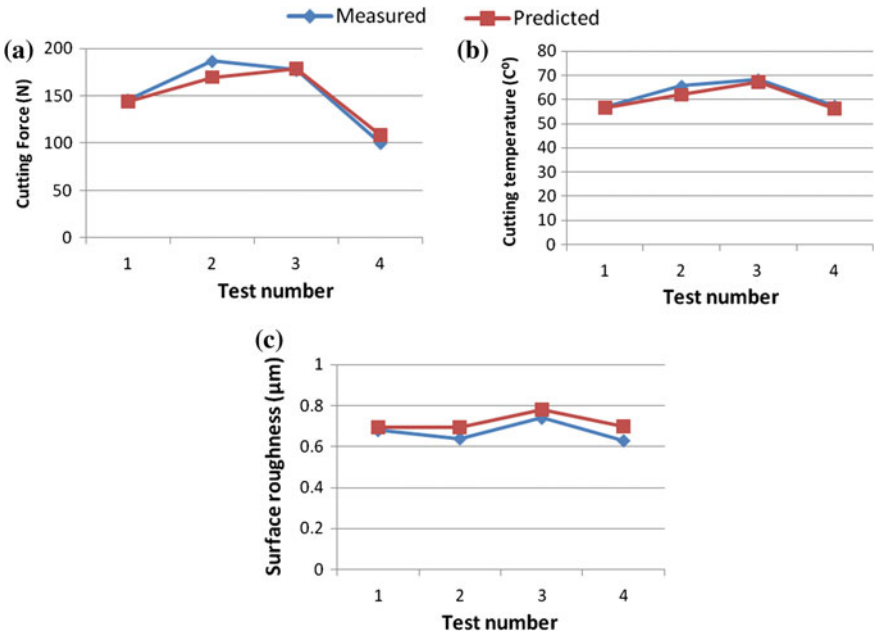


Fig. 8 Comparison of the predicted and measured cutting force, cutting temperature and surface roughness in machining Al7076-T6. **a** Measured and predicted cutting force. **b** Measured and predicted cutting temperature. **c** Measured and predicted surface roughness

8 Discussion

In this study, ANFIS model is proposed to predict cutting force, cutting temperature and surface roughness of Al6061-T6 milling operation using SiO_2 nanolubricant. The result demonstrated settlement between ANFIS model and experimental results with the cutting force, cutting temperature and surface roughness.

As can be observed from the ANFIS model results, extensive dispersed of SiO_2 nanoparticles in cutting oil facilitated by high pressure stream air in cutting zone shows a good performance in reducing cutting force. Tu-Chieh and Yaw-Terng [26] found large interacting force between particle and workpiece would reduce the surface energy of workpiece which is the binding strength between the surface and sub-surface atoms of workpiece [26].

For cutting temperature, ANFIS model results show that, the values is almost similar to the experimental results. In conjunction, in high speed machining operation, the nozzle orientation may be an important factor but very less literature has ever made detail study of the most appropriate nozzle orientation. It may be due to of many conditions which need to be considered such as nozzle specification, cutting operation and desire cutting performance in order to really determine the optimum nozzle orientation. For cutting temperature wise, 15° nozzle angle shows

the optimum, since the measurement of temperature is in workpiece which majority contribution of cutting temperature is from tertiary deformation zone, therefore 15° nozzle orientation successfully withdraw heat from the tertiary zone. However, 30° nozzle angle shows optimum for best surface roughness and chip thickness ratio. It may be related to its orientation is much better in accelerating the cutting oil in the cutting zone and assisting in machining to obtain better surface quality. Again, as mentioned before, the cutting oil in tool-chip interface has negligible on the cutting force and stress in cutting edge, therefore 60° nozzle angle may not be the optimum for chip thickness ratio as cutting force shows optimum in 60° nozzle angle. During the cutting process, the heat generated increases the cutting temperature with an increase of nano-oil concentration, and it may reach the melting temperature of the work material [27]. Associate with the higher temperature, strain energy effects and the present of extreme pressure additive in cutting oil, chemical reaction films are further formed on the machined surface [28].

As per illustrated from ANFIS modeling, the surface roughness are possible to be predicted using soft computing techniques. The surface roughness result illustrated that surface roughness decrease initially and drastically increase when air pressure beyond 2 bar. This may due to the reason which the higher air pressure will lead to formation of welded surface.

Theoretically, machining results reveal that coefficient of friction is reduced at the tool-chip interface by using nanolubricants containing SiO_2 nanoparticles. Reduction of coefficient of friction results lower cutting force. Due to tool wear an increment in cutting force (F_c) and friction force (F_f) is normally observed as presented from Merchant circle as in Fig. 9.

Similarly, the deformation of the chip creates a localized region of intense shear due to the friction at the rake face which is known as secondary shear as it is presented in Fig. 10.

The friction force at the tool-chip interface normally increases due to tool wear which leads to an increment in cutting force and friction force (Fc) component. It is noted that, by adopting the nanolubricant at the tool-chip interface, the coefficient of

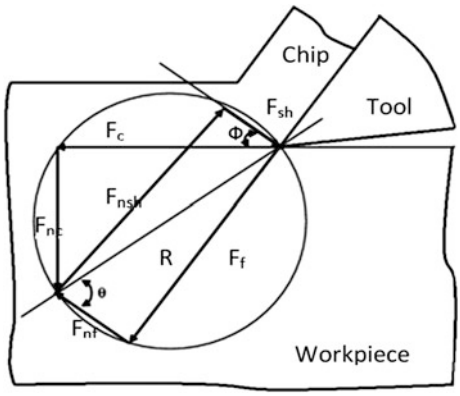


Fig. 9 Merchant circle of cutting mechanism

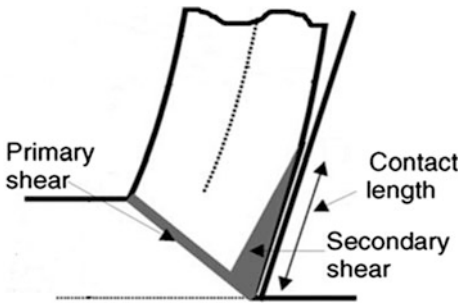


Fig. 10 Shear zones distribution metal cutting process

friction can be reduced. It is believed that nanoparticles deposited on the friction surface and compensate for the loss of mass, which is called 'mending effect'. Due to the porous nature of spherical nanoparticles, it could impart high elasticity, which augments their resilience in a specific loading range and enhances the gap at the tool-workpiece interface. Some particles have rolling effect and some particles are sheared due to very high pressure at cutting zone. The shape of the nanoparticles was changed due to high compression as it is shown in Fig. 11. With increasing the concentration of nanoparticles the degree of shape changing and shearing was increased. For this reason, nanolubrication is capable to reduce the cutting force with less power consumption. Nanoparticles in the mineral oil impart their effect by the combined action of rolling and sliding bearings at the tool chip interface. The rolling and sliding action of the nanoparticles reduces the coefficient of friction

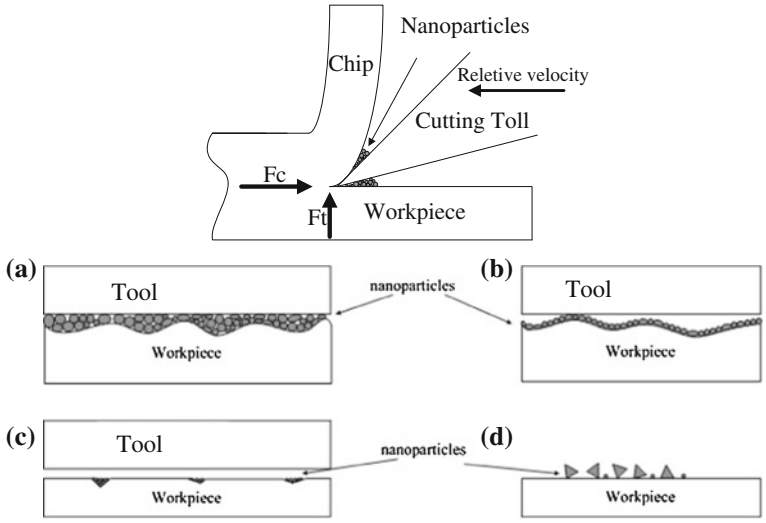


Fig. 11 The presence of nanoparticle reduce the tool-workpiece in contact. **a** Rolling effect. **b** Protective film. **c** Mending effect. **d** Polishing effect

significantly. Therefore, introducing the SiO_2 nanolubricant provides much less friction and superior surface quality due to the tribological properties of these nanoparticles.

9 Conclusion

In this study, ANFIS model has been established to predict the cutting force, cutting temperature and surface roughness of machined surface in Al-6061-T6 milling operation using SiO_2 nanolubricant. The result demonstrated settlement between the ANFIS model and experimental results for cutting force, cutting temperature and surface roughness are 96.195, 98.27 and 91.37 % accuracy, respectively. The close agreement of experimental values of machined surface clearly indicates that the ANFIS model can be used to predict the cutting force, cutting temperature and surface roughness within the range of input parameters under consideration.

Acknowledgments This work was supported by the high impact research (HIR) grant number: HIR-MOHE-16001-00-D000001 from the Ministry of Higher Education, Malaysia.

References

1. Sayuti M, Sarhan AAD, Hamdi M (2014) Performance predictions of using novel SiO_2 nanolubrication in end milling of aerospace AL6061-T6—ANFIS modeling approach. In: Proceedings of the World Congress on Engineering and Computer Science 2014, WCECS 2014, 22–24 Oct 2014. Lecture notes in engineering and computer science, San Francisco, USA, pp 985–991
2. Li Y, Liang SY (1999) Cutting force analysis in transient state milling processes. Int J Adv Manuf Technol 15(11):785–790
3. Reddy NSK, Rao PV (2005) *A* genetic algorithmic approach for optimization of surface roughness prediction model in dry millin*g*. Int J Mach Sci Technol 9(1):63–84
4. Lee J, Cho S, Hwang Y, Cho H-J, Lee C, Choi Y, Ku B-C, Lee H, Lee B, Kim D, Kim SH (2009) Application of fullerene-added nano-oil for lubrication enhancement in friction surfaces. Tribol Int 42:440–447
5. Tao X, Jiazheng Z, Kang X (1996) *T*he ball-bearing effect of diamond nanoparticles as an oil additive. J Phys D Appl Phys 29:2932–2937
6. Lathkar GS, Bas USK (2000) Clean metal cutting process using solid lubricants. In: Proceeding of the 19th AIMTDR conference, Narosa Publishing House IIT Madras, pp 15–31
7. Reddy NSK, Nouari M, Yang M (2010) Development of electrostatic solid lubrication system for improvement in machining process performance. Int J Mach Tools Manuf 50:789–797
8. Ezugwu EO, Bonney J (2005) Finish machining of nickel-base inconel 718 alloy with coated carbide tool under conventional and high-pressure coolant supplies. Tribol Trans 48(1):76–81
9. Sharma VS, Dogra M, Suri NM (2009) Cooling techniques for improved productivity in turning. Int J Mach Tools Manuf 49(6):435–453
10. Nkeki CI, Nwozo CR (2013) Optimal investment under inflation protection and optimal portfolios with stochastic cash flows strategy. IAENG Int J Appl Math 43(2):54–63

11. Sreejith PS, Ngoi BKA (2000) Dry machining: machining of the future. J Mater Process Technol 101(1–3):287–291
12. Masuko M, Aoki S, Suzuki A (2005) Influence of lubricant additive and surface texture on the sliding friction characteristics of steel under varying speeds ranging from ultralow to moderate. Tribol Trans 48(3):289–298
13. Klocke F, Eisenblätter G (1997) Dry cutting. CIRP Ann Manuf Technol 46(2):519–526
14. He S (1982) Solid lubrication materials for high temperature a review. Tribol Int 15:303–314
15. Zalnezhad E, Sarhan AAD, Hamdi M (2013) Adhesion strength predicting of Cr/CrN coated Al7075 using fuzzy logic system for fretting fatigue life enhancement. In: Proceedings of the world congress on engineering and computer science 2013, WCECS 2013, 23–25 Oct 2013. Lecture notes in engineering and computer science, San Francisco, USA, pp 589–595
16. Miaosun S (2000) Solid lubrication materials. China Chemistry Press, Beijing
17. Nakamura T, Tanaka S, Hayakawa K, Fukai Y (2000) A study of the lubrication behavior of solid lubricants in the upsetting process. J Tribol 122:803–808
18. Deshmukh SD, Basu SK (2006) Significance of solid lubricants in metal cutting, in 22nd AIMTDR
19. Alberts M, Kalaitzidou K, Melkote S (2009) An investigation of graphite nanoplatelets as lubricant in grinding. Int J Mach Tools Manuf 49:966–970
20. Peng DX, Kang Y, Hwang RM, Shyr SS, Chang YP (2009) Tribological properties of diamond and SiO_2 nanoparticles added in paraffin. Tribol Int 42:911–917
21. Keblinski P, Phillpot SR, Choi SUS, Eastman JA (2002) Mechanisms of heat flow in suspensions of nano-sized particles (nanofluids). Int J Heat Mass Transf 45:855–863
22. Shamshirband S, Kalantari S, Bakhshandeh Z (2010) Designing a smart multi-agent system based on fuzzy logic to improve the gas consumption pattern. Sci Res Essays 5(6):592–605
23. Ager P, Oseni MI, Yanshio ET (2014) Fuzzy system algorithm on processing methods rheological properties of oil lubricants. In: Proceedings of the world congress on engineering and computer science 2014, WCECS 2014, 22–24 Oct 2014. Lecture notes in engineering and computer science, San Francisco, USA, pp 992–997
24. Tootoonchy H, Hashemi HH (2013) Fuzzy logic modeling and controller design for a fluidized catalytic cracking unit. In: Proceedings of the world congress on engineering and computer science 2013, WCECS 2013, 23–25 Oct 2013. Lecture notes in engineering and computer science, San Francisco, USA, pp 982–987
25. Kasabov NK (1997) Foundations of neural networks, fuzzy systems, and knowledge engineering, vol 33. Computers & mathematics with applications, vol 7. MIT Press, Cambridge. doi:http://dx.doi.org/10.1016/S0898–1221(97)84600-7
26. Tu-Chieh H, Yaw-Terng S (2006) A method for reducing tool wear in a polishing process. Int J Mach Tools Manuf 46:413–423
27. Reza Yousefi YI (2000) A study on ultra—high-speed cutting of aluminium alloy: formation of welded metal on the secondary cutting edge of the tool and its effects on the quality of finished surface. J Int Soc Precisi Eng Nanotechnol 24:371–376
28. Lin YC, So H (2004) Limitations on use of ZDDP as an antiwear additive in boundary lubrication. Tribol Int 37:25–33

Erratum to: Transactions on Engineering Technologies

Haeng Kon Kim, Mahyar A. Amouzegar and Sio-Iong Ao

Erratum to:
H.K. Kim et al. (eds.), *Transactions on Engineering Technologies*, DOI 10.1007/978-94-017-7236-5

The original version of this book was inadvertently published with an incorrect spelling in the editor's name as "Sio-Long Ao", the correct spelling should be "Sio-Iong Ao".

The updated original online version for this book frontmatter can be found at DOI 10.1007/978-94-017-7236-5

H.K. Kim (✉)
Engineering College, Department of Computer and Communication,
Catholic University of DaeGu, DaeGu, Korea, Republic of South Korea
e-mail: hangkon@cu.ac.kr

M.A. Amouzegar
College of Engineering, California State Polytechnic University, Pomona, CA, USA
e-mail: mahyar@csupomona.edu

S.-I. Ao
International Association of Engineers, Hong Kong, Hong Kong SAR
e-mail: publication@iaeng.org

H.K. Kim et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-94-017-7236-5_40

Author Index

H.K. Kim et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-94-017-7236-5

Subject Index

H.K. Kim et al. (eds.), *Transactions on Engineering Technologies*,
DOI 10.1007/978-94-017-7236-5