e - ISSN 2586-9396

Current Applied Science and Technology

Vol. 20 No. 2

May - August 2020

KING MONGKUT'S INSTITUTE OF TECHNOLOGY LADKRABANG

Advisory Board

Current Applied Science and Technology or CAST, formerly KMITL Science and Technology Journal, has been established since its inception as KMITL Science Journal published by King Mongkut's Institute of Technology Ladkrabang (KMITL) in 2001. The journal has been dedicated to publishing advanced and applied knowledge in the form of high-quality research and review articles covering the main areas of Biotechnology, Environmental Science, Agricultural Technology, Food Science and other fields related to Applied Science and Technology. Special issues devoted to important topics in advanced science and technology will occasionally be published.

The journal is an open access peer-reviewed and double blinded journal using Online Journal System (OJS) publishing online academic research and review articles. Previously, articles were published in print on a regular basis (two issues per year) since 2001 and since 2010 onward the articles have been published both in print and electronic forms starting from volume 10. In 2017, the journal title has been changed from *KMITL Science and Technology Journal* to *Current Applied Science and Technology* (CAST) (e-ISSN 2586-9396) to be more identifiable to the international scientific community according to the suggestion of Thai-Journal Citation Index Centre. The journal has been published online only since volume 17(2) (July-December, 2017). In addition, the journal has attracted researchers from other countries more than 22% according to the data. Because of more demands on publication in CAST, the editorial board has decided to publish online original academic research and review articles three issues per year (April, August and December) from 2018 onward.

Furthermore, the advisory board and editorial board comprises honorable and well-known members from around the world in which 50% of editorial board members are from various countries like U.K., Norway, Japan, India, China, Canada, Estonia and Egypt. Only 25% of Thai editorial board members are from the publisher organization and 25% from other publisher organizations. Most of advisory and editorial board members have high H-index according to SCOPUS.

The journal is also committed to maintaining the high level of integrity in the content published and has a Conflict of Interest policy in place. The journal uses plagiarism detection software to screen the submissions. The journal has been working closely with Thai-Journal Citation Index Centre to ensure that the journal complies with international standard of SCOPUS.

Current Applied Science and Technology (CAST)

(formerly KMITL Science and Technology Journal)

Editor

Dusanee Thanaboripat
King Mongkut's Institute of Technology Ladkrabang, Thailand

Editorial Board

CONTENTS

Chicken Feathers Based Keratin Extraction Process Data Analysis Using Response Surface-Box-Behnken Design Method and Characterization of Keratin Product

Kidus Tekleab Welu[1], Surafel M. Beyan[1], Subramanian Balakrishnan[1] and Habtamu Admassu[2]*

[1]Department of Chemical Engineering, College of Biological and Chemical Engineering, Addis Ababa Science and Technology University, Addis Ababa, Ethiopia
[2]Department of Food Process Engineering, College of Biological and Chemical Engineering, Addis Ababa Science and Technology University, Addis Ababa, Ethiopia

Received: 24 October 2019, Revised: 17 January 2020, Accepted: 31 January 2020

Abstract

In the process of production of chicken poultry meat for human consumption, the discarded feather is a solid waste management problem worldwide. There are also millions of tons of feather waste every year from the poultry meat industry in Ethiopia. In general, from this waste feather, keratin constitutes to about 91% which could be augmented for further value-added products such as basic nutrient, medical substance and fertilizer. Hence, this present research is carried out to optimize the extraction and characterization of keratin protein from the waste chicken feather. Keratin extraction using Na_2S as a reducing agent was studied by response surface methodology using the Box-Behnken method. This data analysis is applied for the keratin extraction process to study the effect of the most significant factors such as reducing agent concentration, extraction time and mixing ratio. Applying response surface methodology, a maximum yield of keratin reached 75.39% at sodium sulfide concentration of 0.43 M, extraction time of 5.43 h and mixing ratio of 26.65 g/l. Keratin protein product was characterized using Fourier Transform Infrared spectroscopy (FTIR), UV-Visible spectroscopy, scanning electron microscopy (SEM), and X-Ray Diffraction (XRD). The analysis by FTIR confirmed the presence of chemical compositions such as carboxyl acid and amino groups in the protein samples. The surface morphology studied by scanning electron microscopy analysis showed the formation of porosity and aggregate which were expected on the powder of keratin. Structural studies carried out by X-ray diffraction suggest that sodium sulfide stabilized the β-sheet structure of the protein.

Keywords: chicken feather, keratin protein, sodium sulfide, RSM-BBD
DOI 10.14456/cast.2020.6

*Corresponding author: Tel.: +25 191 321 7366
E-mail: hadtess2009@gmail.com

1. Introduction

A keratinous protein of waste chicken feather has created much attention in recent years. It is the foremost and richest composition of fibrous protein found in chicken feathers, hair, skins, bristles, horns, and hooves [1, 2]. Principally chicken feather contains 91% keratin, which has high mechanical strength [3, 4]. Approximately, billion tons of keratin containing wastes are generated worldwide from the poultry slaughterhouse and wool textile industry every year [5, 6]. Since 5% of the bodyweight of poultry bird is feathers, about three tons of feather waste per day can easily be produced from 50,000 chickens in the slaughterhouse [7]. Chicken feather waste can cause a risk to human health and environmental pollution, therefore, chicken feather discarded in the course of the production of poultry meat for human usage can be a huge problem [8]. There is scant demand for waste chicken feather. Approximately about five billion tons of chicken feather scraps generated annually worldwide by poultry meat producers which are currently disposed by way of landfill or burning or grinding them up to produce farm animals' fodder supplement. Further, feather burnings in special installations are economically expensive [6]. Nowadays, there is a growing interest in manufacturing of value-added materials that are economical and produced from this scrap and renewable resources. Proteins are polymers formed by way of polymerizing different amino acids and the ability to foster intra- and inter-molecular secondary bonds allowing and resulting materials to have a significant difference in their functional groups [7]. Chicken feathers discarded during poultry meat production for human consumption have become a seriously solid and agricultural waste problem world-wide since this waste generates greenhouse gases and also poses danger to the living [8]. There are millions of tons of feather waste generated out of the poultry meat industry in Ethiopia, and it may pose risk to human health and to the environment [3]. The main objectives of this study were to optimize the keratin extraction process using RSM and the characterization of this keratinous protein product.

2. Materials and Methods

2.1 Materials

The feathers of chicken were gotten from Elefora Agro-industries PLC, Debrezeyit, Ethiopia. Sodium sulfide (2M), sodium hydroxide and ammonium sulfate (98%) were acquired from Sigma-Aldrich (St. Louis, MO), USA and the rest of the chemicals used in this research are analytical reagent grade. Products are characterized using the following equipment: UV-visible spectroscopy (CSA C22.2 No 61010-1), Fourier Transform Infrared Spectroscopy (FTIR, PerkinElmer Spectrum IR Version 10.6.1), Scanning Electron Microscope (SEM FEI, INSPCT-F50, and Germany) and X-ray diffractometer (XRD).

2.2 Pre-treatment and keratin extraction

The raw feather was cleaned by soaking in diethyl ether and washed with detergent and dried in an oven at 50^{o}C for 24 h, then it was ground. By this technique, feather sample was cleaned from stains, oil, and grease, etc. The specifics of this method are described by Rouse *et al*. [9] and Sharma *et al*. [10]. Finally, it was stored in a clean closed container at ambient temperature for further study.

2.3 Extraction process optimization using Box Behnken Design (BBD) method

The Box Behnken design is one of the most widely used experimental designs for the Response Surface Methodology (RSM) approach. This is found to be an effective design for sequential experimentation as it provides a reasonable amount of information to test the lack of fit with the required number of experimental values [11]. Prior to following BBD experiments, the levels and chosen parameters for the BBD experiments (mixing ration, extraction time and sodium sulfide concentration) were chosen by results obtained from experimental studies elsewhere through literature survey [12, 13] (Table 1). Sodium sulfide solution at different concentrations, 0.2 M, 0.4 M, and 0.6 M, were prepared using 1l conical flasks. Powdered feathers (20, 25 and 30 g) were added to solution of sodium sulfide. The solutions were continuously mixed for different times (4, 5 and 6 h) at ambient temperature and pH was maintained in the range of 10-13 (alkaline ranges). The solution was filtered and then separated using a centrifuge (10,000 rpm for 5 min) [14]. Then BBD was applied for these experiments. Based on the number of factors and levels required, seventeen experimental runs were conducted to use BBD for the extraction process optimization aided by Design-Expert software 7.0.0 (Stat-Ease, Inc., Minneapolis). To check the adequacy of the variance analysis model (ANOVA) was used. ANOVA was used to obtain desired responses during the course of the experiments. The model used to fit the results of the three-level design is represented by equation 1:

$$Y = b_0 + \sum_{i=1}^{n} b_i X_i + \sum_{i=1}^{n} b_{ii} X_i^2 + \sum_{i=1}^{n-1} \sum_{j=2}^{n} b_{ij} X_i X_j + \varepsilon \quad (1)$$

where X_1, X_2,…, X_n are the input factors that can influence the response Y; n is the number of variables, b_0 is the constant, b_{ii} (i = 1, 2,…n) is the quadratic coefficient, b_{ij} (i = 1, 2,…., n; j = 1, 2,…, n) is the interaction of the coefficient, and ε represents the random error.

Table 1. Levels of variables tested in Box Behnken Design (BBD) for optimization of keratin extraction

Parameters	**Units**	**-1**	**0**	**+1**
Sodium sulfide concentration	M	0.2	0.4	0.6
Extraction time	(h)	4	5	6
Mxing ratio	(g/l)	20	25	30

2.4 Protein precipitation and purification

The feather filtrate solution obtained was transferred into a beaker and then stirred. The solution of ammonium sulphate has been carefully added dropwise [14]. Consequently, the 1:1 ratio of the solution for filtrate and ammonium sulphate was reached. At 4°C, the above solution was subjected to a centrifuge for 5 min at 10,000 rpm and solid particles were stored. With the modification to the method of Shah *et al*. [15], the solid particles were gathered and mixed with de-ionized water till the volume of 100 ml was obtained. The solution was centrifuged at 10,000 rpm for the time duration of 5 min and the solid particles were stored. In 100 ml of 2 M solution of sodium hydroxide, the obtained solid particles were further dissolved. The solution was centrifuged again at 10,000 rpm for 5 min at 4°C and the total liquids were collected and kept for further analysis. Usually, phases of precipitation, washing and dissolution are repeated. The

keratin was precipitated by the isoelectric point that was adjusted with a pH of 3.5 to 4.5 [16, 17]. Finally, it was powdered by a freeze dryer at a temperature of -30ºC for 5 h.

2.5 Characterization of keratin

2.5.1 UV- visible spectroscopy

Freeze-dried keratin was dissolved in distilled water and the solution was analyzed with UV-visible spectrophotometer (CSA C22.2 No 61010-1). The wavelength range was set between 200 nm to 400 nm to determine the desired composition of keratin protein.

2.5.2 Fourier transform infrared spectroscopy (FTIR)

Partially purified freeze-dried keratin powder was used for Fourier Transform Infrared spectroscopy (FTIR) analysis to study functional groups of the products. FTIR run was made between 4000 and 400 cm^{-1} wavenumber range. KBr pellet method was used for performing FTIR analysis.

2.5.3 Scanning electron microscope (SEM)

The morphological structure of the partially purified keratin was studied by using a scanning electron microscope (SEM), model FEI, INSPCT-F50, Germany. The freeze-dried films were fixed on aluminum stubs, coated with a carbon conductive tape and visualized by SEM, operating at 8 kV in a vacuum, and used for identifying the shape and surface morphology of keratin.

2.5.4 X-ray diffraction (XRD)

XRD analysis of keratin was executed using a multi-purpose X-ray diffractometer equipped with a diffraction beam of monochromatic, and a copper marked X-ray tube operated at 0.1 s per step scanning speed to identify the α- helix and β-sheet arrangements in product keratin powder. The data within the scattering angle ranges of 5º to 50º were recorded.

3. Results and Discussion

3.1 Box Behnken Design (BBD) and Response Surface Analysis

Seventeen experiments with a different selection of factor combinations were performed. The experimental and responses of predicted values are given in Table 2. The effects of the factors for all model terms were analyzed by ANOVA. P-values, lack of fit, and R^2-values were used for comparing the models, and accordingly, from a number of possible models, a quadratic model was found to be suitable for the estimation of the given yield as shown by the equation 2 where all variables are given through the coded values.

$$Yield\ (Y) = +74.98 + 1.52 \times A + 0.81 \times B + 0.67 \times C + 0.21 \times A \times B + 2.7 \times B \times C - 5.36 \times A^2 - 1.99 \times B^2 - 2.82 \times C^2 \quad (2)$$

Where, A- Sodium sulfide concentration, B- extraction time and C- mixing ratio.

Table 2. BBD matrix of independent variables used in RSM with corresponding experimental and predicted values of the response

Run no.	Factor 1 A: Concentration of sodium sulfide (M)	Factor 2 B: Extraction time (h)	Factor 3 C: Mixing ratio (g/l)	Yield % Actual (Experimental yield)	Predicted
1	0.2	4	25.00	65.27	65.51
6	0.6	5	20.00	67.45	67.56
8	0.6	5	30.00	68.68	69.09
14	0.4	5	25.00	75.21	74.98
10	0.4	6	20.00	67.48	67.62
3	0.2	6	25.00	66.44	66.71
11	0.4	4	30.00	67.48	67.34
15	0.4	5	25.00	75.45	74.98
17	0.4	5	25.00	73.79	74.98
2	0.6	4	25.00	68.40	68.13
12	0.4	6	30.00	74.52	74.36
13	0.4	5	25.00	75.38	74.98
7	0.2	5	30.00	65.98	65.87
4	0.6	6	25.00	70.41	70.17
9	0.4	4	20.00	71.23	71.39
5	0.2	5	20.00	65.11	64.70
16	0.4	5	25.00	75.09	74.98

3.2 Analysis of variance

The above results were investigated and the calculated determination of coefficient (R^2) for keratin extraction from waste chicken feather was 0.9896 indicating that 98.96% of the response variability could be understood and explained by the statistical model and the model had limitation for extent of 1.04% of the total variation. This implies that the expected values were closer to experimental data and the quadratic polynomial can represent the process for the given experimental domain. The modified R^2 value corrects the R^2 value for the total of factors in the model. The model was found to be highly significant in analyzing the data since the adjusted determination coefficient (Adj R^2 = 0.9761) was found to be very high [18]. "Adeq Precision" determines the signal to noise ratio and if the ratio is higher than four, the model is appropriate. For this study, the ratio of 22.117 showed a tolerable model. Therefore, this model could be useful to navigate the design space. The smaller value of the variation of coefficient, CV= 0.86% indicates the precision with which the experiments were conducted. Similarly, the smaller predicted PRESS statistic shows the better data points fitted the model. Normal probability plots are also a suitable graphical method for judging the normality of the residuals. Residuals normal plot between the normal probability (%) and the internal studentized residuals were obtained to

determine the model approval assumption of analysis of variance. The internally studentized residuals are used to calculate the standard deviations between the experimental and predicted values. Figure 1 showed the relationship between the normal probability (%) and the internal studentized residuals. The straight-line plot indicates that no response variable transformation was required and also apparently there was no problem with normality. The near data straight-line obtained means the residuals were normally distributed and the results were not substantially affected by departures from normality. The predicted values given by a normal distribution as shown in Figure 2 are also plotted against the residual observed, and the points are close to the fitted line that showed a good fit.

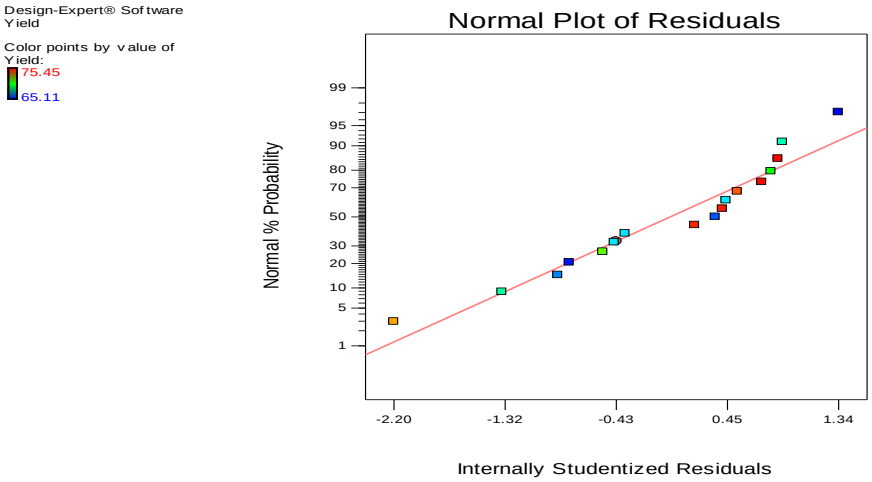


Figure 1. Normal plot of residuals

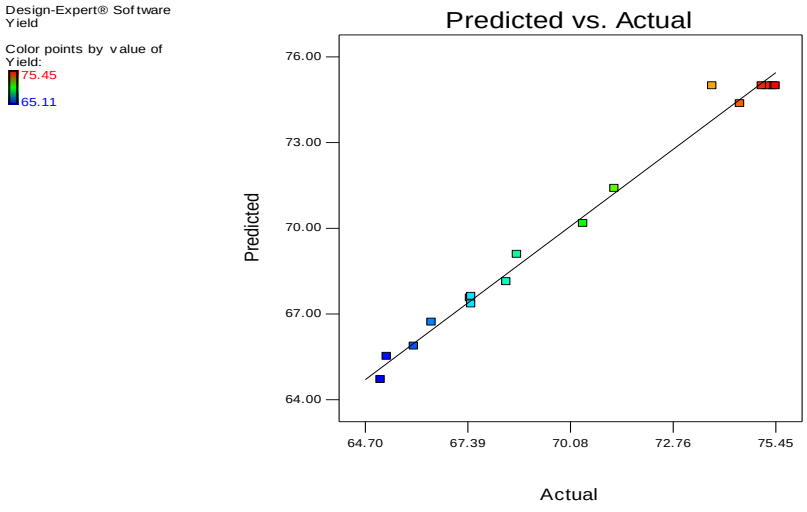


Figure 2. Predicted vs. actual plot of response values

Table 3. ANOVA analysis for response surface quadratic model

Source	Sum of Squares	df	Squares Mean	F value	P value Prob>F	
Model	243.56	9	27.06	73.68	<0.0001	Significant
A	18.42	1	18.42	50.16	0.0002	
B	5.23	1	5.23	14.25	0.0069	
C	3.63	1	3.63	9.89	0.0163	
AB	0.18	1	0.18	0.48	0.5106	
AC	0.032	1	0.032	0.088	0.7751	
BC	29.11	1	29.11	79.25	<0.0001	
A^2	121.11	1	121.11	329.77	<0.0001	
B^2	16.69	1	16.19	45.43	0.0003	
C^2	33.38	1	33.38	90.89	<0.0001	
Residual	2.57	7	0.37			
Lack of Fit	0.71	3	0.24	0.51	0.6979	Not significant
Pure Error	1.86	4	0.47			
Cor Total	246.13	16				

Where: A- Sodium sulfide concentration, B- extraction time and C- mixing ratio

The coefficient term significance is determined by the F-value and p-value. The smaller the value of p and a larger value of F, the more significant is the model. The F-value of 73.68 from the model suggested the model was significant. There is only a 0.10% chance that a "Model F-value" this large might happen because of the noise. Values of "Prob > F" below 0.0500 show that the term of the models are important. In this case, A, B, C, BC, A^2, B^2 and C^2 terms are significant. Lack-of-fit is not significant which represents the model fits well and there is a substantial effect on factors of the output response (Table 3).

3.3 Interaction study

Individual and cumulative effects, as well as the joint interactions between the parameters on the dependent variables, were described using contour plots and response surfaces. In this study, the contour plots and response surface were described by the regression model for BBD which was developed using Design-expert 7.0.0 software. The interaction of concentration of reducing agent (Na_2S) and time of extraction has a progressive effect on the obtained yield of keratin as shown in Figures 3 (a) and (b). It was shown that the keratin yield was increased to 73.79% as Na_2S from 0.2M up to 0.4M and then it was declined when the Na_2S concentrations were further increased. Comparing the lower and the upper limit of the extraction time at a different Na_2S concentration, the keratin with an extraction time of 6 h had a more positive effect on keratin yield than with an extraction time of 4 h. Figures 4 (a) and (b) showed the interaction between concentration of

reducing agent (Na_2S) ($X1$) and mixing ratio ($X2$) with respect to the yield of keratin. Increasing the concentration of reducing agent from 0.325 M to 0.55 M with mixing ratio from 22.25 to 30 g/l enhanced the mass of keratin gained. However, at below 0.325 M and above 0.55 M, a gradual decrease in the response was observed. The interaction effect of extraction time and the mixing ratio was shown in Figures 5 (a) and (b). At the lower limit of the extraction time, the lower value of the mixing ratio (20 g/l) has a high yield of keratin (71.392%) than the higher limit. However, as the extraction time increased a high yield of keratin (74.357%) was obtained at the upper limit of the mixing ratio (30 g/l). Finally, the keratin yield was constant after mixing a ratio of 30 g/l and around the extraction time of 5.75 h.

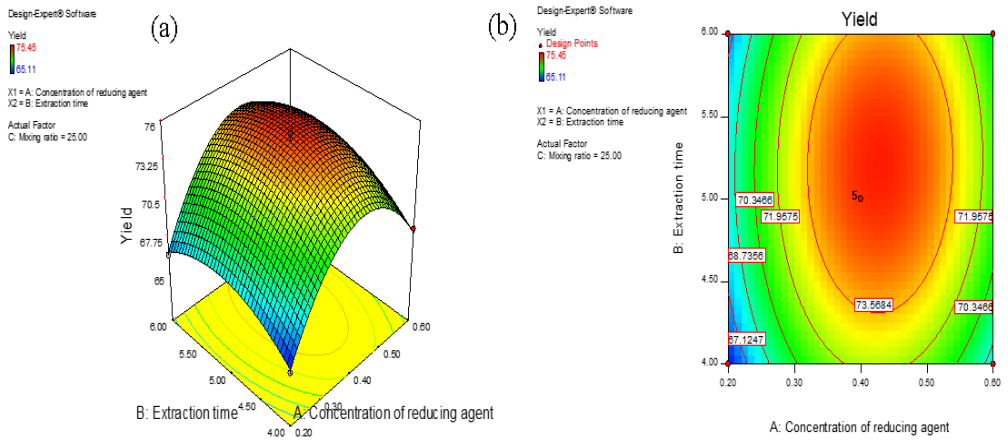


Figure 3. Contour plot (a) and 3D response surface (b) showing the interaction effect of reducing agent and extraction time on the keratin yield

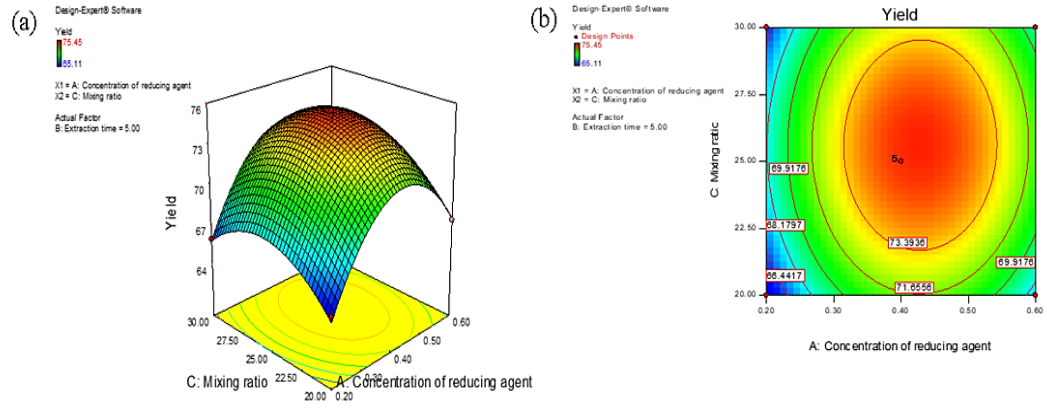


Figure 4. Contour plot (a) and 3D response surface (b) showing the interaction effect of reducing agent and mixing ratio on the keratin yield

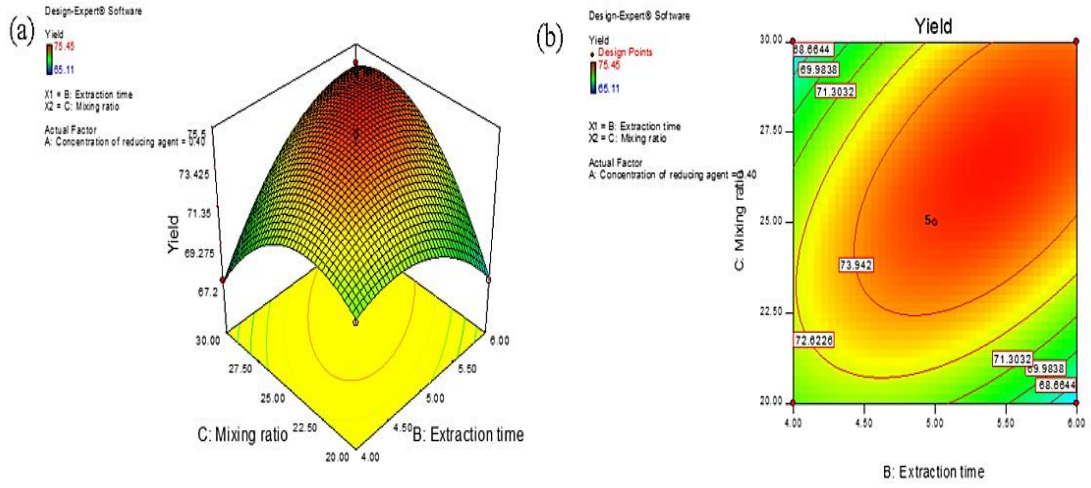


Figure 5. Contour plot (a) and 3D response surface (b) showing the interaction effect of reducing agent and extraction time on the keratin yield

3.4 Optimization of the model and process parameters

The criterion of optimization for the choice of optimal functioning conditions using the quadratic BBD based model was to get the maximum keratin yield with the constrained process factors to the experimental values. Numerical optimization, graphical optimization, and point prediction are ways of expressing optimum process condition and its result. However, all of this optimization method gives similar optimized values for process parameters and the results. A maximum keratin yield of 75.3872% was obtained with the desirability of 0.994 at 0.43M, 5.44 h and 26.65 g/l of Na_2S, extraction time and mixing ratio, respectively.

After optimization, triplicate experiments were performed using the predicted optimized conditions. At these conditions, the mean percentage of keratin yield (74.96%) was obtained. Thus, these results are in correspondence with the predicted values, and hence reflected the applicability of RSM.

3.5 Characterization of keratin protein

3.5.1 UV Visible spectroscopy

The absorption spectra on the solution of keratin exhibited a wide peak in the range of 200-280 nm (Figure 6). UV-Vis absorption measurement of extracted keratin solution showed initial peak wavelength at 220 nm related by the amino acids and carboxylic acid groups forming peptide bond, and the maximum peak at 280 nm caused by the aromatic ring portion of amino acids groups [19]. In general, the fluorescence of keratin was mainly due to tryptophan and tyrosine residues. Keratin is absorbed predominantly in the far UV, but had an absorption extension as far

as 400 nm. The main chromospheres absorbing in the UV region are aromatic compounds of amino acids such as tryptophan, phenylalanine and tyrosine which are existent in the keratin series [20].

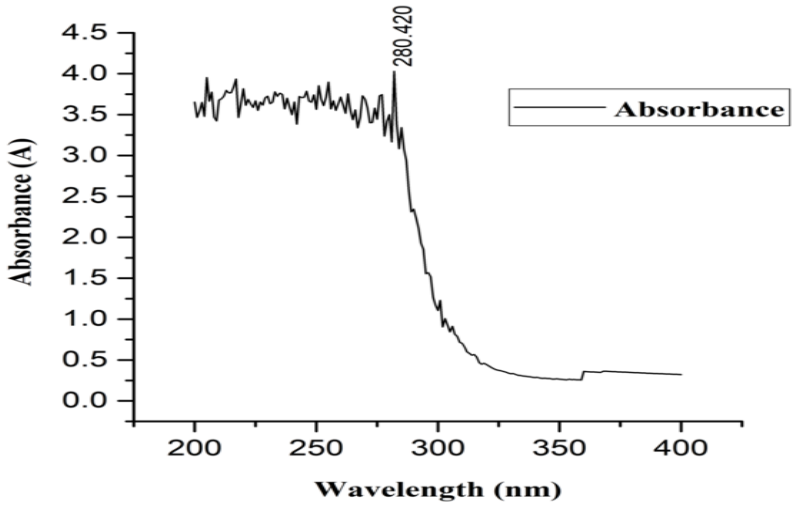


Figure 6. UV visible spectrum of extracted keratin

3.5.2 Fourier transform infrared spectroscopy (FTIR) analysis

The FTIR spectrum obtained for three keratin samples matched the typical spectrum of keratinous protein samples from feathers (Figure 7). All three protein samples show almost similar characteristic peaks. The FTIR characterization of keratin extracts expected to have characteristic peaks corresponding to important functional groups like -CO-NH-, $-NH_2$, -CNH, -C-H. The relatively broad peak in the region of 3300 cm^{-1} corresponds to hydrogen-bonded -N-H and -O-H stretching motion out of amide functionality and absorbed water. The less intense peak in the region of 2900-3100 cm^{-1} signifies -C-H and -N-H groups stretching vibrations. The carbonyl group of amide functionality occurs in the region of 1600-1700 cm^{-1}. An observed peak at 1630 cm^{-1} is assigned to amide carbonyl (-C=O) functional group stretching vibration. The observed peak at 1230 cm^{-1} corresponds to -CNH group comprising -C-N- and -C-C- groups stretching vibrations and -N-H group bending vibration. The bending vibration of $-CH_2$ group occurred at 1450 cm^{-1} [21]. The intense sharp peak at 1525 cm^{-1} corresponds to -C-N-H group bending vibration [22]. The medium intense sharp peaks at 1100 and 920 cm^{-1} correspond to -C-N- group stretching vibration. These FTIR results particularly show the presence of characteristic peaks like amide -N-H, -C=O, -C-N- and -CNH functionalities which confirm building block amino acids forming peptide groups of keratin protein.

3.5.3 Scanning electron microscope (SEM) analysis

The morphology of keratin was observed by SEM with a coated carbon conductive tape in a model FEI, INSPCT-F50, Germany. As shown in Figure 8, the powder keratin was a smooth surface with heterogeneous granulate and texture was merged and embedded. This nature of morphology may be due to the wider poly dispersity of keratin protein. The SEM image of keratin fibers showed a round cross-section and possessed many micropores. This nature of microstructure could be

formed as severe double diffusion keratin filament and small particles in dust form or coagulation bath [18].

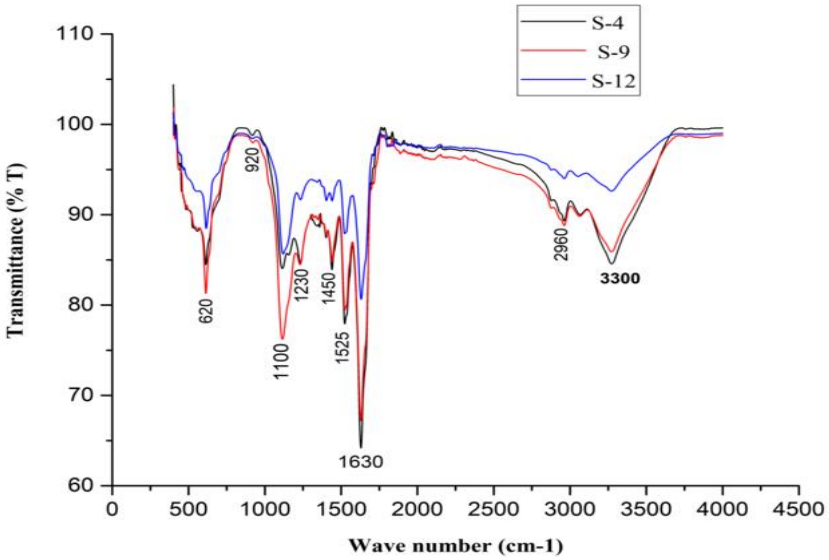


Figure 7. FTIR spectra profile of powdered keratin

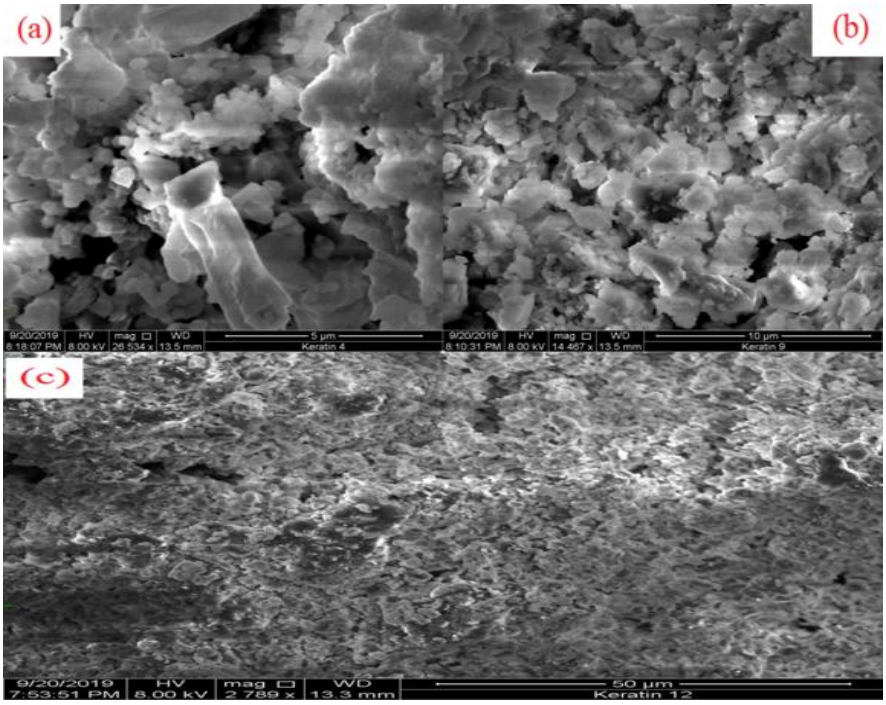


Figure 8. SEM images of keratin:
(a) 26,534× magnification, (b) 14,467× magnification and (c) 2,789 × magnification

3.5.4 X-ray diffraction (XRD) analysis

From the extracted powder keratin, the diffraction angle 2(ɵ) in different peak showed at 5.04º, 9.4º, 11.5º, 17.5º 18.8º, 21.8º and 38.2º (Figure 9). The peak showed in the range of 17.5º to 21.8º represented a protein molecule forming β- sheet structure [5]. The XRD spectra of powder keratin have a slight shoulder or small peak indicating a small ȧ-helix at around 9.4º and most of crystalline characteristics of the keratin have strong intermolecular and intramolecular interaction because of hydrogen bonding [13]. The obtained data showed the effect of the solvent and reducing agent on the crystallization of keratin, and confirmed the tendency of keratin forming β-sheet structure form and sodium sulfide cast form.

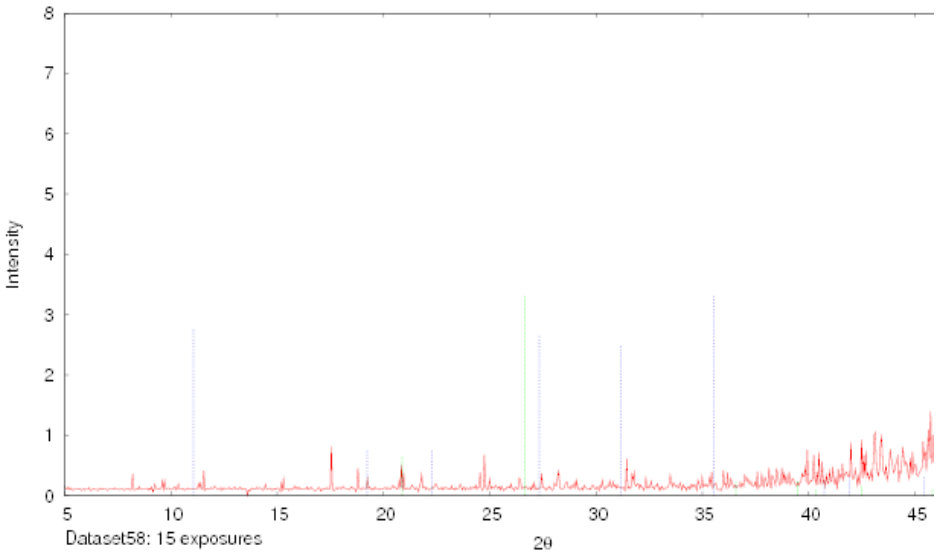


Figure 9. X-ray diffraction of powder keratin

4. Conclusions

Keratin was successfully extracted from chicken feather by using Na_2S as a reducing agent and chemical reduction method. Feathers soaked in diethyl ether, solubilized by sodium sulfide and precipitated with ammonium sulfate and partially purified by NaOH were the most necessary steps to extract keratin. The keratin extraction from the chicken feather waste was optimized for three constraints: reducing agent concentration, extraction time and mixing ratio. The experimental design was performed using the response surface methodology tool and three factors with three levels were used with a total of 17 experiments. The optimum parameters for extraction of keratin were found at reducing agent concentration of 0.43 M, extraction time at 5.53 h and mixing ratio of 26.65 g/l with a keratin yield 75.39% at 0.983 desirabilities using Box- Behnken method. All three parameters had a positive effect on the keratin yield, and from the interaction effects, the extraction time and mixing ratio had a higher significant effect on keratin yield than others. The result of sample characterization showed that non-keratin components were removed through the extraction process. Existence of the fundamental chemical compositions and physical characteristics of keratin was observed using UV-visible spectroscopy, Fourier Transform Infrared Radiation (FTIR) spectroscopy, Scanning Electron Microscopy (SEM) and X-ray Diffraction (XRD) analyses.

References

[1] Esparza, Y., Ullah, A., Boluk, Y. and Wu, J., 2017. Preparation and characterization of thermally crosslinked poly(vinyl alcohol)/feather keratin nanofiber scaffolds. *Matererials and Design*, 135 (5), 1-9.

[2] Reichl, S., Borrelli, M. and Geerling, G., 2011. Keratin films for ocular surface reconstruction. *Biomaterials*, 32, 3375-3386.

[3] Tesfaye, T., Sithole, B. and Ramjugernath, D., 2017. Valorisation of chicken feathers: a review on recycling and recovery route-current status and future prospects. *Clean Technologies and Environmental Policy*, 19 (10), 2363-2378.

[4] Reddy, N. and Yang, Y., 2007. Structure and properties of chicken feather barbs as natural protein fibers. *Journal of Polymer and the Enviroment*, 15, 81-87.

[5] Aluigi, A., Zoccola, M., Vineis, C., Tonin, C., Ferrero, F. and Canetti, M., 2007. Study on the structure and properties of wool keratin regenerated from formic acid. *International Journal of Biological Macromolecules*, 41 (3), 266-273.

[6] Vineis, C., Varesano, A., Varchi, G. and Aluigi, A. 2019. Extraction and characterization of keratin from different biomasses. In: Sharma, S., Kumar, A., eds. *Keratin as a Protein Biopolymer*. Cham: Springer International Publishing, 35-76.

[7] Gupta, A., Perumal, R., Yunus, R.B.M. and Kamarudin, N.B., 2012. Extraction of keratin protein from chicken feather. 8 (6) 732–737. [online] Available at: https://pdfs.semanticscholar.org/ c72d/3cb1ded22997f6aa85ded41f8f1180d9d1aa.pdf

[8] Sharma, S. and Gupta, A. 2016. Sustainable management of keratin waste biomass: applications and future perspectives. *Brazilian Archives of Biology and Technology*, 59: e16150684, http://dx.doi.org/10.1590/1678-4324-2016150684.

[9] Rouse, J.G. and Van Dyke, M.E., 2010. A Review of keratin-based biomaterials for biomedical applications. *Materials*, 3 (2), 999-1014.

[10] Sharma, S., Gupta, A., Chik, S. M.S.T., Kee, C.Y.G., Podder, P.K., Thraisingam, J. and Subramanian, M., 2016. Extraction and characterization of keratin from chicken feather waste biomass : a study. *The National Conference for Postgraduate Research*, Universiti Malaysia Pahang, December, 2016, 693-699.

[11] Yadav, R.N., 2017. A hybrid approach of Taguchi-Response Surface methodology for modeling and optimization of Duplex Turning process. *Measurement*, 100, 131-138.

[12] Eslahi, N., Dadashian, F. and Nejad, N.H., 2013. An investigation on keratin extraction from wool and feather waste by enzymatic hydrolysis. *Prepative Biochemistry and Biotechnology*, 43 (7), 624-648.

[13] Sharma, S., Gupta, A., Kumar, A., Kee, C.G., Kamyab, H. and Saufi, S.M., 2018. An efficient conversion of waste feather keratin into eco-friendly bioplastic film. *Clean Technologies and Environmental Policy*, 20 (10) 2157-2167.

[14] Wingfield, P., 1998. Protein precipitation using ammonium sulfate. *Current Protocols in Protein Science*, Hoboken, NJ, USA: John Wiley & Sons, Inc., A.3F.1-A.3F.8.

[15] Shah, A., Tyagi, S., Bharagava, R.N. and Belhaj, D., 2019. *Keratin as a Protein Biopolymer*. Cham: Springer International Publishing..

[16] Goddard, D. R. and Michaelis, L. 1934. A studies on keratin. *The Journal of biological chemistry*, 106 (2) 605–614.

[17] Dou, Y., Zhang, B., He, M., Yin, G. and Cui, Y., 2016. The structure, tensile properties and water resistance of hydrolyzed feather keratin-based bioplastics. *Chinese Journal of Chemical Engineering*, 24 (3), 415-420.

[18] Kamarudin, N.B., Sharma, S., Gupta, A., Kee, C.G., Chik, S.M.S.B.T. and Gupta, R., 2017. Statistical investigation of extraction parameters of keratin from chicken feather using

Design-Expert. *3 Biotech*, 7 (2), 127, https://doi.org/10.1007/5/3205-017-0767-9.

[19] Idris, A., Vijayaraghavan, R., Rana, U.A., Patti, A.F. and MacFarlane, D.R., 2014. Dissolution and regeneration of wool keratin in ionic liquids. *Green Chemistry*, 16 (5), 2857-2864.

[20] Yamauchi, K., Yamauchi, A., Kusunoki, T., Kohda, A. and Konishi, Y., 1996. Preparation of stable aqueous solution of keratins, and physiochemical and biodegradational properties of films. *Journal of Biomedical Materials Research*, 31 (4), 439-444.

[21] Wojciechowska, E., Włochowicz, A. and Wesełucha-Birczyńska, A., 1999. Application of fourier-transform infrared and raman spectroscopy to study degradation of the wool fiber keratin. *Journal of Molecular Structture*, 511-512, 307-318.

[22] Jose, S., Nachimuthu, S., Das, S. and Kumar, A., 2018. Moth proofing of wool fabric using nano kaolinite. *Journal of Textitle Institute*, 109 (2), 225-231.

Worst Case Analyses of Nearest Neighbor Heuristic for Finding the Minimum Weight *k* - cycle

Tanapat Chalarux and Piyashat Sripratak*

Department of Mathematics, Faculty of Science, Chiang Mai University, Chiang Mai, Thailand

Received: 7 January 2020, Revised: 6 February 2020, Accepted: 12 February 2020

Abstract

Given a weighted complete graph (K_n , w), where w is an edge weight function, the minimum weight k - cycle problem is to find a cycle of k vertices whose total weight is minimum among all k - cycles. Traveling salesman problem (TSP) is a special case of this problem when $k = n$. Nearest neighbor algorithm (NN) is a popular greedy heuristic for TSP that can be applied to this problem. To analyze the worst case of the NN for the minimum weight k - cycle problem, we prove that it is impossible for the NN to have an approximation ratio. An instance of the minimum weight k - cycle problem is given, in which the NN finds a k - cycle whose weight is worse than the average value of the weights of all k - cycles in that instance. Moreover, the domination number of the NN when $k = n$ and its upper bound for the case $k = n - 1$ is established.

Keywords: minimum weight k - cycle, worst case analysis, nearest neighbor heuristic
DOI 10.14456/cast.2020.7

1. Introduction

Traveling salesman problem is one of the most famous problem in mathematics and computer sciences [1]. Let G be a graph and cycle C be a subgraph of G whose vertex set $V(C)$ are the same as $V(G)$, the vertex set of G. C is called a *Hamiltonian cycle* or *tour*. Given a complete graph K_n with a weight function w from the edge set of K_n to the set of positive real numbers, the *symmetric traveling salesman problem* (STSP) seeks for a tour which has the minimum total weight among all tours in the graph. The *asymmetric traveling salesman problem* (ATSP) is defined similarly to the STSP by given a directed complete graph $\vec{K}_n$ instead of a complete graph K_n . The TSP is well-known to be NP-hard [2, 3], so it is difficult to find an optimal solution of the TSP with large number of vertices. There are no polynomial time algorithms to solve either the ATSP or the STSP, unless P = NP. Since the problem has been introduced, many popular heuristics for constructing a tour for the TSP such as greedy heuristics [4], nearest neighbor heuristics [4, 5] and local search heuristics [4, 6] have been proposed.

*Corresponding author: Tel.: +66 61 267 6777
E-mail: psripratak@gmail.com

There are many generalizations of the TSP studied throughout the years. Gutin and Karapetyan [7] have considered the generalized traveling salesman problem. This generalized version is to find a minimum weight cycle C in K_n whose vertex set is partitioned into M partite sets and C is composed of exactly one vertex from each partition set. Khachay and Neznakhina [8] have worked on the generalization of the TSP in the sense of the cycle cover problem (CCP).

In this research, we consider the *minimum weight k - cycle problem*, which is to find a minimum weight k - cycle among all k - cycles in a complete undirected graph K_n with weight function w for a fixed integer $k \le n$. When $k = n$, the minimum weight k - cycle problem and the traveling salesman problem are the same, so we can say that the minimum weight k - cycle problem is a generalization of the STSP. Hence, the minimum weight k - cycle problem is NP - hard.

Gutin *et al*. [9] have shown that greedy type heuristics are not appropriate for the TSP since for each greedy type heuristic they consider, there is an instance of the TSP such that that heuristic constructs a poor result. Precisely, they also show that the domination number of pure greedy heuristic and the NN are 1, and the domination number of the repetitive nearest neighbor heuristic for the STSP is at most 2^{n-3}.

However, when the size of a minimum weight cycle is not n, we cannot ensure that the domination number of the NN is still 1. In this work, we concentrate on the domination number of the NN for some specific values of k, and also consider some other aspects for worst case analyses, namely approximation ratio and no worse than average guarantee.

2. Materials and Methods

We evaluate the NN using three different methods: approximation ratio, no worse than average guarantee and domination number.

The approximation ratio is the most popular way to analyze a heuristic. Many studies use the approximation ratio to analyze heuristics for the TSP. Brecklinghaus and Hougardy [10] find the approximation ratio of the greedy algorithm for some special cases of TSP and that of the Clarke-Wright savings heuristic for the metric TSP. Nilsson [4] refers to the approximation ratio as an evaluation of some of tour construction heuristics and tour improvement heuristics.

We denote $G = (V(G), E(G))$ a graph with vertex set $V(G) = \{1, 2, 3, \ldots, n\}$ and edge set $E(G)$. For any $u, v \in V(G)$, denote $e = \{u, v\}$ an edge from vertex u to vertex v. We call a pair of complete undirected graph K_n and its weight function w an *instance* (K_n, w) of the minimum weight k - cycle problem. Denote $w(u, v)$ a weight of an edge $\{u, v\}$ of graph G and $w(G)$ the sum of the weights of all edges in graph G. We denote $w(S)$ the maximum weight of a k - cycle constructed by heuristic and $w(T)$ the weight of a minimum k - cycle.

Definition 2.1 A heuristic A for the minimum weight k - cycle problem has the *approximation ratio* $\alpha \ge 1$ if for each instance with minimum k - cycle T, the heuristic A finds a k - cycle S such that $\frac{w(S)}{w(T)} \le \alpha$.

We also analyze the heuristic by checking whether it is worse than average. The idea of not worse than average heuristic is introduced in Russian literature. Punnen *et al*. [11] used average value based analysis on some heuristics for the bipartite boolean quadratic programming problem.

Definition 2.2 For each instance I, the average value of weights of all k - cycles is called the *average value* of I, and heuristic A is said to be *not worse than average* if heuristic A constructs a k - cycle of weight less than or equal to the average value of I for all instance I.

The domination number suggested by Glover and Punnen [12] is a new approach for evaluating heuristics. They propose a heuristic for the TSP with complexity $O(n)$ and give the domination number for that heuristic. Gutin *et al*. [9] studied the domination number of some greedy type heuristics for TSP.

For an instance of the minimum weight k - cycle problem, let H and S be k - cycles, we say that H *dominates* S if $w(H) \le w(S)$.

Definition 2.3 The *domination number* of a heuristic A for the minimum weight k - cycle problem on K_n is the maximum integer $d(n, k)$ such that, for any instance of the minimum weight k - cycle problem on n vertices, A produces a k - cycle K which dominates at least $d(n, k)$ cycles in I including K itself.

Among all greedy type heuristics studied in Gutin *et al*. [9], the *nearest neighbor heuristic* (NN) is a heuristic that we are interested since it can be directly applied to our problem. Let u be a vertex in graph G with vertex set $V(G)$. A vertex v is called *nearest vertex* from u if $w(u, v) = \min\{w(u, h): h \in V(G)\}$.

The NN starts constructing a cycle from a fixed vertex i_1, goes to i_2, which is the nearest vertex from i_1 ($w(i_1, i_2) = \min\{w(i_1, j): j \ne i_1\}$), and adds edge $\{i_1, i_2\}$, then to i_3, the nearest vertex from i_2 distinct from i_1 and i_2, and adds edge $\{i_2, i_3\}$. Repeat until we collect k vertices. Then add edge $\{i_k, i_1\}$.

In this work, we represent a path $P = (V(P), E(P))$ where $V(P) = \{v_1, v_2 \ldots, v_k\}$ and $E(P) = \{\{v_1, v_2\}, \{v_2, v_3\}, ..., \{v_{k-1}, v_k\}\}$ as $(v_1, v_2 \ldots, v_k)$. Denote P_n a path with $|V(P)| = n$. We say that P has length $n - 1$. The graph C constructed from a path P by adding an edge $\{v_k, v_1\}$ is called a *cycle*, denoted by $C = (v_1, v_2 \ldots, v_k, v_{k+1} = v_1)$.

3. Results and Discussion

Henceforth, denote $w(S)$ the maximum weight of a k - cycle constructed by the NN and $w(T)$ the weight of the minimum k - cycle.

In the following theorem, we show that approximation ratio is not appropriate for analyzing the NN for the minimum weight k - cycle problem on K_n. For any positive integer $n \ge 3$ and for each $\sigma \ge 1$, we can find an instance of the minimum weight k - cycle problem on K_n such that the ratio of $w(S)$ and $w(T)$ is greater than or equal to σ.

Theorem 3.1 Let n and k be positive integers where $n \ge 4$ and $3 \le k < n$. For any $\sigma \ge 1$, there exists an instance of the minimum weight k - cycle problem on complete graph K_n such that $\frac{w(S)}{w(T)} \ge \sigma$.

Proof Let $\sigma \geq 1$ and $3 \leq k < n$. The following instance of the minimum weight k - cycle problem on K_n is considered. Assume that edges $\{i, i + 1\}$ for $1 \leq i \leq n-1$ and edge $\{n, 1\}$ have weight σ, edge $\{k, 1\}$ has weight $\sigma^2(k+1)-(k-1)\sigma$, and all remaining edges have weight 2σ. Applying the NN starting at vertex 1, the NN constructs the k - cycle $C = (1, 2, 3, \ldots, k, 1)$. Then

$$w(S) \geq w(C) = \sigma(k-1)+(\sigma^2(k+1)-\sigma(k-1)) = \sigma^2(k+1).$$

Next, we show that k - cycle $T = (1, 2, 3, \ldots, k-1, n, 1)$ is a minimum k - cycle. Suppose that H is a k - cycle such that $w(H) < w(T)$. Note that edges of weight σ are the cheapest edges. The weight of $\{k, 1\}$ is $\sigma^2(k+1)-(k-1)\sigma > 2\sigma$. Moreover, T contains $k-1$ edges of weight σ and one edge of weight 2σ. Then H contains k edges of weight σ, which is impossible.

Since $T = (1, 2, 3, \ldots, k-1, n, 1)$ is the minimum k - cycle, we have $w(T) = \sigma(k-1)+2\sigma = \sigma(k+1)$.

Therefore, $\dfrac{w(S)}{w(T)} \geq \dfrac{\sigma^2(k+1)}{\sigma(k+1)} = \sigma$.

In case of $k = n$, there exists an instance of the minimum weight k - cycle problem defined as follows:

Assume that any edge $\{u, u + 1\}$ where $1 \leq u \leq n-1$ and the edge $\{n, 1\}$ have weight σ, the edge $\{n-3, n-1\}$ has weight $\sigma-1$, the edge $\{n-2, n\}$ has weight $\sigma(n\sigma-n+2)$, and all remaining edges have weight 2σ. It is possible that the NN starting at 1 constructs k - cycle $C = (1,2,\ldots,n-3,n-1,n-2,n,1)$ with $w(C) = (n-1)\sigma-1+\sigma(n\sigma-n+2) = \sigma^2 n+\sigma-1$.

Next, we claim that the tour $T = (1,2,\ldots,n,1)$ is a minimum k - cycle with $w(T) = n\sigma$. Suppose that H is a k - cycle such that $w(H) < w(T)$. Thus, H is composed of the only edge that has weight less than σ, which is $\{n-3, n-1\}$ of weight $\sigma-1$, and a Hamiltonian path P from $n-3$ to $n-1$.

If there is an edge of weight 2σ or $\sigma(n\sigma-n+2) > 2\sigma$ in P, then

$$w(H) = w(n-3,n-1)+w(P) \geq \sigma-1+(n-2)\sigma+2\sigma = n\sigma+\sigma-1 > n\sigma = w(T),$$

a contradiction. Hence, all edges in P are of weight σ. Therefore, $n-3$ is adjacent to $n-4$ or $n-2$, and $n-1$ is adjacent to $n-2$ or n. Since $k = n \geq 4$, $n-2$ cannot be adjacent to both $n-3$ and $n-1$. Hence, an edge incident to $n-2$ has weight at least 2σ, a contradiction. Therefore, T is a minimum k - cycle.

Hence, $\dfrac{w(S)}{w(T)} \geq \dfrac{\sigma^2 n+\sigma-1}{n\sigma} > \sigma.$

In the next proposition, we find a general formula of the average value of the weights of all k - cycles for any instance.

Theorem 3.2 Let n and k be positive integers where $n \geq 3$ and $3 \leq k \leq n$. For an instance of the minimum weight k - cycle problem (K_n, w), the average value of the weights of all k - cycles is $\dfrac{2k}{n(n-1)}\sum_{e\in E(K_n)} w(e)$.

Proof We find the number of all k - cycles in a complete graph K_n. There are $\binom{n}{k}$ ways to choose k vertices from n vertices to be in a k - cycle. Since the number of cycles composed of distinct k vertices is $\frac{(k-1)!}{2}$, the number of all k - cycles in a complete graph K_n is $\binom{n}{k}\frac{(k-1)!}{2}$. Next, we find the summation of weights of all k - cycles. We consider an arbitrary edge $\{a, b\}$. To construct a k - cycle, we find a path of k vertices from a to b. There are $\binom{n-2}{k-2}(k-2)!$ possible paths of k vertices from a to b. Thus, the summation of weights of all k - cycles is $\binom{n-2}{k-2}(k-2)!\sum_{e\in E(K_n)} w(e)$. The average value of weights of all k - cycles is

$$\frac{\binom{n-2}{k-2}(k-2)!\sum_{e\in E(K_n)} w(e)}{\binom{n}{k}\frac{(k-1)!}{2}} = \frac{2k}{n(n-1)}\sum_{e\in E(K_n)} w(e).$$

Hence, the average value of the weights of all k - cycles is $\frac{2k}{n(n-1)}\sum_{e\in E(K_n)} w(e)$.

Next, we show that the NN for the minimum weight k - cycle problem can be worse than average by constructing an instance such that the NN constructs a cycle of weight greater than the average value of the weights of all k - cycles in that instance. In the case that $k = n = 3$, the NN constructs the optimal solution for all instances since there is only one possible k - cycle in each instance. Then we consider the case when $n \geq 4$ and $3 \leq k \leq n$.

Theorem 3.3 Let n and k be positive integers where $n \geq 4$ and $3 \leq k \leq n$. There is an instance such that the NN for the minimum weight k - cycle problem is worse than average.

Proof Consider an instance of the minimum weight k - cycle problem on K_n such that all edges have weight 1 except edge $\{1, k\}$ where $w(1,k) = 1+\frac{n}{2}(n-1)$. The average value of this instance is

$$\frac{2k}{n(n-1)}\{\frac{n}{2}(n-1)-1+\frac{n}{2}(n-1)+1\} = 2k.$$

Apply the NN starting at vertex 1. One of the cycles that we can obtain from the NN is the k - cycle $S = (1, 2, 3, \ldots, k, 1)$ with weight $k-1+1+\frac{n}{2}(n-1) = k+\frac{n}{2}(n-1) > 2k$. Thus, the NN for the minimum weight k - cycle problem is worse than average.

We consider the domination number of the NN for $k = n - 1$ and $k = n$. We first show an upper bound of the domination number of the NN for the case $k = n - 1$.

Theorem 3.4 Let n be a positive integer where $n \geq 4$. The domination number of the NN for the minimum $(n - 1)$ - cycle problem on K_n is at most $\frac{1}{2}(n-2)!+1$.

Proof For any edge $\{i, j\} \in E(K_n)$, we define weight function w as follow:

$$w(i,j) = \begin{cases} in; & 1 \le i \le n-2,\ j = i+1 \\ in+1; & 2 \le i \le n-2,\ j \ge i+2 \\ n+1; & i = 1, 3 \le j \le n-2 \text{ or } j = n \\ (n-1)(n-2)n+1; & i = 1,\ j = n-1 \\ 1; & i = n-1,\ j = n. \end{cases}$$

Suppose that the NN starts at vertex 1. Then the $(n-1)$ – cycle constructed by the NN is $C_{NN} = (1,2,3,\ldots,n-1,1)$ where $w(C_{NN}) = n\sum_{i=1}^{n-2} i + (n-1)(n-2)N + 1$. We consider an $(n-1)$ - cycle H that does not contain edge $\{1, n-1\}$. Since

$$w(1,n-1) > (n-1)\cdot max\{w(i,j) \mid \{i,j\} \in E(K_n) \setminus \{\{1,n-1\}\}\}, w(H) < w(C_{NN}).$$

We can see that the number of $(n-1)$ - cycles which do not contain edge $\{1, n-1\}$ is

$$\frac{n}{2}(n-2)! - (n-2)!.$$

In the next step, we consider $(n-1)$ - cycle $C \ne C_{NN}$ which contains edge $\{1, n-1\}$. This $(n-1)$ - cycle C is composed of an edge $\{1, n-1\}$ and the path P of length $n-2$ starting from 1 and ending at $n-1$, where $P = (1 = v_1, v_2, v_3, \ldots, v_{n-1} = n-1)$. Assume that C does not contain vertex n. Let $B = \{\{v_i, v_{i+1}\} \in E(P) \mid v_i > v_{i+1}\}$. Since $C \ne C_{NN}$, B is not empty. Then

$$w(C) \le w(C_{NN}) + N \sum_{\{v_i, v_{i+1}\} \in B} (v_{i+1} - v_i) + \varepsilon(n-2) < w(C_{NN}).$$

Thus, the $(n-1)$ - cycle $C \ne C_{NN}$, which contains edge $\{1, n-1\}$ and does not contain vertex n, is not dominated by C_{NN}. The number of these $(n-1)$ - cycles is $(n-3)! - 1$.

Next, we consider the case when C contains vertex n. Assume that $v_i = n$. Let u be a vertex that does not show in $(n-1)$ - cycle C and $B' = \{\{v_j, v_{j+1}\} \in E(P) \mid v_j > v_{j+1}, v_j \ne n\}$. Then

$$w(C) \le w(C_{NN}) + (v_{i+1} - u)N + \sum_{\{v_j, v_{j+1}\} \in B'} (v_{j+1} - v_j)N + (n-2)\varepsilon.$$

Since $\sum_{\{v_j, v_{j+1}\} \in B'} (v_{j+1} - v_j)N \le 0$, we have $w(C) \le w(C_{NN}) + (v_{i+1} - u)N + (n-2)\varepsilon$. If $v_{i+1} - u < 0$, we can conclude that $w(C) < w(C_{NN})$. Hence, we count the number of $(n-1)$ - cycles which contain edge $\{1, n-1\}$ and vertex n where $v_{i+1} < u$. We see that the number of ways to choose vertices u and v_{i+1} satisfying the condition $u > v_{i+1}$ is $\frac{1}{2}(n-3)(n-4)$. Since the path P is a sequence of $n-1$ vertices starting with $v_1 = 1$ and ending with $v_{n-1} = n-1$, we just need to fill in the remaining $n-3$ positions by the $n-3$ vertices other than vertices 1, u and $n-1$, so that n is followed by the given v_{i+1}. There are $(n-4)!$ ways to complete this step. Thus, the number of ways to construct $(n-1)$ - cycle satisfying the condition is $\frac{1}{2}(n-3)!(n-4)$.

From all cases, there are at least

$$\left(\frac{n}{2}(n-2)! - (n-2)!\right) + \left((n-3)! - 1\right) + \left(\frac{1}{2}(n-3)!(n-4)\right) = \frac{n}{2}(n-2)! - \frac{n-2}{2}(n-3)! - 1$$

$(n-1)$ - cycles that are not dominated by C_{NN}. Note that the number of all $(n-1)$ - cycles in K_n is $\frac{n}{2}(n-2)!$. Then C_{NN} can dominate at most $\frac{n}{2}(n-2)!-(\frac{n}{2}(n-2)!-\frac{n-2}{2}(n-3)!-1)=\frac{1}{2}(n-2)!+1$ $(n-1)$ - cycles including itself. Hence, the domination number of the NN for the minimum weight $(n-1)$ - cycle problem is at most $\frac{1}{2}(n-2)!+1$.

We can slightly modify the instance used in the proof of Theorem 3.4 to give an instance in which the NN gives the unique maximum weight k - cycle. As a result, the domination number of the NN for the minimum weight k - cycle problem for $k = n$ is 1. Since our problem becomes the STSP when $k = n$, the next theorem offers a verification for the domination number of the NN for the STSP, which is mentioned without proof by Gutin *et al*. [9]. Hence, the proof is omitted.

Theorem 3.5 Let n be a positive integer where $n \geq 3$. The domination number of the NN for the minimum weight n - cycle problem on K_n is 1.

4. Conclusions

We point out that approximation ratio is not an appropriate method to analyze the NN for the minimum weight k - cycle problem. We show that for any $\sigma \geq 1$, there is an instance of the minimum weight k - cycle problem such that the ratio of the maximum weight of k - cycle constructed by the NN and the weight of a minimum k - cycle is greater than or equal to σ.

Secondly, we find the average value of the weights of all k - cycles for each instance, and construct an instance of the minimum weight k - cycle problem such that the weight of a k - cycle constructed by the NN is greater than the average value of the weights of all k - cycles in the instance. Thus, the NN for the minimum weight k - cycle problem can be worse than average.

Finally, we establish an upper bound $\frac{1}{2}(n-2)!+1$ for the domination number of the NN for the minimum $(n-1)$ - cycle problem. Moreover, we prove that the domination number of the NN for the STSP is 1.

Even the NN can give the unique worst solution for some instance of the TSP, the domination analysis shows that it is more promising when $k \neq n$. The output from the NN can be used as an initial solution in more complicated heuristics to get a better solution.

References

[1] Lawler, E.L., Lenstra, J.K., Rinnooy Kan, A.H.G. and Shmoys, D.B., 1985. *The Traveling Salesman Problem*. Chichester: John Wiley and Sons Ltd.

[2] Gutin, G., 2001. TSP tour domination and Hamilton cycle decomposition of regular digraphs. *Operations Research Letters*, 28, 107-111.

[3] Gutin, G., Punnen, A.P. and Lenstra, J.K., 2002. *The Traveling Salesman Problem and Its Variations*. Boston: Kluwer Academic Publishers.

[4] Nilsson, C., 2003. *Heuristics for the traveling salesman problem*. [online] Available at: http://160592857366.free.fr/joe/ebooks/ShareData/Heuristics%20for%20the%20Traveling%20Salesman%20Problem%20By%20Christian%20Nillson.pdf

[5] Khan, A.A. and Agrawala, H., 2016. Comparative study of nearest neighbour algorithm and genetic algorithm in solving traveling salesman problem. *International Research Journal of Engineering and Technology*, 3, 234-238.
[6] Wu, Y., Weise, T. and Chiong, R., 2015. Local search for the traveling salesman problem: a comparative study. *Proceedings of IEEE 14th International Conference on Cognitive Informatics and Cognitive Computing*, 213-220.
[7] Gutin, G. and Karapetyan, D., 2009. Generalized traveling salesman problem reduction algorithms. *Algorithmic Operations Research*, 4, 144-154.
[8] Khachay, M. and Neznakhina, K., 2016. Approximability of the minimum-weight k -size cycle cover problem. *Journal of Global Optimization*, 66, 65-82.
[9] Gutin, G., Yeob, A. and Zverovich, A., 2002. Traveling salesman should not be greedy: domination analysis of greedy-type heuristics for the TSP. *Discrete Applied Mathematics*, 117, 81-86.
[10] Brecklinghaus, J. and Hougardy, S., 2015. The approximation ratio of the greedy algorithm for the metric traveling salesman problem. *Operations Research Letters*, 43, 259-261.
[11] Punnen, A.P., Sripratak, P. and Karapetyan, D., 2015. Average value of solutions for the bipartite boolean quadratic programs and rounding algorithms. *Theoretical Computer Science*, 565, 77-89.
[12] Glover, F. and Punnen, A.P., 1997. The traveling salesman problem: new solvable cases and linkages with the development of approximation algorithms. *Journal of the Operational Research Society*, 48, 502-510.

Algorithm for Solving Parallel Machines Scheduling Problem to Minimize Earliness and Tardiness Costs

Pensiri Sompong*

Kasetsart University, Chalermphrakiat Sakon Nakhon Province Campus,
Sakon Nakhon, Thailand

Received: 24 December 2019, Revised: 6 February 2020, Accepted: 14 February 2020

Abstract

Algorithm for parallel machines scheduling problem to minimize the earliness and tardiness costs is proposed in this study. The problem is associated with the assignment of jobs to machines and determination of staring time for each job in a given sequence. Population-based incremental learning (PBIL) algorithm is used to allocate the jobs to machines. The optimal timing algorithm based on the minimum block cost function calculation is then employed to decide the starting time of jobs on each machine. To illustrate the performance of proposed algorithm, numerical examples generated randomly are tested. The numerical results obtained from PBIL combined with optimal timing algorithm called PBILOTA are compared to EDDPM (Earliest Due Date for Parallel Machines) to indicate the decrease in penalty cost. From the experimental results, it is shown that PBILOTA is an efficient algorithm for solving parallel machines scheduling problem with earliness-tardiness costs minimization.

Keywords: population-based incremental learning algorithm, scheduling, parallel machines, earliness, tardiness
DOI 10.14456/cast.2020.8

1. Introduction

Study related to minimization of earliness and tardiness costs plays an important role in production system due to the Just-in-Time (JIT) production philosophy. The problem associated with the objective to minimize earliness and tardiness penalties has been focused on algorithm design making the jobs finished exactly on their due dates or as close as possible. Lee and Choi [1] considered a job scheduling problem for a single machine to minimize early-tardy penalty costs. An optimal timing algorithm was used to decide the optimal starting time of each job in a given sequence generated by genetic algorithm. The optimal starting time was obtained by shifting a block to a point giving the minimum block cost function. Bauman and Józefowska [2] proposed an algorithm for solving a single machine scheduling problem with linear earliness and tardiness costs.

*Corresponding author: Tel.: +66 42 72 5033 Fax: +66 42 72 5034
E-mail: pensiri.so@ku.th

The objective was to find a vector of job completion time such that the total cost is minimum. Kedad-Sidhoum and Sourd [3] proposed fast neighborhood search based on a block representation of schedule. To get a larger neighborhood search, random swap and earliness-tardiness perturbation were performed. Feasible solutions were considered as a vector of the completion times of all jobs. Kianfar and Moslehi [4] developed a branch and bound algorithm for solving a single machine scheduling problem to minimize the weighted quadratic earliness and tardiness penalties such that no machine idle time was allowed. An arc-time-indexed formulation and a branch and bound algorithm were presented by Keshavarz *et al*. [5] to investigate a single machine sequence dependent group scheduling problem combined with earliness and tardiness considerations. The single machine scheduling problem with distinct time windows and sequence dependent setup time was addressed by Rosa *et al*. [6]. The problem involved the determination of job sequence and starting time of each job in the sequence. Implicit enumeration and general variable neighborhood search algorithms were proposed to determine the job sequence and idle time insertion algorithm was then used to determine the starting time for each job. In addition, parallel machines system has been considered. Kayvanfar *et al*. [7] studied scheduling problem on unrelated parallel machines with the objective to minimize earliness-tardiness costs and makespan simultaneously. Therefore, no inserted idle time was allowed. ISETP was used to assign the jobs on parallel machines and PNBC-NBE heuristic was then applied to acquire the optimal set of jobs compression and expansion processing time in a given sequence. Scheduling problem on unrelated parallel machines with sequence dependent setup time was studied by Zeidi and Mohammad Hosseini [8]. An integrated meta-heuristic algorithm consisting of genetic algorithm and simulated annealing method was proposed to solve the problem in which simulated annealing method was used as a local search to improve the quality of solutions. Alvarez-Valdes *et al*. [9] proposed hybrid heuristic algorithm combining priority rules for assigning jobs to machine and local search for solving the one-machine subproblem. Path relinking and scatter search were also applied to obtain high quality of solutions. Hung *et al.* [10] addressed the scheduling jobs with time windows on unrelated parallel machines with sequence dependent setup time. Machine and job dependent processing times were also considered. Three solution methods based on mixed integer programming (MIP) called HCMIP, ETMIP and HIMIP were proposed for solving the problem. Wu *et al.* [11] investigated unrelated parallel machine scheduling with consideration of job rejection and earliness-tardiness penalties. Hybrid algorithm combining genetic algorithm and tabu search were presented to solve the problem.

In the literature, the exact method and heuristic algorithm were proposed for solving scheduling problem. This study is associated with the job assignment to parallel machines and determination of starting time for each job in a given sequence. Population-based incremental learning (PBIL) algorithm explored by Baluja [12] is used to create feasible solution indicating a sequence of jobs on each machine. PBIL is an evaluation algorithm combining the mechanism of genetic algorithm and competitive learning. Probability vector is applied to define a population representing solution of problem. To obtain high quality of solution, high values of probability in probability vector are updated based on the best solution at each iteration. Once a job sequence is given, optimal timing algorithm is applied to decide optimal starting time for each job in the sequence as one-machine subproblem.

The remainder of this paper is organized as follows. Section 2 describes the problem description and some notations used throughout in this paper. Three algorithms for assigning and sequencing jobs to parallel machines, dispatching rule based on earliest due date, optimal timing and population-based incremental learning algorithms are also explained in this section. Section 3 presents the experiment results obtained from the proposed algorithm. Finally, the conclusions are provided in section 4.

2. Materials and Methods

2.1 Problem description

Given n jobs and m parallel machines, the job i for $i=1,2,...,n$ can be processed on any machines k for $k=1,2,...,m$ without interruption. A machine can perform only one job at a time and the processing time of job *i*, denoted by p_i, is identical for all machines. Let d_i and c_i be due date and completion time of job *i*, respectively. Job *i* is said to be early if $c_i < d_i$ and the earliness cost is $\alpha_i(d_i - c_i)$ where α_i is earliness weight while it is said to be tardy if $c_i > d_i$ and the tardiness cost is $\beta_i(c_i - d_i)$ where β_i is tardiness weight. Job *i* is on time if $c_i = d_i$. All jobs are available to process at time zero and can be started immediately after the predecessor job is completed. The inserted idle time is allowed to determine the starting time of each job in which the completion time is close to its due date. The cost of job *i* processed completely before or after its due date is represented by equation (1).

$$f(c_i) = \alpha_i E_i + \beta_i T_i \tag{1}$$

Where $E_i = \max\{0, d_i - c_i\}$ and $T_i = \max\{0, c_i - d_i\}$. The objective of the study is to determine the job schedule that minimize the total cost calculated by the following equation.

$$\text{Total cost} = \sum_{i=1}^{n} f(c_i) \tag{2}$$

2.2 Dispatching rule

Definition 2.1 A set of jobs $\{J_1, J_2, ..., J_n\}$ is in the order of earliest due date (EDD) if the jobs are sequenced according to the non-decreasing due date, i.e., $d_{J_1} \le d_{J_2} \le ... \le d_{J_n}$.

In order to obtain the job sequence on each machine, the jobs in EDD order are assigned to the machines depending on total completion time. The steps for the job assignment called Earliest Due Date for Parallel Machines (EDDPM) are as follows.

Algorithm 1: EDDPM

Step 1: Let $S = \{J_1, J_2, ..., J_n\}$ be a set of jobs sequenced by EDD rule and $S' = \varnothing$ be a set of assigned jobs.

Step 2: Let $TC_k = 0$ be the total completion time on machine k, $k = 1, 2, ..., m$. Assign the first job J_1 to the first machine and update $TC_1 = p_{J_1}$. Put J_1 in S to S', $S' = S' \cup \{J_1\}$ and $S = S - \{J_1\}$

Step 3: Assign the next job J_i in S to all machines and calculate temporary $TC_k^{'} = TC_k + p_{J_i}$. Job J_i is processed on a machine giving minimum $TC_k^{'}$, in the case that $TC_k^{'} = TC_p^{'}$, where $k \ne p$, a machine with minimum index is chosen. Update $TC_k = TC_k^{'}$, $S' = S' \cup \{J_i\}$ and $S = S - \{J_i\}$.

Step 4: Continue step 3 until $S' = \{J_1, J_2, ..., J_n\}$ and $S = \varnothing$.

To present the job sequence and penalty cost resulting from applying EDDPM, the input data for 10 jobs given in Table 1 are used for the calculation and the results are shown in example 2.1.

Table 1. Input data for 10 jobs

Job	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
p_i (day)	6	5	1	4	3	10	3	3	4	4
d_i (day)	11	10	12	15	15	8	13	14	14	10
α_i ($/day)	1.9	1.8	1.8	1.1	2.3	1.4	1.9	1.2	1.8	0.9
β_i ($/day)	3.8	3.7	3.4	1.4	2.7	1	0.9	0.8	1.8	3.3

Example 2.1 Consider the input data given in Table 1. A set of job ordered by EDD rule is $S=\{6,2,10,1,3,7,8,9,4,5\}$. By applying EDDPM, job sequence for each machine is shown in Figure 1.

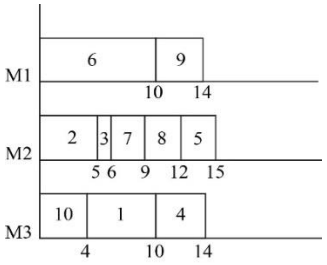


Figure 1. Job sequences for each machine resulting from applying EDDPM algorithm

It can be seen from Figure 1 that job 6 is completed after the due date for two days, i.e. $E_6=\max\{0,d_6-c_6\}=0$ and $T_6=\max\{0,c_6-d_6\}=2$ while job 9 is completed on its due date, i.e. $E_9=\max\{0,d_9-c_9\}=0$ and $T_9=\max\{0,c_9-d_9\}=0$. The costs of job 6 and job 9 calculated by equation (1) are as follows:

$$f(c_6)=(1.4)(0)+(1)(2)=2$$
$$f(c_9)=(1.8)(0)+(1.8)(0)=0$$

Therefore, the penalty cost of machine 1 calculated by equation (2) is \$2. For machine 2, jobs 2, 3, 7 and 8 are finished before the due date for 5, 6, 4 and 2 days, respectively, and job 5 is finished on its due date. Then,

$$E_2=\max\{0,d_2-c_2\}=5,\ T_2=\max\{0,c_2-d_2\}=0,$$
$$E_3=\max\{0,d_3-c_3\}=6,\ T_3=\max\{0,c_3-d_3\}=0,$$
$$E_7=\max\{0,d_7-c_7\}=4,\ T_7=\max\{0,c_7-d_7\}=0,$$
$$E_8=\max\{0,d_8-c_8\}=2,\ T_8=\max\{0,c_8-d_8\}=0,$$
$$E_5=\max\{0,d_5-c_5\}=0,\ T_5=\max\{0,c_5-d_5\}=0,$$

and the costs for all jobs are as follows:

$$f(c_2) = (1.8)(5) + (3.7)(0) = 9$$
$$f(c_3) = (1.8)(6) + (3.4)(0) = 10.8$$
$$f(c_7) = (1.9)(4) + (0.9)(0) = 7.6$$
$$f(c_8) = (1.2)(2) + (0.8)(0) = 2.4$$
$$f(c_5) = (2.3)(0) + (2.7)(0) = 0$$

Thus, the penalty cost of machine 2 is \$29.8. Similarly, the penalty cost of machine 3 is \$8.4 and the total cost for all machines is \$40.2.

2.3 Optimal timing algorithm

Optimal timing algorithm is an algorithm used to decide the optimal starting time of each job. Lee and Choi [1] proposed an optimal timing algorithm to determine the starting time of the jobs in a given sequence without total cost evaluation. The minimum block cost function is calculated to determine the extreme point which decides the starting time of the first job in the block. The extreme point is a point where the slope begins to be greater than or equal to zero. Due to the earliness and tardiness penalties, idle time can be inserted between blocks to shift the entire block toward the minimum point. In order to apply the optimal algorithm proposed by Lee and Choi [1] to this study, the steps of the algorithm can be concluded as follows.

Algorithm 2: Optimal timing algorithm (OTA)

Step 1: Let $S = \{J_1, J_2, ..., J_n\}$ be a set of jobs sequenced by EDD rule and $S' = \varnothing$ be a set of assigned jobs.

Step 2: Put the first job J_1 in block B_1, $s_{J_1} = \max(0, d_{J_1} - p_{J_1})$ and $c_{J_1} = \max(d_{J_1}, p_{J_1})$, where s_{J_1} and c_{J_1} are starting and completion times of job J_1, respectively. $S = S - \{J_1\}$ and $S' = \{J_1\}$.

Step 3: For any block B_t and job J_i, $i = 2, ..., n$, consider the following three cases.

(1) If $c_{J_{i-1}} + p_{J_i} < d_{J_i}$, then $s_{J_i} = d_{J_i} - p_{J_i}$, $c_{J_i} = d_{J_i}$, $t = t + 1$, $B_t = \{J_i\}$, $S = S - \{J_i\}$ and $S' = S' \cup \{J_i\}$.

(2) If $c_{J_{i-1}} + p_{J_i} = d_{J_i}$, then $s_{J_i} = c_{J_{i-1}}, c_{J_i} = d_{J_i}$, $B_t = B_t \cup \{J_i\}$, $S = S - \{J_i\}$ and $S' = S' \cup \{J_i\}$.

(3) If $c_{J_{i-1}} + p_{J_i} > d_{J_i}$, then $s_{J_i} = c_{J_{i-1}}$, $c_{J_i} = s_{J_i} + p_{J_i}$ and $B_t = B_t \cup \{J_i\}$, $S = S - \{J_i\}$ and $S' = S' \cup \{J_i\}$.

Step 4: Determine the minimum block cost function by comparing the slopes of block which can be obtained from the earliness and tardiness weights and shift the entire block B_t toward the minimum point until one of the following cases occurs.

(1) s_{J_1} in block B_1 is zero.

(2) The minimum point is reached.

(3) s_{J_i} in block B_t equals to $c_{J_{i-1}}$ in block B_{t-1}. In this case, blocks B_t and B_{t-1} are concatenated and the new minimum block cost function has to be calculated.

Step 5: Continue steps 3 and 4 until $S' = \{J_1, J_2, ..., J_n\}$ and $S = \varnothing$.

Example 2.2 To determine the optimal starting time of the jobs on each machine as shown in Figure 1 with the given input data in Table 1, algorithm 2 is utilized. In this example, only block cost function of machine 2 is calculated.

Job 2:

Step 1: Let $S = \{2,3,7,8,5\}$ be the job sequence of machine 2 and $S' = \varnothing$.

Step 2: Set $B_1 = \{2\}$, $s_2 = 5$, $c_2 = 10$, $S = \{3,7,8,5\}$ and $S' = \{2\}$. In this case, job 2 has no penalty.

Job 3:

Step 3: Since $c_2 + p_3 = 11 < d_3$, $s_3 = 11$, $c_3 = 12$, $B_2 = \{3\}$, $S = \{7,8,5\}$ and $S' = \{2,3\}$. Job 3 has no penalty.

Job 7:

Step 3: Since $c_3 + p_7 = 15 > d_7$, $s_7 = 12$, $c_7 = 15$, $B_2 = \{3,7\}$, $S = \{8,5\}$ and $S' = \{2,3,7\}$.

Step 4: Compare the slope of block B_2, $-\alpha_3 - \alpha_7 = -3.7$, $-\alpha_3 + \beta_7 = -0.9$ and $\beta_3 + \beta_7 = 4.3$. Thus, job 3 is still completed on its due date and block B_2 is not shifted.

Job 8:

Step 3: Since $c_7 + p_8 = 18 > d_8$, $B_2 = \{3,7,8\}$, $s_8 = 15$, $c_8 = 18$, $S = \{5\}$ and $S' = \{2,3,7,8\}$.

Step 4: Compare the slope of block B_2 as follows.

$$-\alpha_3 - \alpha_7 - \alpha_8 = -4.9$$
$$-\alpha_3 - \alpha_7 + \beta_8 = -2.9$$
$$-\alpha_3 + \beta_7 + \beta_8 = -0.1$$
$$\beta_3 + \beta_7 + \beta_8 = 5.1$$

Job 3 is finished on its due date and block B_2 is placed at the present position.

Job 5:

Step 3 : Since $c_8 + p_5 = 21 > d_5$, $B_2 = \{3,7,8,5\}$, $s_5 = 18$, $c_8 = 21$, $S = \varnothing$ and $S' = \{2,3,7,8,5\}$.

Step 4: By comparing the slope of block B_2, job 7 is finished on its due date and it has to be shifted for 2 time units left. Because $c_7 - d_7 > s_3 - c_2$, block B_2 can be shifted only one time unit. Now, block B_1 and B_2 are concatenated such that $s_3 = 10$, $c_3 = 11$, $s_7 = 11$, $c_7 = 14$, $s_8 = 14$, $c_8 = 17$, $s_5 = 17$, $c_5 = 20$ and $B_1 = \{2,3,7,8,5\}$. Due to the concatenation of block, the new minimum block cost function of B_1 is calculated and it is found that job 7 is finished on its due date. Thus, block B_1 is shifted for one time unit left. Now, $S = \varnothing$ and $S' = \{2,3,7,8,5\}$, therefore, the algorithm is terminated.

The optimal starting time of jobs on each machine is shown in Figure 2. The earliness and tardiness costs of machine 2 is reduced to \$17.8 and the total cost for all machines is \$24.3.

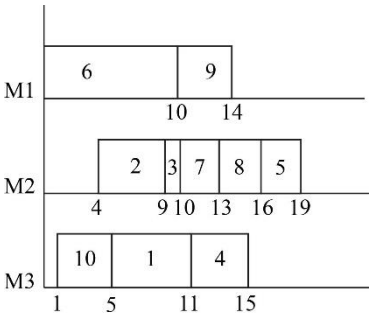


Figure 2. Optimal starting time for each job after applying algorithm 2

2.4 Population-based incremental learning algorithm

Population-based incremental learning (PBIL) algorithm is an evaluation algorithm using a probability vector to describe the population representing the solution of the problem. Each position of solution is represented by either 0 or 1 which can be obtained by the probability vector. Suppose that probability generating 1 for *k*-th position is 0.6, the probability generating 0 for *k*-th position is 0.4 resulting from subtracting 0.6 from 1. The following definition describes the job assignment to the machines used in this study.

Definition 2.2 Let $A=[a_{ki}]$ be an $m\times n$ matrix such that $a_{ki}\in\{0,1\}$. A is satisfied by the properties that $\sum_{k=1}^{m} a_{ki}=1$ and $\sum_{k=1}^{m}\sum_{i=1}^{n} a_{ki}=n$. If $a_{ki}=1$, then machine k operates job i and A is called population.

Example 2.3 Given population matrix A as follow.

$$A=\begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Population matrix A is 3×10 matrix and it indicates that there are 3 machines and 10 jobs such that machine 1 operates jobs 6 and 9, machine 2 operates jobs 2, 3, 5, 7 and 8 and the last machine operates the remaining jobs satisfying job assignment as shown in Figure 1. It can be seen that population matrix A only specifies the jobs which will be processed by a machine. Thus, EDD rule is used to determine the sequence of jobs.

The procedure of PBIL combined with optimal timing algorithm called PBILOTA applied to scheduling problem to minimize earliness and tardiness costs is as follows.

Algorithm 3: PBILOTA

Step 1: Let $1\times n$ matrix P be the initial probability vector.
Step 2: Create a set of feasible solution called population according to probability vector P representing the job assignment on machines.
Step 3: Sort the jobs by EDD rule to determine the job sequence on each machine. The optimal timing algorithm is applied to find the starting time of jobs in each block and total cost is then

calculated for all solutions in step 2 based on eq. (2). The best solution is a population giving the minimum total cost.

Step 4: Update probability vector P by using the following equation.

$$prob_i^{(t)} = prob_i^{(t-1)}\left(1-LR\right)+Best_i^{(t-1)}\left(LR\right) \tag{3}$$

For $i=1,2,...,n$ and $Best_i^{(t-1)}$ is bit integer (0 or 1) of the best solution at iteration $t-1$ and LR is learning rate.

Step 5: Continue step 2 to step 4 until stopping criterion is satisfied. The maximum number of iteration, 500 iterations, is set to be stopping criteria. When the algorithm is terminated, probability values in probability vector are approached to either 0 or 1.

Example 2.4 In order to reduce earliness and tardiness costs by using PBILOTA, the initial probability vector is set based on the solution obtained from EDDPM combined with optimal timing algorithm in example 2.2.

After applying PBILOTA with the input data given in table 1 to determine optimal starting time for each job, the total penalty cost is reduced from \$24.3 to \$12.3 and job schedule for all machines is shown in Figure 3.

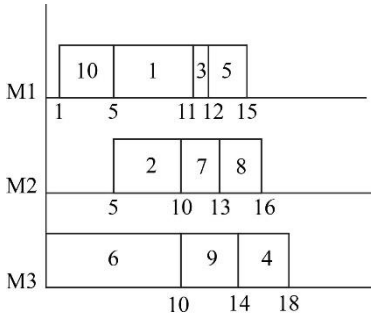


Figure 3. Job schedule obtained from applying PBILOTA

3. Results and Discussion

3.1 Validation

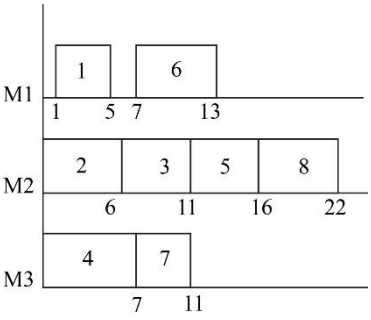
To validate and indicate that the proposed PBILOTA can reduce total cost, the job information presented in Kayvanfar *et al.* [7] is employed to test the algorithms. The information is shown in Table 2 consisting of processing time (day), due date (day), earliness and tardiness weights (\$/day) for each job. This information is used to illustrate ISETP performance for 8 job with 3 machines proposed by Kayvanfar *et al.* [7]. The results obtained from using ISETP, EDDPM and PBILOTA with the given information are shown in Figure 4.

Figure 4 shows the jobs schedule and total cost obtained from applying ISETP, EDDPM and PBILOTA. ISETP is a procedure proposed by Kayvanfar *et al.* [7] for assigning the jobs on parallel machines to minimize earliness-tardiness penalties and makespan simultaneously, thus no inserted idle time is allowed. To satisfy the objective of this study, optimal timing algorithm is utilized to find the starting time of each job to reduce the cost for comparison. As seen in Figure 4(a), the cost of ISETP after applying optimal timing algorithm is \$3.5. The solution of EDDPM shown in Figure 4(b) is set to be one of the populations in PBILOTA such that the cost is set to be

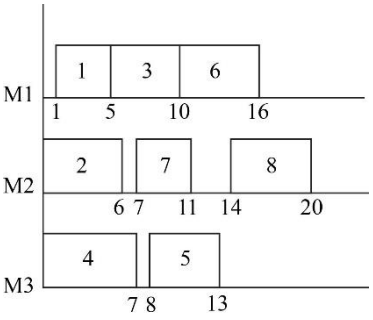
upper bound. Although the cost of EDDPM is more than the cost of ISETP, it is decreased from \$5 to \$2.5 after PBILOTA is performed as shown in Figure 4(c).

Table 2. Job information presented in Kayvanfar *et al.* [7].

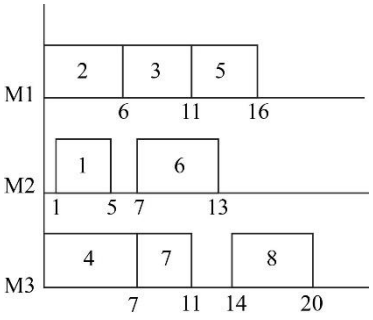
Job	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
p_i	4	6	5	7	5	6	4	6
d_i	5	5	11	6	13	13	11	20
α_i	0.5	1	1	1.25	1.5	1	1.5	0.5
β_i	0.5	0.5	1.25	0.5	0.5	1	3	0.5



(a) ISEPT: cost \$3.5



(b) EDDPM: cost \$5



(c) PBILOTA: cost \$2.5

Figure 4. The results obtained from (a) ISETP, (b) EDDPM and (c) PBILOTA with job information in Table 2

3.2 Parameter Setting

To illustrate the performance of the proposed method, numerical examples are generated. In order to generate the parameters used for the random problems, the procedures found in the literatures are applied. The proposed algorithm is tested on the problem instances consisting of 20, 40, 60, 80 and 100 jobs. Due to parallel machines, the number of machines is related to the number of jobs

such that $m = \lfloor n/\gamma + 0.5 \rfloor$, where $\gamma = 4, 5, 6$ and $\lfloor x \rfloor$ is the greatest integer less than x [10]. Processing time of job i is calculated based on the product of based processing time and machine adjusting factor, $b_i \times \delta_i$ [13]. In this study, b_i and δ_i are randomly generated from the uniform distribution with the range of [1,7] and [0.5,1.5], respectively. An integer due date is generated from the uniform distribution $\lfloor (1-T-R/2)P, (1-T+R/2)P \rfloor$ [4] for $P = \left(\sum_{i=1}^{n} p_i\right) \Big/ m$. Tardiness factor T is a parameter indicating the opportunity that a job is tardy and parameter R is due date range. To avoid crash due date in this study, normal due date can be randomly generated by setting $T = 0.2$ and $R = 0.5$. The earliness and tardiness weights are randomly generated from the range of [0.5,2.5] and [0.5,4.5], respectively [7]. For PBILOTA, population size is set to 100 and the initial probability vector is set based on the solution obtained from EDDPM. The learning rate represents the speed of convergence. As the learning rate is increased, the speed of convergence is increased while the search portion is decreased. To increase the chance for finding feasible solution, $LR = 0.05$ is used in this study and the maximum number of iterations, 500 iterations, is set to be the stopping criteria.

3.3 Results

By the parameter setting addressed in previous section, 5 random problems are generated for each level of γ and 5 run times of PBILOTA are performed for each problem. Thus, total of 375 problem combinations are tested. Since no exact numerical results are available, the results obtained from PBILOTA are compared to EDDPM combined with optimal timing algorithm to evaluate the performance.

Table 3 shows the average relative percentage deviation resulting from the comparison of PBILOTA and EDDPM algorithms mentioned in section 2. The first column shows the number of jobs with up to 100 jobs. The number of jobs divided by the different three levels of γ in the second column results in the different number of machines shown in the third column. To indicate that the application of PBILOTA can minimize the earliness and tardiness penalties, the relative percentage deviation (RPD) for all instances is calculated as follow:

$$RPD = \frac{sol_{EDDPM} - sol_{PBILOTA}}{sol_{EDDPM}} \times 100 \qquad (4)$$

where sol_{EDDPM} and $sol_{PBILOTA}$ are the solutions obtained from EDDPM and PBILOTA algorithms, respectively. Because five problems are randomly generated for each level of γ and PBILOTA is applied for five run times in each problem, the average RPD for each level is computed as shown in the last column of Table 3. The average RPD indicates the decrease in total cost compared to EDDPM. For 20 jobs with $\gamma = 4$ and $\gamma = 5$, total cost is reduced more than the other cases because the jobs can normally be allocated to machine resulting in the increase in the chance for shifting the job near to the due date. When the number of machines is decreased by the increase of γ, the average RPD is decreased because the different cost between EDDPM and PBILOTA is lower indicating that EDDPM creates good initial population. As seen from Table 3, the average RPD trends to be decreased except for the case of 60 jobs. In the case that the number of machines is decreased and the number of jobs is increased simultaneously, the number of jobs per machine is increased causing extended block size of job. Therefore, jobs at the beginning and at the end parts of block are completed far from their due dates. However, the better solution can be obtained

by PBILOTA based on the improvement of probability values in probability vector. Overall, PBILOTA can reduce the cost for all instances.

Figure 5 shows the convergence of proposed PBILOTA method for some problems of 20 to 100 jobs with different γ. It can be seen that the solutions are converged to a single point that the probability values in probability vector are approached to either 0 or 1.

Table 3. Average RPD obtained from the comparison of EDDPM and PBIL algorithms.

Jobs	γ	Machines	Average RPD (%)
$n = 20$	4	$m = 5$	42.23
	5	$m = 4$	44.34
	6	$m = 3$	24.73
$n = 40$	4	$m = 10$	29.29
	5	$m = 8$	39.07
	6	$m = 7$	31.81
$n = 60$	4	$m = 15$	27.08
	5	$m = 12$	31.78
	6	$m = 10$	32.11
$n = 80$	4	$m = 20$	33.28
	5	$m = 16$	31.73
	6	$m = 13$	24.90
$n = 100$	4	$m = 25$	28.29
	5	$m = 20$	24.94
	6	$m = 17$	23.85

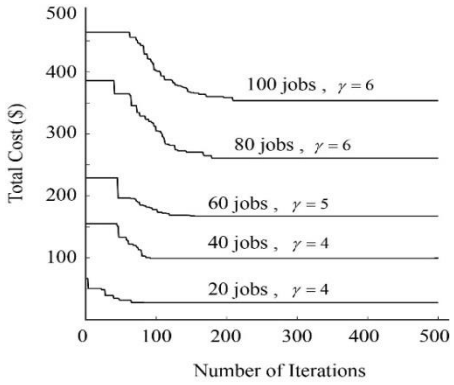


Figure 5. Convergence of solutions obtained from PBILOTA

4. Conclusions

Parallel machine scheduling problem to minimize earliness and tardiness cost is studied. Population-based incremental learning (PBIL) algorithm combined with optimal timing algorithm called PBILOTA is proposed for solving this problem. Computational results demonstrate that the better solution for each iteration can be obtained by the improvement of probability values. In the case that the number of machines is decreased and the number of jobs is increased simultaneously, PBILOTA is still able to find higher quality of solution. The results also show that EDDPM

creates good initial solution resulting in faster improvement of PBILOTA solution. It can be concluded that PBILOTA is an efficient algorithm and suitable for solving parallel machines scheduling problem with earliness and tardiness cost minimization. Developing other dispatching rules to create initial population or integrating heuristic algorithms in the proposed method may be performed for the extension of the study to improve the solution quality. Furthermore, additional conditions such as job sequence dependent setup time and machine dependent processing time are the options for further research.

References

[1] Lee, C.Y. and Choi, J.Y., 1995. A genetic algorithm for job sequencing problems with distinct due dates and general early-tardy penalty weights. *Computers & Operations Research*, 22(8), 857-869.

[2] Bauman, J. and Józefowska, J., 2006. Minimizing the earliness-tardiness costs on a single machine. *Computers & Operations Research*, 33(11), 3219-3230.

[3] Kedad-Sidhoum, S. and Sourd, F., 2010. Fast neighborhood search for the single machine earliness-tardiness scheduling problem. *Computers & Operations Research*, 37(8), 1464-1471.

[4] Kianfar, K. and Moslehi, G., 2012. A branch-and-bound algorithm for single machine scheduling with quadratic earliness and tardiness penalties. *Computers & Operations Research*, 39(12), 2978-2990.

[5] Keshavarz, T., Savelsbergh, M. and Salmasi, N., 2015. A branch-and-bound algorithm for the single machine sequence-dependent group scheduling problem with earliness and tardiness penalties. *Applied Mathematical Modelling*, 39(20), 6410-6424.

[6] Rosa, B.F., Souza, M.J.F., de Souza, S.R., de França Filho, M.F., Ales, Z. and Michelon, P., 2017. Algorithms for job scheduling problems with distinct time windows and general earliness/tardiness penalties. *Computers & Operations Research*, 88, 203-215.

[7] Kayvanfar, V., Komaki, GH.M., Aalaei, A. and Zandieh, M., 2014. Minimizing total tardiness and earliness on unrelated parallel machines with controllable processing time. *Computers & Operations Research*, 41, 31-43.

[8] Zeidi, J.R. and Mohammad Hosseini, S., 2015. Scheduling unrelated parallel machines with sequence-dependent setup times. *International Journal of Advanced Manufacturing Technology*, 81(9-12), 1487-1496.

[9] Alvarez-Valdes, R., Tamarit, J.M. and Villa, F., 2015. Minimizing weighted earliness-tardiness on parallel machines using hybrid metaheuristic. *Computers & Operations Research*, 54, 1-11.

[10] Hung, Y.-F., Bao, J.-S. and Chen, Y.-E., 2017. Minimizing earliness and tardiness costs in scheduling jobs with time windows. *Computers & Industrial Engineering*, 113, 871-890.

[11] Wu, R., Guo, S. and Li, X., 2017, Unrelated parallel machine scheduling with job rejection and earliness-tardiness penalties. *Proceedings of the 36th Chinese Control Conference*, Dalian, China, July 26-28, 2017, 2846-2851.

[12] Baluja, S., 1994. Population-based Incremental Learning: A Method forIintegrating Genetic Search Based Function Optimization and Competitive Learning. [online] Available at: https://www.ri.cmu.edu/pub_files/pub1/baluja_shumeet_1994_2/baluja_shumeet_1994_2.pdf

[13] Joo, C.M. and Kim, B.S., 2015. Hybrid genetic algorithms with dispatching rules for unrelated parallel machines scheduling with setup time and production availability. *Computers & Industrial Engineering*, 85, 102-109.

Healthcare Service Network Analysis: Northern Region's Healthcare Service Network of Cleft Lip and Cleft Palate

Supaksiri Suwiwattana[1,2], Chompoonoot Kasemset[2]* and Krit Khwanngern[3]

[1] Graduate Program in Industrial Engineering, Department of Industrial Engineering, Faculty of Engineering, Chiang Mai University, Chiang Mai, Thailand
[2]Department of Industrial Engineering, Faculty of Engineering, Chiang Mai University, Chiang Mai, Thailand
[3]Department of Surgery, Faculty of Medicine, Chiang Mai University, Chiang Mai, Thailand

Received: 20 January 2020, Revised: 12 February 2020, Accepted: 14 February 2020

Abstract

This research focused on the analysis of healthcare network of cleft lip and cleft palate patients at Chiang Mai University Craniofacial One Stop Service Center as the treatment hub for the northern part of Thailand. The delay was founded during the treatment plan. Patient classification was carried out using K-mean technique. The results from the classification show that the best case patient is the case with the average distance to the treatment center (64.20 km) without complication symptoms. To improve the service of this network, the proposed solution is to increase the number of treatment centers. Two hospitals, Fang and Chomthong hospitals, were proposed as the potential hospitals in this network. The mathematical model was proposed for patient assignment. Twelve districts were assigned to Maharaj Nakorn Chiang Mai hospital and 6 districts were assigned to Fang hospital and Chomthong hospital. The average distances of three groups were 32.84, 55.55 and 62.28 km less than the best case distance at 64.20 km. When the real implementation is considered, both hospitals have to increase their capabilities by hiring more specialist for this service that incurs 2.64 million baht/year for Fang hospital and 2.16 million baht/year for Chomthong hospital. With the proposed solutions, the critical time of the treatment plan can be reduced from 2,663.6 to 2,504.5 days that was 5.97%.

Keywords: Why-Why analysis, PERT technique, K-means technique, mathematical model
DOI 10.14456/cast.2020.9

1. Introduction

Cleft lip and cleft palate are birth facial malformations, that is palate and lip of the baby do not join together properly. Cleft lip and cleft palate results in living, physical, eating, speech and breathing problems and may cause various diseases, for example dental and oral diseases.

*Corresponding author: Tel.: +66 53 94 4125 Fax: +66 53 94 4185
E-mail: chompoonoot.kasemset@cmu.ac.th

However, the effectiveness of the treatment depends on receiving standardized treatment continuously at the proper age of infants by a multidisciplinary team involved in the planning and providing treatment [1, 2].

Chiang Mai University Craniofacial One Stop Service Center or CMU CF center is the treatment center at Maharaj Nakorn Chiang Mai hospital established in 2011 by combining plastic surgeons, neurosurgeons, otolaryngologists, ophthalmologists, anesthesiologists, orthodontists and nurses to establish a multidisciplinary team for the treatment of patients with cleft lip and cleft palate. This center plays a role in providing the treatment service for cleft lip and cleft palate patients since the womb until growing up to allow the patient to have a good quality of life at any age, until normally living and working with other people [1].

This research aims to study the treatment guidelines according to the treatment plans and analyze the patient's route to identify causes of delay along the treatment plan using the data of the 48 patients at CMU CF center as sample data. The highlights of this study are to (1) study the treatment path using CPM/PERT techniques to identify the critical route along the treatment plan, (2) apply one of data mining techniques for classifying patients leading to propose appropriate solutions, and (3) propose solutions using the mathematical model and provide the related aspects, for example investments and resource limitations, when considering the real implementation.

2. Materials and Methods

Program evaluation and review technique (PERT) is a technique adopted by organizations to analyze and present the activity in a project, and to illustrate the flow of events in a project. PERT is a method to evaluate and estimate the time required to complete a task within deadlines [3]. PERT technique was applied widely in many areas, such as in planning the processes of electrical power systems and electrical communication systems in order to reduce time of each activity as addressed in [4]. PERT technique was used to calculate the estimated times of each activity when uncertainty was included.

Data mining is the process of sorting through large data sets to identify patterns and to establish relationships among big data for solving problems. Data mining tools allow enterprises to predict future trends [5]. K-means clustering is a method commonly used to automatically partition a data set into k groups. It proceeds by selecting k initial cluster centers and then iteratively refining them as follows; instance (d_i) is assigned to its closest cluster center and cluster center (C_j) is updated to be the mean of its constituent instances [6]. Data mining was widely applied in healthcare problems. For example, a decision support system for analyzing the risk of chronic disease in case of diabetes and hypertension was presented in Jongkasikit [7] using decision tree and k-nearest neighbor techniques. The proposed model from this research was tested with 10-fold cross validation. The results showed that applying decision trees technique generated the most accurate results.

Facility location problem is to find locations for new facilities such that the conveying cost from facilities to customers is minimized [8]. One application of this problem was presented in Thanakorn and Phongsathon [9]. The distribution centers were located in order to achieve the lowest costs of transportation. The mathematical model of single facility location problem was applied Then the optimal solution was compared with the location selection by center of gravity technique. The results from the mathematical model were superior in term of decreasing shipping distance.

In field of healthcare service management, performance measurements are usually time-base and normally uncertainty. PERT is appropriate to measure treatment time especially activities on a critical paths of treatment plans that can influence completion time of the treatment. Because

of the complication of patients, patient classification is needed to propose different solutions for each group of patients, K-mean technique was applied in this research for classifying patients based on their characteristics. Then the proposed solutions were provided based on the mathematical model of facility location problem for increasing capability of healthcare service of cleft lip and cleft palate treatment service network.

The methodology of this study was as follows:

1) Preliminary Study: The standard treatment protocol for cleft lip and cleft palate was studied on patients aged 0-5 years including activities, time (age) fences, and multidisciplinary team. Patient's route was presented using PERT and CPM techniques to identify the treatment critical path and total treatment time along the standard protocol.

2) Data Collection and Analysis: The data of 48 patients, aged between 3-7 years, without complications, residing in Chiang Mai province, were collected. Based on available data, K-means technique was applied to classify groups of patients. Then problem identification was carried out based on results of K-means.

3) Model Formulation and Optimization: Location problem was applied to this research for locating additional healthcare service centers for the patients residing in Chiang Mai. The proposed model was solved using Lingo optimizer. The optimal solution from the case study was interpreted and compared with the current situation.

4) Result Discussion: The optimal solution was discussed in term of real implementation considering available resources and additional investment for improving this healthcare service network.

3. Results and Discussion

3.1 Cleft lip and cleft palate standard protocol

The standard treatment protocol for patients (0 - 6 months) was presented in Table 1. At each age, specific treatments are needed by multidisciplinary specialists. Any delay occurring within the treatment plan is affected to the treatment effectiveness and time plan.

Table 1. The standard treatment protocol for patients (0 - 6 months)

Age (months)	**Treatment plan of cleft lip and cleft palate**	**Multidisciplinary teams**
0-4	Preliminary diagnosis to patients	Plastic Surgeons
	Providing advice and knowledge in patient care	Officer of CF Center
	Congenital abnormally evaluation	Pediatrics
	Micrognathia and Airway obstruction evaluation	Plastic Surgeons
	Feeding evaluation	Nurse
	Cleft lip nose deformity and Alveolar cleft evaluation	Plastic Surgeons
	Insertion tools for adjusting nose structure (Nasoalveolar molding: NAM)	Dentists
	Inserting Palatal Obturator	Dentists
5	Cheiloplasty or Cleft lip repair	Plastic Surgeons
	Follow Up #1 after Cleft lip repair 7- 14 days	Plastic Surgeons
	Inserting Nasoform	Dentists
6	Follow Up #2 after Cleft lip repair 1 month	Plastic Surgeons

From Table 1, it can be seen that the treatment processes for cleft lip and cleft palate requires multidisciplinary team with expertise in various fields including plastic surgeons, otolaryngologists (ENT), ophthalmologists, radiologists, pediatricians, anesthesiologists and dentists. In addition, patients are required to undergo treatment in every treatment process, which will result in patients continuing to receive treatment until the end of treatment plan. Patients living in Chiang Mai are admitted to CMU CF center. Treatment path analysis of cleft lip and cleft palate patients aged 0-5 years admitted to this center is presented in Figure 1. Critical path and the treatment duration from CPM/PERT technique were summarized as shown in Table 2.

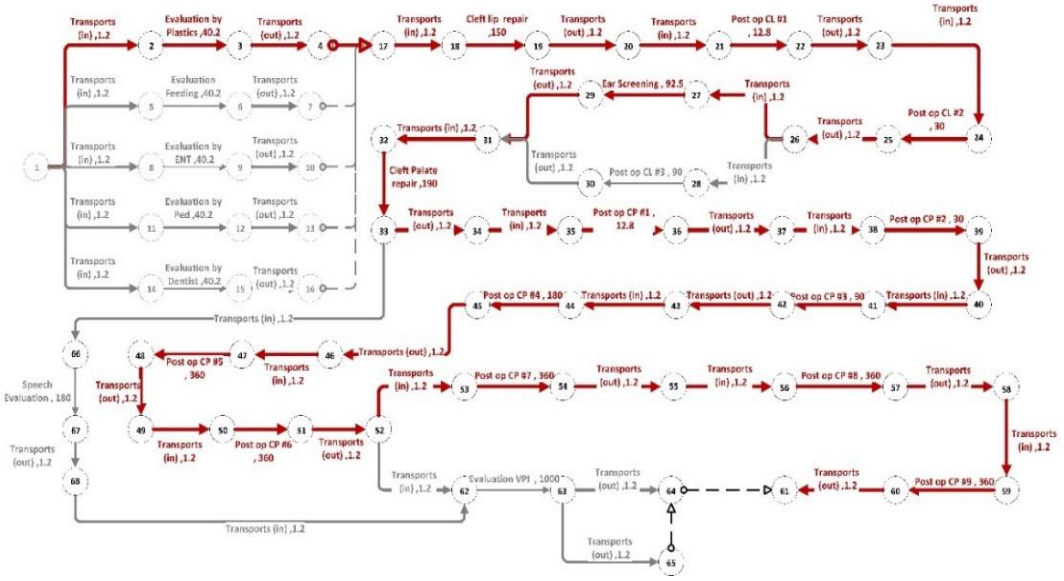


Figure 1. The network identifying the critical path of treatment for cleft lip and cleft palate during aged 0-5 years

Table 2. Treatment path analysis of cleft lip and cleft palate patients aged 0-5 years

Detail	Amount
Critical Time (Days)	2,663.3
The Number of Activities Of The Treatment Processes	22
The Number of Onward Travel Activities	23
The Number of Return Travel Activities	23
The Traveling Expenses (Baht/Person)	30,921.2

From Table 2, patients with cleft lip and cleft palate have a critical treatment duration of 2,663.3 days, with a total of 22 treatment activities, 23 onward travel activities and 23 return travel activities. The traveling expenses is 30,921.2 baht per patient.

3.2 Patient classification using K-means

From data collection, 64.10% of patients with cleft lip and 96.67% of patients with cleft palate were behind the treatment protocol. To identify the cause of delay, data of sample patients with available parameters including the starting age of receiving treatment (days), traveling distance (km) from residential to CF, case's complication (yes/no), and residential area (Chiang Mai/ others), were used in patient classification.

Applying k-means technique via Rapid Miner program, the appropriate number of groups (K) was 5 at the elbow point as shown in Figure 2. Five groups of patients were presented in Table 3.

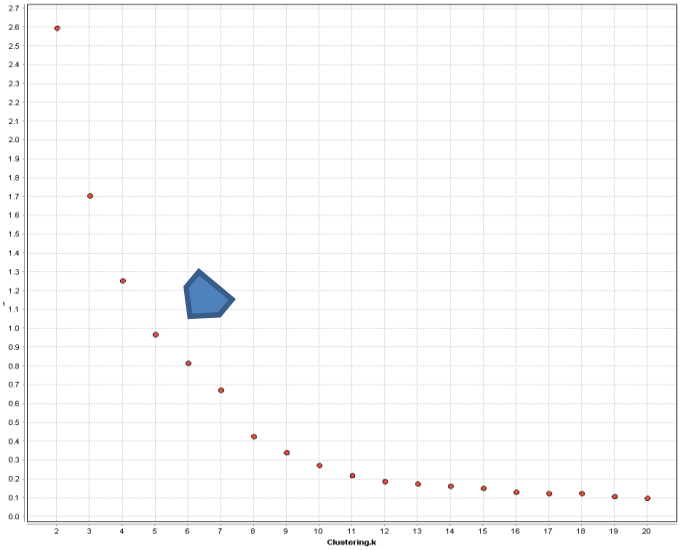


Figure 2. Relationship graph between the number of groups and the average within centroid distance

Table 3. Patient classification of 5 groups

Group	Average age starting to receive treatment (days)	Average travel distance (km.)	Complications	Domiciles
Group 0	237.38	237.62	No	Other provinces
Group 1	21.76	64.204	No	Chiang Mai
Group 2	64.50	163.71	No	Other provinces
Group 3	25.83	110.76	Yes	Chiang Mai and others
Group 4	178.064	136.10	No	Chiang Mai

The first visit for starting the treatment based on the protocol should be within 30 days after childbirth. Thus the 1st and 3rd groups were satisfied with this condition. Comparing these two groups, the third group is the case with complications, so patients need to contact quickly with CF center, while the first group is patients living near CF center. The first group was used as the

best case. From the classification, without complication as group 0, 2, and 4, traveling distance has an effect on delay on the first visit for the treatment. Considering only patients living in Chiang Mai, the 4th group was considered for the improvement.

3.3 The proposed solution

Patients in group 4 live in suburb area of Chiang Mai province, thus the distance between their homes and CF center is rather far. The best case is the group 1 having satisfied starting time of the treatment (within 30 days) with the average distance to CF center as 64.2 km. The location problem was applied to assigned patients at additional possible treatment centers. The mathematical formulation was explained below.

Set:

- i Districts in Chiang Mai province
- j 3 network-centric Hospitals, namely Maharaj Nakorn Chiang Mai Hospital, Fang Hospital, Chomthong Hospital (Proposing by the CF team members considering capability of the hospital in each district)

Parameter:

- D_{ij} Travel distance from district i to network hospital j (km)
- C_{ij} Traveling cost from district i to network hospital j (baht/time)

Decision Variable:

X_{ij} is decision variable used for assigning patients of each district to network-centric hospitals, which is equal to 1 when the patients are sent from District i to District Network Hospital j, and is equal to 0 when no patient is selected.

Mathematical Model:

$$\text{Minimize} \quad \sum_{i=1}^{m} \sum_{j=1}^{n} D_{ij}X_{ij} + 2C_{ij}X_{ij} \qquad (1)$$

Subject to

$$\sum_{j=1}^{n} X_{ij} = 1 \quad \forall i \qquad (2)$$

$$X_{ij} \in \{0,1\} \quad \forall\, i, j \qquad (3)$$

Equation (1) is the objective that combined two objectives together to minimize (i) the total distance and (ii) the traveling cost, considering patients in different districts to the network-centric hospitals. To combine these two objectives, the first objective is assigned to multiply by 1 when the second objective is assigned to multiply by 2 (suggestion form CF team members for prioritizing two objectives). Equation (2) is to assign patients in each district to only one network-centric hospital. Equation (3) is binary constraint for X_{ij}.

In this research, the Lingo program was used to find the optimal solution. The patient assignment was presented in Table 4.

Table 4. Patient assignment to each network-centric hospital

Maharaj Nakorn Chiang Mai Hospital		**Fang Hospital**		**Chom Thong Hospital**	
District	Distance (km)	District	Distance (km)	District	Distance (km)
Mueang	3	Fang	0	Chom Thong	0
Doi Saket	20.1	Mae Ai	30.1	Mae Chaem	78
Mae Taeng	59.8	Phrao	89.7	Hod	67.5
Mae Rim	32	Wiang Haeng	121	Doi Tao	77.8
Samoeng	80.5	Chai Prakan	22.1	Om Koi	132
San Pa Tong	27.1	Chiang Dao	70.1	Doi Lo	18.4
San Kamphaeng	27.8				
San Sai	14.3				
Hang Dong	14.7				
Saraphi	15				
Mae Wang	60.1				
Mae On	39.7				
Average	**32.84**		**55.55**		**62.28**

From Table 4, there were three groups based on each network-centric hospital. Twelve districts were assigned to Maharaj Nakorn Chiang Mai hospital and 6 districts were assigned to Fang hospital and Chom Thong hospital. The average distance of three groups were 32.84, 55.55, and 62.28 km. Comparing with the best case from patient classification, group 1 with average distance as 64.20 km., the optimal solutions were under the best case average distance as shown in Figure 3.

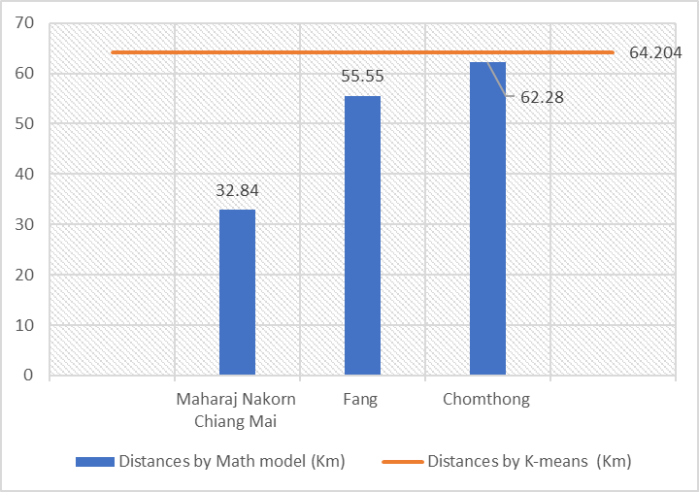


Figure 3. Average traveling distance comparison

Afterwards, treatment path analysis for the optimal solution was carried out to evaluate the duration of treatment as shown in Figure 4.

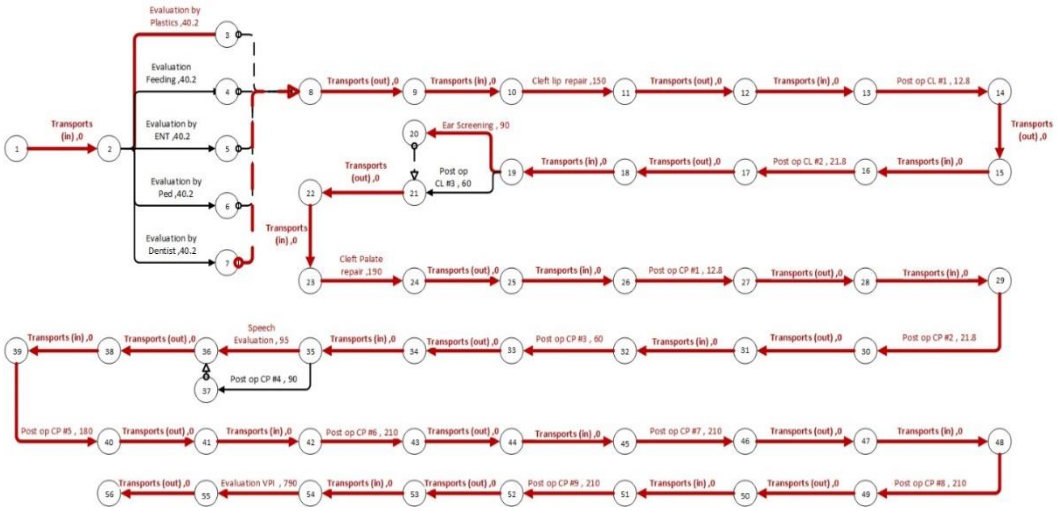


Figure 4. Treatment path analysis for the improvement

Proposing three network-centric hospitals, the treatment protocol was improved by using ECRS technique. The evaluation steps by plastic surgeons, nutritionists, otolaryngologists, pediatrician and dentists are at the same time as 40.2 days. These activities can make appointments at the same day to reduce the number of nodes on activities from 68 to 56 nodes.

The results of patient's path analysis compared between current situation and proposed solution were presented in Table 5.

Table 5. Patient's path analysis results comparison

Parameter	Current	Proposed	Percentage
Critical Time (Days)	2,663.6	2,504.5	↓ 5.97
The Number of Activities of the Treatment Processes	22	22	-
The Number of Onward Travel Activities to Maharaj Nakorn Chiang Mai Hospital	23	0	↓ 100
The Number of Return Travel Activities to Maharaj Nakorn Chiang Mai Hospital	23	0	↓ 100
The Traveling Expenses (Baht/Person)	30,921.2*	0	↓ 100

* From Tawanan [10], the cost of traveling to receive the treatment at Maharaj Nakorn Chiang Mai hospital was 1,344.41 Baht/time.

From Table 5, the critical time of the treatment was reduced from 2,663.3 days to 2,504.5 days or 5.97% without reducing the number of treatment activities. The traveling activities to the Maharaj Nakorn Chiang Mai hospital can be eliminated and the expenses can also be reduced approximately as 30,921.2 baht per person.

3.4 Discussion

The proposed solution is to locate more treatment centers not only Maharaj Nakorn Chiang Mai hospital, but also considering Fang hospital and Chom Thong hospital. To enhance the treatment's capability of both hospitals, some expense is needed. In this study, the cost for hiring specialists was estimated at 2,640,000 Baht/year for Fang hospital and 2,160,000 Baht/year for Chom Thong hospital. Table 6 presented the comparison among cost and benefit of the proposed solution.

From Table 6, total benefits from the proposed solution were approximately as 376,961.60 and 202,979.33 Bath/year for Fang and Chom Thong hospitals, respectively. Comparing with additional cost, there were 14.28% and 9.40% of additional costs for both hospitals. Although monetary benefit is small comparing with the cost, other benefits, for example social and people's well-being, should be considered in term of public service provided by the government.

Table 6. Cost-benefit comparison for the proposed solution

Hospitals	**Cost**	**Benefit (Bath/Year)**	
	Yearly Cost (Millon Baht)	Traveling Expenses	Surgery Income
Fang	2.64	45,773.60	331,188.00
Chom Thong	2.16	24,647.33	178,332.00

4. Conclusions

The cleft lip and cleft palate healthcare service network of the northern region of Thailand has been analyzed with one center at Chiang Mai University Craniofacial One Stop Service Center located at Maharaj Nakorn Chiang Mai hospital. The hospital provides the treatment since patients are newborn until they can live as normal people. Due to the data collection, the delay in the treatment standard protocol is obviously observed. After patient classification using K-mean technique, the results presented that long distance from resident area to the CMU CF center has an effect on the delay of the starting treatment plan causing the delay along the treatment protocol for patients without complication. Thus the solution is to propose additional treatment centers. Two potential hospitals are proposed by specialists as Fang and Chom Thong hospitals. The assignment of patients was carried out based on the proposed mathematical model. The optimal solutions presented that three groups of patients were assigned to each center when the averages of total distance to centers were below the best case of patient group from K-mean technique. The critical path for the treatment was reduced as 5.97%. For the real implementation, due to the limitation of additional hospitals, some cost incurs for enhancing their treatment capabilities.

5. Acknowledgements

The authors would like to acknowledge Faculty of Engineering, Chiang Mai University for supports under Research Assistant scholarship (RA) and the Graduate School of Chiang Mai University for financial support under Teacher Assistant and Research Assistant scholarship (TA/RA) and Research scholarship. The authors would also like to express the thankfulness to CMU CF center for supporting the data and information for this research.

References

[1] WebMD LLC., 2019. *Cleft Lip and Cleft Palate*. [online] Available at: https://www.webmd.com/oral-health/cleft-lip-cleft-palate

[2] Khwanngern, K., 2011. *Treatment Protocol of Lip and/or Palate CMU Craniofacial Center*. (in Thai) [e-book] Faculty of Medicine: Chiang Mai University. Available through: https://w1.med.cmu.ac.th/surgery/

[3] Techopedia Inc., 2020. *Program Evaluation and Review Technique (PERT)*. [online] Available at: https://www.techopedia.com/definition/12112/program-evaluation-and-review-technique-pert

[4] Singsri, D., Worrarat, S., and Phumarj, T., 2017. Planning time reduction of electrical power systems and electrical communication system case study: Prapokklao Hospital. *Dhurakij Pundit University Graduate School Journal* (in Thai), 3, 1055-1069.

[5] Witten, I.H., Frank, E. and Hall, M.A., 2011. *Data Mining: Practical Machine Learning Tools and Techniques.* 3rd ed. San Francisco: Morgan Kaufmann Publications.

[6] Wagstaff, K., Cardie, C., Rogers, S. and Schröedl, S., 2001. Constrained K-means clustering with background knowledge. In *Proceedings of the Eighteenth International Conference on Machine Learning*, San Francisco, CA, USA, 577-584.

[7] Jongkasikit, N.N., 2016. A decision support system for analyzing the risk of two chronic diseases: Diabetes mellitus and hypertension. *Industrial Technology Lampang Rajabhat University Journal*, 9 (2), 11-19.

[8] Liu, B., 2009. *Theory and Practice of Uncertain Programming*. 2nd ed. Berlin: Springer.

[9] Thanakorn, W. and Phongsathon, E., 2018. *The Location Selection of Distribution Center by Applying Mathematical Model and Center of Gravity Method: The Food and Beverage Manufacturing Case study*. B.Eng. Dhurakij Pundit University.

[10] Tawanan, A., 2014. *The Study of the Economics Cost of Cleft Lip and Cleft Palate Patients Aged 0-8 Years Admitted to Maharaj Nakorn Chiang Mai Hospital*. B.Econ. Chiang Mai University.

Sago Palm Genome Size Estimation via Real-Time Quantitative PCR

Hairul Azman Roslan[1]*, Md Anowar Hossain[2], Ngieng Ngui Sing[1] and Ahmad Husaini[1]

[1]Faculty of Resource Science and Technology, Universiti Malaysia Sarawak, Sarawak, Malaysia
[2]Department of Biochemistry and Molecular Biology, University of Rajshahi, Rajshahi, Bangladesh

Received: 9 December 2019, Revised: 23 February 2020, Accepted: 27 February 2020

Abstract

Sago palm, *Metroxylon sagu* Rottb., is an underutilized indigenous food crop that can be found mainly in the South East Asia and Pacific regions. It is a main starch producer and socioeconomically important crop in the South East Asia region including Malaysia. The sago starch provides for considerable potential to food security in the places where it is grown. However, not many molecular works have been reported thus far. In the post genomic era, sago plant genome sequencing is very important for sustainable starch development in these regions. Therefore, determination of the genome size is prerequisite to full genome sequencing and assembly. Here we report on the use of real-time quantitative polymerase chain reaction (qPCR) in determining the genome size. For this work, we calculated the genome size, Γ (bp) of *M. sagu* based on qPCR-derived copy number of two single copy genes. *Pichia pastoris,* with a known genome size, was used as a control to estimate sago palm genome size. With this technique, the genome size of *M. sagu* was calculated to be 1.87 Gbp. This genome size information would be beneficial for subsequent molecular work including genome sequencing and analysis on this economically important crop plant.

Keywords: Genome size, *Metroxylon sagu*, real-time PCR, copy number, *Pichia pastoris*
DOI 10.14456/cast.2020.10

1. Introduction

Sago palm (*Metroxylon sagu* Rottb.) is a palm that is widely distributed in the South East Asia region. It is a starch-producing crop and economically important to the state of Sarawak in Malaysia. Sarawak has the largest acreage planted with sago palm and exports sago starch to countries such as Taiwan, Japan, Singapore and others, generating incomes of up to US$10.8 million/year [1]. Despite its importance, little research in molecular biology has been completed on *M. sagu* molecular biology, and it has not been sequenced, because genome size is not known for this emerging crop. The genome size is a prerequisite in genome sequencing project where it is needed to calculate the number of clones needed to be generated in shotgun sequencing and library

*Corresponding author: Tel.: +60 82 58 3038 Fax: +60 82 58 3160
E-mail: hairulroslan@hotmail.com

screening. The information would also be beneficial to estimate read coverage in next generation sequencing of the genome and its subsequent analyses [2].

An organism genome size is defined as the amount of its DNA and commonly denoted as "C" value or "Γ". The content represents the amount of DNA per haploid genome and is calculated in picograms. Several techniques can be used to estimate genome size such as pulse field gel electrophoresis (PFGE) where the genome was cut into smaller pieces and directly measured [3]. PFGE have been successfully used to determine smaller genomes typically prokaryotes and unicellular eukaryotes [3]. Earlier methods used to estimate genome size are based on either determining the phosphate contents in DNA strands of a defined number of cells or on re-association kinetics of high molecular weight of genomic DNA (C_ot assay) [3]. More recently, D'Hondt and co-workers reported the use of flow and absorption cytometry that determined genome size characterization via DNA-tagged fluorescent dyes [4].

In this work, to estimate the genome size of sago palm, we followed a real-time quantitative polymerase chain reaction (qPCR) technique described by Mounsey and co-workers in 2012 [5] where the amount of genetic materials can be determined based on a known amount of DNA. The real time qPCR combines the nucleic acid amplification and identification in one step and finally quantify the product in a real time by fluorescent detection. This method uses a known DNA mass that was determined by UV spectrophotometry and relatively pure DNA content, that is then used to calculate the C-value by dividing the mass of the DNA sample by the copy number of single copy genes [5]. Using this simple strategy, the genome sizes of various eukaryotic organisms have been estimated such as fungus [6], lichen [7], house-fly [8], fruit fly [9], mites [10], wasp [11], human liver fluke [12], Venus flytrap [13] and three different eukaryotes (yeast, swordtail fish and human) [14]. Here we report on the work to determine the genome size of sago palm using the real-time quantitative PCR technique.

2. Material and Methods

2.1 Collection of plant materials

Fresh young leaf samples (5 months old) of *M. sagu* were collected from the Faculty of Resource Science and Technology garden located at GPS coordinates of 1.469852°N, 110.428175°E. The midribs and stem were removed and then washed with distilled water to remove any debris. The leaves were then surface sterilized using 70% ethanol, blotted dry and cut into small pieces. The processed samples were either used immediately or stored at -80ºC.

2.2 *Metroxylon sagu* DNA isolation and purification

Total genomic DNA isolation was carried out using the CTAB method according to Wee and Roslan [1]. Approximately 0.1 g of the *M. sagu* leaves were ground to powder in liquid nitrogen and added into 1 ml of cetyl trimethylammonium bromide (CTAB) buffer (2% CTAB; 1.4 M NaCl; 20 mM EDTA; 100 mM Tris-HCl, pH 8.0) preheated together with 2 µl of β-mercaptoethanol at 65°C. The mixture was incubated for 1 h at 65°C before gently mixed with 400 µl of 24:1 ratio of chloroform: isoamyl alcohol and centrifuged at 13,000 rpm for 5 min.

The upper aqueous layer containing DNA was transferred to a new tube and treated with 0.2 µl of RNase A (10 µg/µl) (Qiagen, USA) at 37ºC for 1 h, to digest any possible RNA. Genomic DNA was then precipitated overnight at -20°C with equal volume of ice-cold propan-2-ol. The mixture was then centrifuged at 13,000 rpm for 2 min and the resulting pellet was washed

with 1 ml of 70% ethanol, air-dried and re-suspended in 50 µl of 15 mM TE buffer. The DNA samples were stored at -80°C for use in quantification and quality checking.

2.3 DNA isolation from *Pichia pastoris*

Pichia pastoris (strain GS115) was used as the positive control for the method. Three milliliter of *P. pastoris* was cultured overnight in yeast peptone dextrose (YPD) medium. Yeast genomic DNA was isolated using the method described by Mounsey and co-workers [5]. Genomic DNA was extracted by homogenizing *P. pastoris* with a conical grinder in a microcentrifuge frozen with liquid nitrogen. Two milliliter of digestion buffer (800 mM guanidine HCl; 30 mM Tris-Cl, pH 8.0; 30 mM EDTA, pH 8.0; 5% Tween-20; 0.5% Triton X-100) was added to the homogenate containing 2 µl RNase A (10 µg/µl) (Fermentas, Lithuania) and the mixture was incubated at 37°C for 30 min. To further digest all proteins, 100 µl of proteinase K (Fermentas, Lithuania) was added and incubated at 50°C for 1 h. The lysate was then centrifuged at 4000 x g for 10 min at 4°C. The resulting supernatant was transferred into a Genomic-tip 20/G (Qiagen, USA) and DNA extraction was performed according to the manufacturer's instructions. The DNA pellet was then re-suspended in 50 µl of 15 mM Tris-EDTA buffer and stored at -80°C prior to further quantification and quality checking.

2.4 DNA quality control

The concentration and purity of all extracted DNA samples were determined by measuring absorbance at 260 nm using U2900 UV-VIS spectrophotometer (Hitachi, Japan). The DNA purity with A260:A280 ratio of between 1.8 and 1.9 was regarded as having good quality, where one A260 unit correspond to 50 µg/ml DNA.

2.5 PCR primer design and preparation of qPCR standards

Sago palm primers were designed based on sago palm EST sequence available in the National Center for Biotechnology Information database (GenBank). Two sago palm single-copy nuclear genes were selected for this study, the beta-actin (GI: JK731311.1) and elongation factor (EF3) (GI: JK731265.1) genes. For the positive control using *P. pastoris* (genome size 9.43 Mbp), the primers selected were the genes actin [AF216956] and elongation factor 3 (EF3) [FN392322] [5]. Detailed information for the primers used is shown in Table 1. The PCR conditions were optimized to determine that the optimum annealing temperatures were between 55-65°C. The PCR reactions contained 0.2 mM dNTPs, 0.4 µM of each primer, 2.5 mM $MgCl_2$, 0.2 U *Taq* Polymerase, and 1 µl template gDNA. PCR cycle conditions include an initial denaturation at 95°C for 2 min, followed by 35 cycles of a denaturation-step of 94°C for 30 s, annealing-step of 55-65°C for 30 s, and elongation-step of 72°C for 1 min, and a final extension for 10 min at 72°C. The PCR products were visualized on 1.5% agarose gels, purified and ligated into pGEM-T vector (Promega, USA). The PCR products were then sequenced to confirm the genes had been amplified as intended. The plasmid was linearized with *NcoI* restriction enzyme and purified. The DNA was then quantified using NanoDrop 2000/2000c spectrophotometer (Thermo Scientific, USA).

Table 1. Primers used for qPCR in this study.

Gene	Primers	Sequence (5'→3')	PCR Size	Tm
Beta-actin (*P. pastoris*)	PpAct-F	GAT GCG ATC TCC GTT TGA TT	246 bp	60.04
	PpAct-R	GGA AGC GTA CAG GGA CAA AA		60.11
Elongation Factor 3 (*P. pastoris*)	PpEF3-F	CCA TTT GGA CAC TGT CAA CG	246 bp	60.0
	PpEF3-R	CCT GGT TCT GGG AAC TTG AA		60.08
Beta-actin (*M. sagu*)	MsAct-F	GCA CGA TTG AAG GAC CAC TT	185 bp	60.12
	MsAct-R	TGC TGA TCG TAT GAG CAA GG		59.97
Elongation Factor (*M. sagu*)	MsEF-F	CTC TCA CAG CAA AAC GAC CA	216 bp	60.02
	MsEF-R	GTT ATG CCC CTG TGC TTG AT		59.96

2.6 Real-time quantitative PCR

Real-time qPCR was utilized to quantify the amount of a targeted genomic DNA according to modified method described by Mounsey and co-workers [5]. The quantity of amplified product monitored in qPCR was achieved via calculating the fluorescence signal intensities upon integration into DNA strands and are proportional to the amount of amplified PCR. The signal curves that are produced by the standards and target samples, in the same run, were then used to quantify the target DNA.

Real-time qPCR experiment was conducted using Rotor-Gene Q Cycler (Qiagen, USA). A PCR total volume of 25 µl reaction mixture was prepared containing 12.5 µl of 2X SYBR green master mix (Applied Biosystems, USA), 1.0 µl 10 µM of each primer, and 1.0 µl template DNA. The PCR cycle conditions include an initial denaturation-step at 95°C for 3 min followed by 40 cycles of denaturation-step at 95°C for 15 s, annealing-step at 60°C for 20 s and extension-step at 72°C for 30 s. Two reaction runs were conducted with each run consisted of a series of linearized plasmid standards (six, 10-fold dilutions of standard DNA as template), genomic DNA and a no-template control. After the PCR extension step, SYBR Green I fluorescence was calculated and a melting curve analysis was measured. The melting curve confirms the amplification of specific product and was achieved with a temperature gradient of 0.2 Ks^{-1} from 62°C to 95°C.

A DNA weight-to-moles calculator was used to calculate the copy number (Practical Molecular Biology, http://molbiol.edu.ru/eng/scripts/h01_07.html). This calculation determined the number of template concentration and size. To calculate the copy number of targeted gDNA, a standard curve was generated using the C_T values for plasmid amplifications. In order to estimate the genome size, only qPCR runs with standard curve efficiency of more than 90% and duplicates of less than 0.5 C_T standard deviation were used. The C value of a single copy gene of a nuclear genome (picograms) was calculated by dividing the input of the template concentration by the qPCR-derived copy number [5]. The genome size of the unknown gDNA was estimated using the genome size formula described by Mounsey *et al*. [5] and Dolezel *et al*. [15]:

$$\text{Genome size (bp)}, \Gamma = (0.978 \times 10^9) \times C \text{ (pg)}$$

Where, the mean weight of one nucleotide base pair is 1.023×10^{-9} pg.

3. Results and Discussion

The technique used in this work requires the calculation of the absolute amount of single-copy genes present in a DNA sample that is then used to estimate the C values and Γ. To achieve this aim, DNA dilution with a known concentration of a standard (single-copy and gene specific) was required. Two pairs of primers for *M. sagu* single copy genes of beta-actin and translation elongation factor (EF3) were designed and used to amplify the single-copy genes. The PCR products were cloned into a cloning vector, products were confirmed via sequencing and the plasmids were linearized by restriction enzyme *NcoI*. The linearized DNAs were used as standard DNAs in real-time PCR to determine the copy number of genomic DNA present in genomic DNA samples. To validate the protocol, positive control real-time PCR using the beta-actin and elongation factor (EF3) primers was carried out to estimate *P. pastoris* genome size.

An amplification curve generated for elongation factor gene sequence quantification of *P. pastoris* is shown in Figure 1, where fluorescence signals from SYBR Green I were measured at 520 nm. The different color curves are the standard PCR amplified from 10^6 to 10^1 copies of the genes, while the curves for the targeted genomic sample, measured six times, are indicated by arrows. The red curve represents the amplification profile of the no-template control (NTC). A C_T value calibration curve of the standard versus the concentrations in copies per microliter, was plotted using Q-Rex software and used to determine the target copy numbers based on their fractional cycle number or C_T values (Figure 2), where the values are proportional to the log of initial target concentrations. To determine the concentrations of targeted gDNA, a calibration curve of the C_T values of the standard dilution series versus the concentrations was then calculated [5, 14].

Melting curve analysis was done by slowly increasing the PCR amplification temperature from 72° to 90°C. The signal was continuously recorded to ensure that non-specific products were not amplified before the signal curves reached the threshold. The cooperative melting process of the dsDNA causes an abrupt decrease in fluorescence signal that can be seen as a clear peak in the negative derivative (±dF/dT) of the melting curves. The melting curve analysis of the specific PCR product showed that the standard melted at approximate 85°C (Figure 3). The same melting temperature of 85°C was also observed for the PCR product of the targeted genomic DNA, therefore confirming the specificity of the amplification that gave rise to the same PCR product.

The methanotrophic yeast *P. pastoris* strain GS115 was chosen as the positive control for the method due to the availability of its complete genome sequence which was reported to be 9.43 Mbp [16] and 8.7 Mbp using similar technique [5]. Comparatively via real-time qPCR, the genome size of *P. pastoris* was estimated to be 8.52 Mbp (Table 2) which was within 10% of the size reported by De Schutter *et al*. [16] and closer to the size by Mounsey *et al*. [5]. Quantitative real-time PCR of two single copy genes of *M. sagu* gave a mean genome estimate of 1.87 Gbp for *M. sagu* (Table 2) with no large differences in genome size estimated using either of the two specific primers. Assuming the 10% upper error limit for the genome size of *P. pastoris*, the actual genome size of *M. sagu* is therefore estimated to be 2.06 Gbp.

This estimated genome size of *M. sagu* is very much larger than that of any other economically important plant of known genome size such as agarwood (*Aquilaria malaccensis*) which range from 894.65 to 938.88 Mbp [17]. The genome size for the Para rubber tree (*Hevea brasiliensis*) has been estimated to range from 1.37 to 1.47 Gbp [18] and the genome size for oil palm (*Elaeis guineensis*) closely related to *M. sagu* and from the same family Arecaceae was estimated at the 1.8 Gbp [19]. However, genome size across angiosperms has been revealed to be

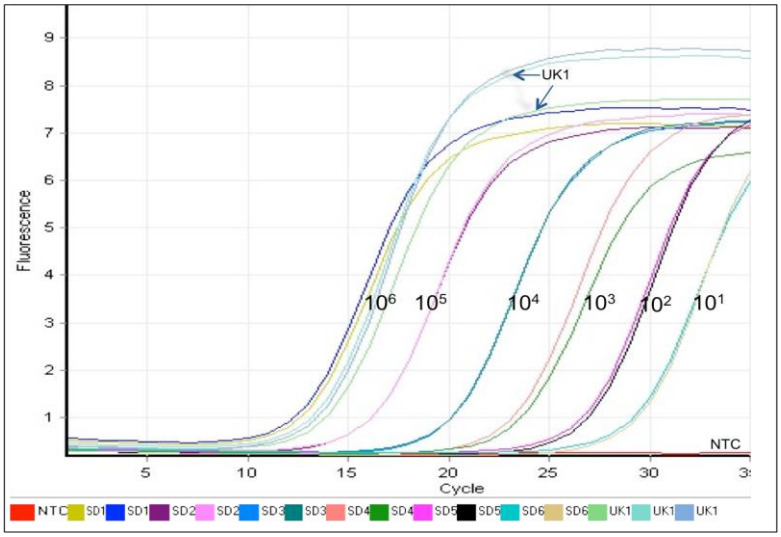


Figure 1. Signal curves obtained from real-time PCR. An amplification curve used for quantification of elongation factor gene from *P. pastoris*. Fluorescence signals from SYBR Green 1 were measured at 520 nm. Curves obtained for the standards with 10^6, 10^5, 10^4, 10^3, 10^2 copies of the standards PCR products as templates are shown in different colors (SD1 to SD6, each duplicate samples). Samples curves are shown by black arrows (UK1, triplicate samples); NTC indicates the no-template control.

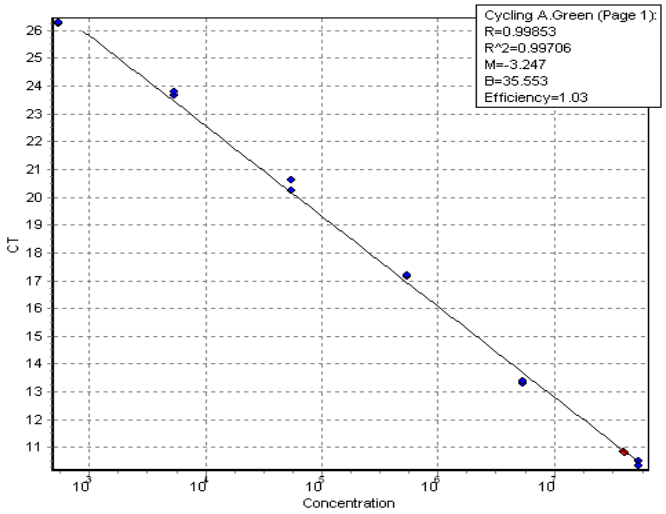


Figure 2. C_T values of serially 10-fold diluted standard of known concentrations (copy number) of elongation factors (EF3) gene sequence and unknown genomic samples of *P. pastoris.* Standard and unknown samples C_T values are indicated by blue and red colors, respectively.

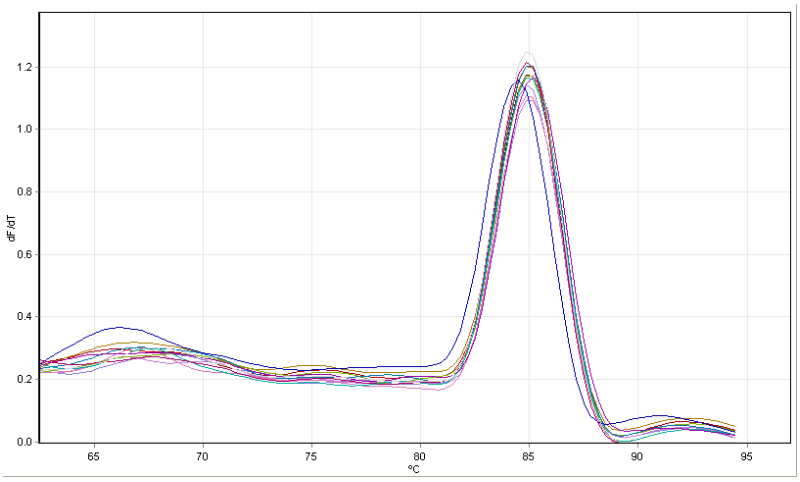


Figure 3. Melting curve analysis of standard and samples after amplification with the PpEF3 specific primers (Amplification curve is shown in Figure 1). Melting curve analysis was carried directly after the PCR by slowly increasing the temperature at 0.2 Ks^{-1} from 72 to 90°C while the signal was taken continuously. Cooperative melting process of the dsDNA causes a steep decrease in the fluorescence signal around the melting temperature of the PCR products. The temperature where the top value is reached is considered as the melting temperature: the specific PCR product obtained from the amplification of the standard PCR product melts at 85°C. Same melting temperature was also determined for the PCR product of the genomic DNA sample.

Table 2. Quantitative real time PCR based genome size estimates of *P. pastoris* and *M. sagu*

Species	**Sample conc.** (ng/µl)	**Gene**	**qPCR products size** (bp)	**Target copies/µl**	**C** (Mean, ±SD, pg)	**Genome Size, Γ** (Mean, ±SD)	**Average Genome size** (Γ)
P. pastoris	134	Beta-actin	246	$15.5x10^6$	0.0086 ± 0.0006	8.45 (± 0.5) Mbp	8.52 (± 0.42) Mbp
	254	Elongation factor 3	246	$28.81x10^6$	0.0088 ± 0.0004	8.62 (± 0.4) Mbp	
M. sagu	151	Beta-actin	185	$7.95x10^4$	1.89 ± 0.11	1.85 ± 0.09 Gbp	1.87 (± 0.08) Gbp
	160	Elongation factor	209	$8.20x10^4$	1.95 ± 0.12	1.90 ± 0.08 Gbp	

diverse with differences in C value ranging up to 2000-fold [20]. The genome size estimated here for *M. sagu* of 1.87 Gbp, is therefore still considered small compared to the monocot *Trillium hagae* (*Melanthiaceae*) that is known to have the largest genome size (129.54 Gbp) among all angiosperms analyzed to date [21]. The relatively small size of *M. sagu* genome nevertheless

allows a fair challenge to its genome sequence project. Although flow cytometry remains the most ideal method for determination of genome size, this method is not available to many investigators, and the PCR method we use here is demonstrated to be quite reliable. Up until now, there has been no reported analyses of *M. sagu* genome size. Therefore, this is the first report for genome size estimation for *M. sagu*, which will be helpful for future *M. sagu* genome sequencing and subsequent molecular analyses projects.

4. Conclusions

Quantitative real time PCR is an invaluable tool that could be used in many applications including in evaluating the genome size of an underutilized crop plant, *Metroxylon sagu*. The qPCR technique indicated that the genome size of *M. sagu* is determined to be 1.87 Gbp. With the determination of genome size of *M. sagu*, this would enable other molecular research of the species, such as genome sequencing, and further drive molecular analysis of this important but underutilized crop plant.

5. Acknowledgements

This study was funded by grants derived from UNIMAS (F07/DPD01/1119/2014(01)) and partially funded by the Malaysia Ministry of Higher Education (FRGS/2/2013/ST03/UNIMAS/02/2). The authors would also like to thank the Faculty of Resource Science and Technology (UNIMAS) for providing the research facilities.

References

[1] Wee, C.C. and Roslan, H.A., 2012. Expressed sequence tags (ESTs) from young leaves of *Metroxylon sagu*. *3 Biotech*, 2, 211-218.

[2] Hare, E.E. and Johnston, J.S., 2011. Genome size determination using flow cytometry of propidium iodide-stained nuclei. *Methods in Molecular Biology*, 772, 3-12.

[3] Zhang, J.Z. and Fan M.Y., 2002. Determination of genome size and restriction fragment length polymorphism of four Chinese rickettsial isolates by pulsed-field gel electrophoresis. *Acta Virologica*, 46, 25-30.

[4] D'Hondt, L., Hofte, M., Van Bockstaele, E. and Leus, L., 2011. Applications of flow cytometry in plant pathology for genome size determination, detection and physiological status. *Molecular Plant Pathology*, 12, 815-828.

[5] Mounsey, K.E., Willis, C., Burgess, S.T.G., Holt, D.C., McCarthy, J. and Fischer, K., 2012. Quantitative PCR-based genome size estimation of the astigmatid mites *Sarcoptes scabiei, Psoroptes ovis* and *Dermatophagoides pteronyssinus. Parasites & Vectors*, 5, 3. https://doi.org/10.1186/1756-3305-5-3.

[6] Henk, D.A. and Fisher, M.C., 2012. The gut fungus *Basidiobolus ranarum* has a large genome and different copy numbers of putatively functionally redundant elongation factor genes. *Plos One*; 7 (2), e31268. https://doi.org/10.1371/journal.pone.0031268

[7] Armaleo, D. and May, S., 2009. Sizing the fungal and algal genomes of the lichen *Cladonia grayi* through quantitative PCR. *Symbiosis*, 49, 43, https://doi.org/10.1007/s13199-009-0012-3

[8] Gao, J. and Scott, J.G., 2006. Use of quantitative real-time polymerase chain reaction to estimate the size of the house-fly *Musca domestica* genome. *Insect Molecular Biology*, 15, 835-837.
[9] Tsoumani, K.T. and Mathiopoulos, K.D., 2012. Genome size estimation with quantitative real-time PCR in two Tephritida species: *Ceratitis capitata* and *Bactrocera oleae*. *Journal of Applied Entomology*, 136, 626-631.
[10] Kim, J.H., Roh, J.Y., Kwon, D.H., Kim, Y.H., Yoon, K.A., Yoo, S., Noh, S.J., Park, J.H., Shin, E.H., Park, M.Y. and Lee, S.H., 2014. Estimation of the genome sizes of the chigger mites *Leptotrombidium pallidum* and *Leptotrombidium scutellare* based on quantitative PCR and k-mer analysis. *Parasites & Vectors*, 7, 279. https://doi.org/10.1186/1756-3305-7-279.
[11] Park, B. and Kim, Y., 2012. Genome size estimation of an endoparasitoid wasp, *Cotesia plutellae*, using quantitative real-time polymerase chain reaction. *Journal of Asia-Pacific Entomology*, 15, 349-353.
[12] Kaewkong, W., Imtawil, K., Maleewong, W., Intapan, P.M., Sri-Aroon, P., Wongkham, S. and Wongkham, C., 2012. Genome size estimation of liver fluke *Opisthorchis viverrini* by real-time polymerase chain reaction based method. *Parasitology International*, 61, 77-80.
[13] Jensen, M.K., Vogt, J.K., Bressendorff, S., Seguin-Orlando, A., Petersen, M., Sicheritz-Pontén, T. and Mundy, J., 2015. Transcriptome and genome size analysis of the Venus flytrap. *Plos One*, 10: e0123887. https://doi.org/10.1371/journal.pone.0123887
[14] Wilhelm, J., Pingoud, A. and Hahn, M., 2003. Real-time PCR-based method for the estimation of genome sizes. *Nucleic Acids Research*, 31: e56. https://doi.org/10.1093/ nar/gng056
[15] Dolezel, J., Bartos, J., Voglmayr, H. and Greilhuber, J., 2003. Nuclear DNA content and genome size of trout and human. *Cytometry. Part A ,* 51, 127-128.
[16] De Schutter, K., Lin, Y-C., Tiels, P., Van Hecke, A., Glinka, S., Weber-Lehmann, J., Rouzé, P., Van de Peer, Y. and Callewaert, N., 2009. Genome sequence of the recombinant protein production host *Pichia pastoris*. *Nature Biotechnology*, 27, 561-566.
[17] Siti Suhaila, A.R., Mohd Saleh, N., Norwati, M., Mahani, M.C., Namasivayam, P. and Kandasamy, K.I., 2018. DNA content and genome size of highly valued malaysian agarwood, *Aquilaria malaccensis* LAMK. *Malaysian Applied Biology*, 47, 13-21.
[18] Tang, C., Yang, M., Fang, Y., Luo, Y., Gao, S., Xiao, X., An, Z., Zhou, B., Zhang, B., Tan, X., Yeang, H.Y., Qin, Y., Yang, J., Lin, Q., Mei, H., Montoro, P., Long, X., Qi, J., Hua, Y., He, Z., Sun, M., Li, W., Zeng, X., Cheng, H., Liu, Y., Yang, J., Tian, W., Zhuang, N., Zeng, R., Li, D., He, P., Li, Z., Zou, Z., Li, S., Li, C., Wang, J., Wei, D., Lai, Z.Q., Luo, W., Yu, J., Hu, S. and Huang, H., 2016. The rubber tree genome reveals new insights into rubber production and species adaptation. *Nature plants*, 2, 16073 https://doi: 10.1038/nplants.2016.73.
[19] Singh, R., Ong-Abdullah, M., Low, E-L., Manaf, M.A., Rosli, R., Nookiah, R., Ooi, L.C., Ooi, S.E., Chan, K.L., Halim, M.A., Azizi, N., Nagappan, J., Bacher, B., Lakey, N., Smith, S.W., He, D., Hogan, M., Budiman, M.A., Lee, E.K., DeSalle, R., Kudrna, D., Goicoechea, J.L., Wing, R.A., Wilson, R.K., Fulton, R.S., Ordway, J.M., Martienssen, R.A., Sambanthamurthi, R., 2013. Oil palm genome sequence reveals divergence of interfertile species in Old and New worlds. *Nature*, 500, 335-339.
[20] Pellicer, J., Fay, M.F. and Leitch, I.J., 2010. The largest eukaryotic genome of them all? *Botanical Journal of the Linnean Society* 164, 10-15.
[21] Zonneveld, B.J.M., 2010. New record holders for maximum genome sizes for monocots and eudicots. *Journal of Botany*, 2010, article ID 527357. https://doi.org/ 10.1155/2010/527357.

Two-Dimensional Cutting Stock Problems with a Modified Column Generation Method

Sirirat Juttijudata* and Phissanu Sudjarittham

Applied Optimization Research Unit, Faculty of Engineering Sriracha,
Kasetsart University, Sriracha Campus, Chonburi, Thailand

Received: 17 January 2020, Revised: 6 February 2020, Accepted: 5 March 2020

Abstract

The two-dimensional cutting stock problems pose mathematical challenges due to the nature of mixed integer linear programming resulting in NP-hard problems. At the same time, the problems are industrially important in manufacturing, logistic and supply chain industries. The ability to solve large-scale two-dimensional cutting stock problems could have a great impact on research community as well as industries. The objective of this work is to develop a framework of solution method for two-dimensional cutting stock problems using a modified column generation method. Two-stage Guillotine cutting patterns are considered. The relationship between the cutting patterns in the first and second stages gives rise to additional constraints that are not previously found in one-dimensional cutting stock problems. As a result, the column generation method was modified to handle these additional constraints. In order to further simplify the problem, LP relaxation is used in conjunction with the column generation. Integer solution can be obtained by rounding of LP solution. The lower bound of the problems may be estimated from the minimization of LP problem; allowing the optimality of solution obtained to be assessed in terms of, for example, the worst performance ratio. With the instance problem studied in this work, the modified column generation method performs well and produces the optimal result that is only 1% less than optimal solution obtained from the exact algorithm, which is the effect of rounding. In terms of speed, the proposed method requires only 1/200 floating point operations compared to the full problem with all feasible solutions (from the instance problem studied here). The proposed method may be further fine-tuned both in terms of rounding techniques with some tweaks in the column generation method in the future. The special structures of the problems should be further exploited for the advantage of the solution methods for large-scale problems.

Keywords: two-dimensional cutting stock problems, Guillotine constraints, column generation
DOI 10.14456/cast.2020.11

*Corresponding author: Tel.: +66 16 46 5636
E-mail: srcsrr@ku.ac.th

1. Introduction

There is a wide variety of industrial applications of the two-dimensional cutting stock problems such as those in wood, glass and paper industries. The goal of a two-dimensional cutting stock problems is to minimize the number of standard-size stock sheets to be cut into pieces of specified sizes. In general, the standard-size stock sheets can be cut into either regular (rectangle, circular, etc.) or irregular shapes depending on the applications. The mathematical model and solution procedures of the cutting-stock problems can also be applied in other problems such as bin-packing, knapsack, vehicle loading, pallet loading and car loading problems. According to Delorme *et al*. [1], the number of articles having the titles related to either "cutting stock problems," "bin-packing problems," or both increased from approximately 30 articles per year before 2000 to 130 articles per year in 2014 in Google Scholar database.

For the two-dimensional rectangular guillotine cutting stock problems considered in this study (further explained in the next section), Gilmore and Gomory in 1965 [2] proposed the solution procedure based on an integer programming model with a cutting pattern (column) generation procedure by solving two one-dimensional knapsack problems. In 1966, Gilmore and Gomory [3] further elaborated the cutting pattern generation procedure based on dynamic programming recursions. Since the cutting-stock problems are NP-hard problems, it becomes more complex to generate all feasible cutting patterns as the size of problem increased, which is typically found in practice. Heuristic and metaheuristic procedures represent more practical approaches, e.g., the partial enumeration heuristics of all feasible patterns by Benati in 1997 [4], the linear programming (LP) relaxation with the rounding procedure by Johnson in 1986 [5], the three-stage-sequential heuristics by Suliman in 2005 [6], genetic algorithms by Hopper and Turton in 1999 [7], and the ant colony optimization by Levine and Ducatelle in 2004 [8].

Even though the heuristic and metaheuristic approaches could alleviate some of our difficulties in terms of the complexity of large-scale cutting stock problems in practice, they are still not an absolute mean to solve such problems. Furthermore, there is no guarantee that the solution obtained will be optimal or at least close to optimal. Because of this gap, the objective of this paper is to develop a solution procedure for large-scale two-dimensional rectangular guillotine cutting stock problems. The method should also provide the lower bound as a gauge to check the optimality of the solution. Consequently, this study proposes the solution procedure based on the LP relaxation i.e. the column generation and the rounding.

2. Materials and Methods

2.1 Two-dimensional (2D) cutting stock problem

2.1.1 Mathematical model

For two-dimensional cutting stock problems, the standard-size stock sheets are cut into pieces of specified sizes subjected to 2 stage guillotine constraint-uninterrupted cuts going from one end to the opposite end of the sheet in two perpendicular directions sequentially as shown in Figure 1 while minimizing the number of standard-size stock sheets used. Gilmore and Gomory [2] developed the mathematical model for n-stage cutting stock problems of two and more dimensions. Their 2 stage two-dimensional cutting stock (with trimming) mathematical model is adopted in this paper. In the first stage cut, the standard size sheet of $W \times L$ will be cut into strips with given cutting patterns shown in Figure 1.

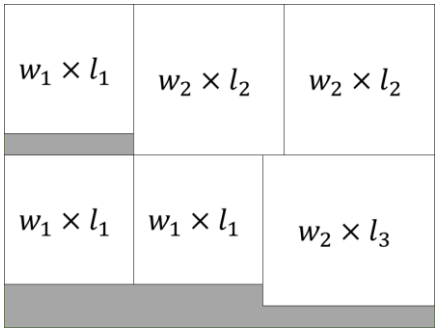


Figure 1. The 2-stage guillotine cutting pattern with trimming wasted material shown in the rendered area. The standard size sheet was $W \times L$ and pieces with sizes $w_i \times l_j$

The number of total strips N_{w_i} with w_i width obtained from the first stage cut will then be:

$$N_{w_i} = \sum_{p=1}^{P} a_{ip} x_p$$

where x_p was the number of standard sheets cut with p pattern at stage 1. The p cutting pattern a_{ip} is defined as the number of strips w_i cut along the width of standard size. Note that i is the index for the widths of strips running from $i = 1, \dots, I$, where I represents the total number of widths of strips. Figure 1 showed the strips of sizes $w_1 = 6$ units and $w_2 = 7$ units, i.e. $I = 2$. If $p = 1$ in Figure 1, the cutting pattern in the first stage for strips w_1 and w_2 is:

$$a_{i1} = \begin{Bmatrix} 0 \\ 2 \end{Bmatrix}.$$

Note that the summation of the strip widths in the cutting pattern must be less than W, i.e.

$$\sum_{i=1}^{I} w_i a_{ip} \leq W, \forall p. \tag{1}$$

In the second stage, the cutting will be performed in the perpendicular direction, i.e. the vertical direction along the w_i strips with q cutting pattern b_{ijq}. Noted that j is the index of piece lengths running from $j = 1, \dots, J$. For w_2 strips in Figure 1, the cutting pattern $q = 1, 2$ for the pieces with l_1, l_2 and l_3, respectively, are:

$$b_{2j1} = \begin{Bmatrix} 1 \\ 2 \\ 0 \end{Bmatrix} \text{ and } b_{2j2} = \begin{Bmatrix} 2 \\ 0 \\ 1 \end{Bmatrix}.$$

Remember that the strip with w_i width can also be trimmed in to narrower pieces as illustrated in Figure 1. Constraint (1) must also be satisfied along the length of the strip w_i :

$$\sum_{j=1}^{J} l_j b_{ijq} \leq L, \forall i, q. \tag{2}$$

As the total number of strip w_i is limited by N_{w_i} from the first stage, the relationship between the number of strips w_i, y_q, and x_p is related via:

$$\sum_{q=1}^{Q} y_q \leq \sum_{p=1}^{P} a_{ip} x_p \,, \forall i. \tag{3}$$

After the cutting process in both stages, the number of pieces $w_i \times l_j$ must be at least equal to the demand of such pieces d_{ij} :

$$\sum_{q=1}^{Q} b_{ijq} y_q \geq d_{ij}, \forall i, j \tag{4}$$

with the objective to minimize the total number of standard-size stock sheet used in the cutting process:

$$\min z = \sum_{p=1}^{P} x_p \,. \tag{5}$$

In summary, the mathematical model for a two-stage cutting stock problem in two-dimensions is:

$$\min z = \sum_{p=1}^{P} x_p \,.$$

$$\textit{Subjected to} \quad \sum_{p=1}^{P} a_{ip} x_p - \sum_{q=1}^{Q} y_q \geq 0, \forall i$$

$$\sum_{q=1}^{Q} b_{ijq} y_q \geq d_{ij}, \forall i, j$$

$$x_p, y_q \geq 0$$

$$x_p, y_q \in I,$$

$$\textit{where} \quad \sum_{i=1}^{I} w_i a_{ip} \leq W, \forall p \text{ and } \sum_{j=1}^{J} l_j b_{ijq} \leq L, \forall i, q.$$

2.1.2 Instance problem

Let us consider the instance problem of 2-stage cutting stock in two-dimension with the standard-size stock sheets of width W and length L of 15 and 20 units, respectively. The demands d_{ij} for pieces with size of $(w, l) = (6,6), (6,7), (6,8), (7,6)$, (7,7) and (7,8) are 30, 15, 5, 5, 15 and 25, respectively. The problem instance is summarized in Table 1.

Table 1. The summary of problem instance

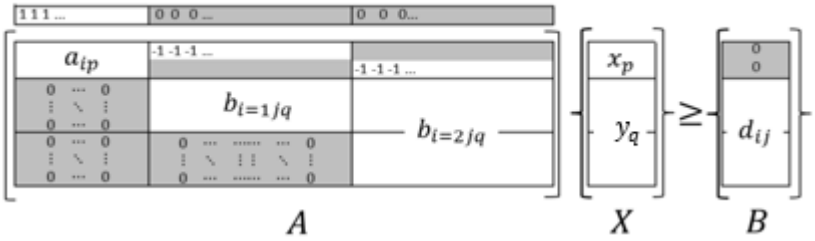
Standard-size stock sheet		
W		15
L		20
Demand size and demand		
The width of the strip	*The length of the piece*	*Demand*
$w_1 = 6$	$l_1 = 6$	30
	$l_2 = 7$	15
	$l_3 = 8$	5
$w_2 = 7$	$l_1 = 6$	5
	$l_2 = 7$	15
	$l_3 = 8$	25

The mathematical model of the problem with all feasible cutting patterns following constraints (1), (2) is formulated in Microsoft Excel 2013 in the matrix form as:

$$\min z = c^T X$$
$$A\,X \leq B$$
$$X \geq 0 \; and \; X \; \in I$$

Where:

$c^T =$



All feasible cutting patterns a_{ip} and b_{ijq} are generated by means of the search tree resulting in the number of cutting patterns for $a_{ip}, P = 5$ and the number of cutting patterns for $q_{1jq} + q_{2jq}, Q = 13 + 49 = 62$. Therefore, the total number of decision variables x_p and y_q, and constraints are 67 and 8, respectively. The optimal solution was determined by Microsoft Excel Solver with Simplex LP, constraint precision of 1×10^{-6}, automatic scaling, and integer optimality of 1%. For the optimal solution, the total number of standard-size stock sheets needed to be cut to meet the demands exactly of 175. The cutting patterns in the first stage a_{ip} and x_p for the optimal solution are:

$$a_{ip} = \begin{bmatrix} 0 & 1 \\ 2 & 1 \end{bmatrix} \text{with } x_p = \begin{Bmatrix} 75 \\ 100 \end{Bmatrix}$$

and the cutting patterns in the second stage b_{ijq} and y_q for the optimal solution are as follows:

Strip $w_1 = 6$:

$$b_{1jq} = \begin{bmatrix} 1 & 2 \\ 2 & 0 \\ 0 & 1 \end{bmatrix} \text{with } (y_q)_1 = \begin{Bmatrix} 50 \\ 50 \end{Bmatrix}$$

and strip $w_2 = 7$:

$$b_{2jq} = \left[\begin{array}{cccc} 0 & 0 & 1 & 2 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ \hline 0 & 1 & 0 & 0 \\ 0 & 2 & 1 & 0 \\ 2 & 0 & 0 & 1 \end{array}\right] \text{with } (y_q)_2 = \begin{Bmatrix} 100 \\ 50 \\ 50 \\ 50 \end{Bmatrix}$$

where the cutting patterns in the upper section of b_{2jq} for strips $w_2 = 7$ corresponds to the cutting patterns for the strips with $w_1 = 6$ while the lower section corresponds to those for the strips with $w_2 = 7$. Remember with the strips of $w_2 = 7$, pieces with the width of either $w_1 = 6$ or $w_2 = 7$ can be cut from them.

2.2 Column generation method

2.2.1 One-dimensional (1D) cutting stock problems

As the size of cutting stock problems increase, the number of all feasible cutting patterns grows rapidly. It becomes impractical, even impossible, to include all the cutting patterns into the problems even it is a one-dimensional cutting stock problem. Instead, the column generation algorithm is used to solve the problem. In column generation algorithm [9], the problem is formulated as a restricted master problem (RMP) with as few decision variables (as well as cutting patterns) as possible, i.e. the model with $P \subset P_{all}$ and $Q \subset Q_{all}$. The new decision variables and cutting patterns are brought into the basis as needed in a similar manner to the simplex method via the following sub-problem:

$$\min RCC = \min\left(1 - \sum_{i=1}^{I} \pi_i a_{ip}\right) = 1 - \max\left(\sum_{i=1}^{I} \pi_i a_{ip}\right)$$

$$\textit{Subjected to} \quad \sum_{i=1}^{I} w_i a_{ip} \leq W, \forall p$$

$$a_{ip} \in I, \forall i, p.$$

RCC and π_i are a reduced cost coefficient and a shadow price from the restricted master problem. This sub-problem is indeed a knapsack problem which has been studied extensively and may be solved efficiently by many algorithms, e.g., dynamic programming. As in the simplex method, the criterion for possible improvement is $\min RCC < 0$ or $\max\left(\sum_{i=1}^{I} \pi_i a_{ip}\right) > 1$ and the corresponding column a_{ip} , i.e. the cutting pattern will be enter the master problem as the basis and at the optimal solution, $\min RCC \geq 0$.

2.2.2 Two-dimensional (2D) cutting stock problems

For two-dimensional cutting stock problems, the Guillotine constraint and relationship between x_p and y_q impose certain constraints on column a_{ip} and $b_{ijq}$of the restricted master problem as seen in the matrix form of the instance problem discussed above. Two modifications are proposed here to handle the two-dimensional cutting stock problems.

First, as the structures of columns of matrix A are having different structures, the sub-problem for each structure shall be formulated individually. Let us take the example of the instance problem. There are three different structures for columns of A in response to the cutting patterns for the first stage, those for the second stage with the strips of w_1 width, and those for the second stage with the strips of w_2; there will be three different formulations for the Knapsack sub-problems. The decision variables for sub-problems are only the cutting patterns a_{ip} and b_{ijq} while constants 0 and -1 in those columns are treated as a constant contribution in sub-problems.

The second modification comes from the fact that there will be three columns generated at each stage rather than just one column, in the case of the instance problem discussed above; thereby three newly generated columns will be added as the bases to the problem concurrently at the end of each iteration.

The solution of master and sub-problem continues until the optimal solution has arrived - either when $RCC \geq 0$ or there are no new independent columns generated.

It should be noted that to get the shadow price for the knapsack problem, the (restricted) master problem has to be relaxed to LP problem - the integer constraint was removed from the problem. As a result, the LP solution obtained may not necessary be integer and required further rounding procedure to round them into integer. Elaborate methods such as branch-and-bound may be used but as the preliminary study of the extension of column generation algorithm for two-dimensional cutting stock problems, simply rounding to the nearest integer is adopted at the end of optimal solution from relaxed LP problem.

The optimal solution of LP problem may also be used as a lower bound to evaluate the worst performance ratio $WCPR = z^*/z_{LP}$ of the solution procedure.

3. Results and Discussion

The modified column generation technique was applied to the instance problem in section 2.1.2. The procedure started from as few decision variables as possible but sufficient to generate feasible solutions. Master and sub-problems were solved sequentially over and over until the optimal solution converged. Among some different initial conditions experimented in this study, the column generation algorithm took less than 10 iterations to arrive at the relaxed optimal solution. The LP optimal solutions for all the cases are 175 coincide well with the exact solution in section 2.1.2. The optimal solutions with cutting patterns, x_p and y_q are:

$$a_{ip} = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} \text{ with } x_p = \begin{Bmatrix} 62.5 \\ 112.5 \end{Bmatrix}$$

$$b_{1jq} = \begin{bmatrix} 1 & 2 \\ 2 & 0 \\ 0 & 1 \end{bmatrix} \text{ with } (y_q)_1 = \begin{Bmatrix} 75 \\ 50 \end{Bmatrix}$$

$$b_{2jq} = \begin{bmatrix} 1 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 2 & 0 & 0 & 0 \\ 0 & 1 & 1 & 2 \end{bmatrix} \text{ with } (y_q)_2 = \begin{Bmatrix} 75 \\ 25 \\ 25 \\ 100 \end{Bmatrix}$$

With the rounding, the (sub) optimal solution from the solution procedure is 176. With the worst performance ratio = 176/175 = 1.006, it is clear that the solution from the procedure is only 0.6% higher than the lower bound, i.e. the LP solution, which is typically lower than the integer programming model; the solution procedure produces satisfactory result. More elaborate rounding procedure is necessary to improve our method and shall be studied in the future.

The column generation method proposed here only adds four more cutting patterns b_{ijq} (columns), i.e. adding four new decision variables to the problem. In total, there are only 12 decision variables to be considered in the column generation procedure. Comparing the number of decision variables to those from the exact LP problem in section 2.1.2, the column generation method only deals with 12 decision variables or only about 20% of the full problem resulting in almost 200 times faster in terms of SIMPLEX floating point operations. Moreover, with the column generation method, it is not necessary to generate all feasible cutting patterns which can be rather challenging for large-scale problem.

Along the course of solution determination, it is observed that at the beginning, the columns corresponding to the cutting patterns in the first stage of cutting process a_{ip} are generated and dominated the whole process. This suggests the method tried to generate stocks of strips with different widths as the resource of the second-stage cutting process. Later on, columns in response to the cutting patterns in the second stage are added. Most of these cutting patterns are for wider strips, w_2, with cutting patterns for pieces with mixture of $w_1$and w_2. This should be the result that the wider strips are more flexible in terms of patterns to be cut; the solution is driven into that direction.

4. Conclusions

The main contribution of this paper is to develop an alternative algorithm to solve large-scale two-cutting stock problems based on the column generation with LP relaxation. The mathematical models of such problems are derived from Gilmore and Gomory [2]. Guillotine constraint and the relationship between x_p and y_q impose certain constraints on columns a_{ip} and b_{ijq}; hence the column generation algorithm has to be modified to cope with the change. The restricted master problems (relaxed LP) and knapsack sub-problems were solved sequentially until the optimal solution are reached. Rounding was applied at the end of the process resulting in slightly less than 1% optimal solutions compared to the full exact solution. The lower bound of the optimal solution may be estimated from z^* of LP. Consequently, the worse-performance ratio can be evaluated and the optimality of the obtained solution can be gauged. It should be noted that in terms of floating point operations, the column generation method requires less 1/200 of those required by exact solution. In other words, the column generation method is 200 times faster than the full exact solution.

There are several aspects that can be done in the future: to refine the column generation method, to improve the rounding technique such as branch-and-bound method, and to exploit the special structures of problems for large-scale problems. Further comparison with other solution procedure shall be carried out to benchmarking the performance of the proposed method.

References

[1] Delorme, M., Iori, M. and Martello, S., 2016. Bin packing and cutting stock problems: mathematical models and exact algorithms, *European Journal of Operational Research*, 225 (1), 1-20.

[2] Gilmore , P.C. and Gomory, R.E., 1965. Multistage cutting stock problems of two and more dimensions, *Operations Research*, 13, 94-120.

[3] Gilmore , P.C. and Gomory, R.E., 1965. The theory and computation of knapsack functions, *Operations Research*, 14, 1045-1074.

[4] Benati, S., 1997. An algorithm for a cutting stock problem on a strip. *Journal of the Operational Research Society*, 39, 288-294.

[5] Johnson, R.E., 1986. Rounding algorithms for cutting stock problems. *Asia-Pacific Journal of Operational Research*, 3, 166-171.

[6] Suliman, S.M. A., 2005. A sequential heuristic procedure for the two-dimensional cutting stock problem. *International Journal of Production Economics*, 99 (1-2), 177-185.

[7] Hopper, E. and Turton, B., 1999. A genetic algorithm for a 2D industrial packing problem. *Computers and Industrial Engineering*, 37 (1-2), 375-378.

[8] Levine, J. and Ducatelle, F., 2004. Ant colony optimization and local search for bin packing and cutting stock problems. *Journal of Operational Research Society*, 55(7), 705-716.

[9] Lasdon, L.S., 1979. *Optimization Theory for Large Systems*. London: MacMillan Publishing.

Dynamic Maintenance Scheduling with Fuzzy Data via Biogeography-based Optimization Algorithm and Its Hybridizations

Pasura Aungkulanon[1]*, Busaba Phruksaphanrat[2] and Pongchanun Luangpaiboon[2]

[1]Department of Materials Handling and Logistics Engineering, Faculty of Engineering, King Mongkut's University of Technology North Bangkok, Bangkok, Thailand
[2]Research Unit in Industrial Statistics and Operational Research, Department of Industrial Engineering, Faculty of Engineering, Thammasat University, Pathumthani, Thailand

Received: 10 January 2020, Revised: 6 February 2020, Accepted: 12 March 2020

Abstract

A multi-objective maintenance problem of a plaza building is presented using a dynamic fuzzy maintenance scheduling model (DFMS). There are multiple component machines and jobs with different fuzzy processing time. Generally, it aims to simultaneously minimize total labor cost on regular, overtime and subcontract including equipment cost and minimize the makespan of all jobs, teams and consecutive time periods under fuzzy natures. Nature-inspired intelligence algorithms have become increasingly popular to implement complex problems. Some features of biogeography-based optimization algorithm (BBO) are unique among biology-based methods. This study applied the BBO and its hybridizations based on the variable neighborhood search (BBOVNS) and particle swarm optimization (BBOPSO) mechanisms to the DFMS. Analytical findings indicated that the proposed BBOPVNS is powerful in terms of dispersion effects. The proposed DFMS demonstrates an efficient compromise method and the overall levels of decision making satisfaction with the multi-objective problem.

Keywords: dynamic maintenance, fuzzy data, metaheuristic, biogeography-based optimization, variable neighborhood search, particle swarm optimization
DOI 10.14456/cast.2020.12

1. Introduction

In this case, this study is a major distributor of consumer goods. In the company's complex plaza, there are many machines and equipment requiring constant maintenance and repairs for full efficiency. Therefore, maintenance was found to reduce the likelihood of machine errors and prevent losses caused by machine and equipment damage.

*Corresponding author: Tel.: +66 25 87 4336 Fax: +66 25 87 4336
E-mail: pasurachacha@hotmail.com

Furthermore, maintaining machines and equipment for functionality based on needs enabled the company to boost work efficiency to meet standards in the quantitative and qualitative aspects with safety and minimum losses. Maintenance need to incur the lowest costs possible.

Thus, the agency needs to make suitable maintenance systems divided into two zones consisting of the office building and the complex. Buildings have many systems such as air conditioning systems, fire extinguishing systems, lighting systems and sanitation systems. Each system is important and has effects on work, safety and customer convenience when customers come to use services.

Therefore, the researcher's interest is studying the aforementioned problem using maintenance data from the case study company as an example. Existing problems in the Maintenance Department consist of differences in expertise and experience among technicians and engineers, repair and maintenance work diversity and differences in priority, urgency and working hours. This has made work schedule organization, work orders and repair and maintenance work assignments complicated with the effect of preventing repairs from being completed on time while incurring high costs. Therefore, the goals of this study were to reduce the number of work delays involved in daily repair and maintenance assignments, and to organize technicians' work schedules by using maximum cost-efficiency within the specified period to improve repair and maintenance efficiency.

Current problem situations in repair and maintenance agencies with work volume as shown in Table 1, which shows the work volume and types of repair and maintenance tasks for the company in one week. Tasks are assigned to four work teams. At present, work is not completed on time due to employees' work delays. The previous data from maintenance unit, overtime and subcontract cost accounted for 30-40%. Furthermore, some teams have received excessive assignments, thereby causing significant overtime costs and causing maintenance times to exceed specifications. In addition, outside contractors are needed to be hired to perform repairs.

A study of the repair work order distribution organization had the following conditions: Repair work did not have fixed characteristics. However, work was assigned twice per day at 7:30 a.m. and at 3:30 p.m. Each repair task has different time requirements for repair standards and different delivery times. Each repair work required one team of technicians responsible. Each repair work task had different score weights and differences in priority. Technicians must receive work assignments to be responsible for full-time work based on daily workloads with eight working hours per day. The company had four teams of technicians in the department who had different working hours.

In this paper, we introduce a recent metaheuristic tool to solve dynamic fuzzy maintenance scheduling optimization problems; the method has been referred as Biogeography-based Optimization Algorithm (BBO) which is inspired by the migration of species between different habitats, and the evolution and extinction of species. The sequential searching procedure of this algorithm mainly relies on the generation of new candidates. The source and number of new candidates can be determined by the users. Since real-world optimization problems are complicated, evolutionary algorithms or metaheuristics are powerful approaches to solve those optimization problems.

There are both single solution and population based algorithms have been developed. The first solution consists of Simulated Annealing (SA) and Tabu Search (TS) algorithms, while Firefly Algorithm (FA), Elephant Herding Optimization (EHO), and Virus Optimization (VO) are applied to multiple solutions with their own evolutionary heuristic tools. The well-performed metaheuristic tools balance among algorithm complexity of exploration and exploitation, computation time, and solution quality including problem complexity and difficulty. With wide applications and their good performance working with high dimensional problems, BBO is then compared with two widely used algorithms.

2. Materials and Methods

2.1 Problem description

Infrastructure systems in the plaza include all equipment and assets that help society function. These infrastructure services like air cool chiller, cooling tower, water pumps of primary and secondary chiller, pumps of transfer, booster, fountain, overflow, drainage, submersible pump, air blower pumps, fire pump and generator, AHU, pressurized, smoke and exhaust fans including load center and control building, are in demand. Engineers and maintenance professionals including subcontracting agencies need to monitor and maintain these existing infrastructures to extending their useful life and ensuring safety in the present as well as in the future. Based on the recent development of call centers reporting service requests, an allocation of maintenance resources properly target these user defined needs. However, with higher demands for services infrastructure with limited budgets, it is vital to maintain these infrastructure jobs over all period. Engineers and others who maintain these systems must develop more flexible ways to meet growing demand while using fewer resources via the fitted mathematical model.

2.2 Proposed model

The case study is conducted in the plaza company in Thailand. The maintenance plan is applied for 12 months. The maintenance workload is over its current capacity. There are 21 workers separated into three teams. Workers can be categorized to S6, S5 and S4 with the skill of worker of 0.5, 0.3 and 0.2, respectively. The summary of skill worker per team is greater than 1.5. There are 30 work groups. The works were assigned twice per day at 7:30 a.m. and at 3:30 p.m. Each repair task had different time requirements for repair standards and different delivery times. Regular working hours are 8 hours per day and overtime are 4 hours per day. Based on the current information, Table 1 shows the frequency (f), minimal (a), standard (m) and maximal (b) fuzzy processing times per unit (T) and number of machines (MC) used in the equipment. There are now nearly 200 items. However, based on the equipment used, they can be categorized into 26 and 4 jobs for main and head quarter buildings, respectively. In each team, regular and subcontract cost per hour are 700 and 1000 baht, respectively. Equipment cost per job is 2000 baht.

A mathematical model of dynamic fuzzy maintenance scheduling (DFMS) is presented in this section. The first step after identifying the problem in a mathematical model formulation is to establish the decision variables. Then an objective function is identified that should satisfy all related constraints on those decision variables (Table 2). The objective function for this model is to minimize total maintenance cost from all N-job (Z1) and to minimize total maintenance time or makespan from all M-team (Z2). The total maintenance cost consists of full-time worker salary, overtime payment and subcontract cost including equipment cost during planning periods in the planning horizon (T). There are some DFMS constraints related to full-time workers, part-time worker, tool or equipment, production, overtime, and subcontracting constraints. The number of full-time workers must be between the minimal and maximal limits. If it is lower than the minimal limit, the maintenance unit cannot proceed. Also, if it exceeds the maximal limit, some workers will be idle [1,2].

Table 1. Maintenance jobs categorized by equipments

Equipment	f	T			MC
		a	m	b	
Air Cool Chiller	2	2	2.5	3	2
Cooling Tower	2	2	2.5	3	4
Primary Chiller Water Pump	2	0.5	0.8	1	4
Secondary Chiller Water Pump	2	0.5	1	1.5	5
Condenser Water Pump	2	0.5	0.8	1	4
Transfer Pump	2	0.5	2	3	4
Booster Pump	2	0.5	1	1.5	2
Fountain Pump	2	0.5	1	1.5	6
Overflow Pump	2	0.5	1	1.5	4
Drainage Pump	2	0.5	1	1.5	10
Submersible Pump	2	0.5	1	1.5	15
Air Blower Pump	2	0.5	1	1.5	10
Fire Pump & Generator	2	0.5	1	1.5	5
AHU Building A	2	1.5	1	2	28
AHU Building B	2	1.5	1	2	36
AHU Building C	2	1.5	1	2	8
FCU Building A	2	1.5	2	2.5	16
FCU Building B	2	1.5	2	2.5	6
FCU Building C	2	1.5	2	2.5	8
FCU Parking	2	1.5	2	2.5	26
Pressurized Fan	2	1.5	2	2.5	6
Smoke Fan	2	1.5	2	2.5	4
Exhaust Fan	2	0.5	1.5	2	6
Load Center and Control Building A	2	1	2	3	9
Load Center and Control Building B	2	1	2	3	4
Load Center and Control Building C	2	1	2	3	2
Case A AHU	2	2	2.5	3	8
Case B PUMP	2	0.5	1	1.5	20
Case C Electrical Other	2	0.5	1	1.5	100
Case D Other	2	0.5	0.7	1	200

Table 2. Variables of the DFMS model

Variable	Symbol
Labor cost per hour for job j in period t	LC_{tj}
Working time per team for job j in period t	W_{tj}
Overtime cost for job j in period t	OC_{tj}
Overtime per team for job j in period t	WOH_{tj}
Subcontract cost per hour for job j in period t	SC_{tj}
Subcontract working time for job j in period t	SH_{tj}
Equipment cost for job j	EC_j
Makespan for team i in period t	MS_{ti}
Amount of overtime in period t	OT_t
Amount of subcontract in period t	SUB_t
Number of full-time workers	W
Skill Score for team i	SS_i

The limit is calculated from the total number of workers, working days, and working hours in each day that the overtime and subcontract can be applied. The total overtime and subcontract man-hours should be lower than the maximum limit. The number of skill levels per team should be higher than the minimal limit; otherwise the maintenance unit cannot function properly. Also, it should be below the maximum limit. The skill scores are the same for all teams. Finally, the total time man-hours should not over the maximum allowable limit. MAX W, MAX OT, MAX SUB, MAX SS and MAX T are the maximal number of workers, overtime, subcontract, skill score and maintenance time in the maintenance unit, respectively. MIN W, MIN SS and MIN T are the minimal levels of workers, skill score, and maintenance time in the maintenance unit, respectively.

$$\text{Min } Z1 = \sum_{t=1}^{T}\sum_{j=1}^{N} LC_{tj} * W_{tj} + \sum_{t=1}^{T}\sum_{j=1}^{N} OC_{tj} * WOH_{tj} + \sum_{t=1}^{T} SC_{tj}SH_{tj} + \sum_{j=1}^{N} EC_j \quad (1)$$

$$\text{Min } Z2 = \text{Max}(\sum_{t=1}^{T} MS_{t1}, \sum_{t=1}^{T} MS_{t2}, \dots, \sum_{t=1}^{T} MS_{tM}) \quad (2)$$

Subject to

$$\sum_{t=1}^{T} MS_{ti} = \sum_{t=1}^{T}\sum_{j=1}^{N/M} W_{tj}; i = 1, 2, \dots, M \quad (3)$$

$$\text{MIN W} \le W \le \text{MAX W} \quad (4)$$

$$OT_t \le \text{MAX OT} \quad (5)$$

$$SUB_t \le \text{MAX SUB} \quad (6)$$

$$\text{MIN SS} \le SS_i \le \text{MAX SS} \quad (7)$$

$$\text{MIN T} \le \sum_{t=1}^{T}\sum_{j=1}^{N} W_{tj} \le \text{MAX T} \quad (8)$$

2.3 Fuzzy natures

2.3.1 Fuzzy inputs

In this paper, each process time is uncertain, so triangular fuzzy number represented by a triplet [a, b, c] is used as presented in Figure 1. The level of membership function is defined as

$$\mu_{\psi_i}(x) = \begin{cases} 0, x < a \text{ or } x > c \\ \frac{x-a}{b-a}, a \le x \le b \\ \frac{c-x}{c-b}, b \le x \le c \end{cases} \quad (9)$$

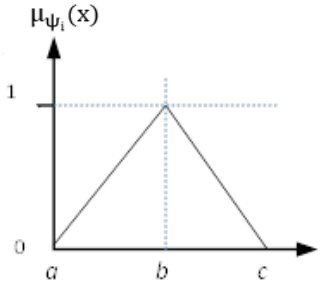


Figure 1. Fuzzy Maintenance Time

Next, the central gravity defuzzification can be applied to transform uncertain data of the fuzzy maintenance time (ψ_i) to crisp equivalent value using the approximated equation [3].

2.3.2 Fuzzy linear programming

In multi-objective functions, several conflicting objectives are considered. The problem is named Multiple Objective Decision Making (MODM) problem. There are many kinds of methods to solve this problem. Fuzzy linear programming (FLP) was applied in this research to make easiness of allowing vague aspirations of the DMs [4-7]. Generally, in MODM model all constraints are restricted as shown in the following

$$\text{MIN } Z_k = c_k^T x \quad (10)$$
$$\text{Subject to } Ax \leq b; x \geq 0$$

In reality, input data are usually imprecise because of incomplete information. In order to solve this kind of problem, the fuzzy membership is then used. A symmetric fuzzy linear programming model can be represented by

$$c_k^T x \lesssim Z$$
$$Ax \leq b; x \geq 0 \quad (11)$$

Here $\lesssim$ represents the fuzzified version of $\leq$ and has the linguistic interpretation *"*essentially smaller than or equal to*"*. It is assumed to be linearly decreasing over the tolerance interval, p_k and k =1, 2, …, K. Then, the membership function can be denoted by

$$\mu_k(x) = \begin{cases} 1 \text{ for } c_k^T x \leq Z_k^{PIS} \\ 1 - \frac{c_k^T x - Z_k^{PIS}}{p_k} \text{ for } Z_k^{NIS} \geq c_k^T x \geq Z_k^{PIS} \\ 0 \text{ for } c_k^T x \geq Z_k^{NIS} \end{cases} \quad (12)$$

; where $p_k = \left|Z_k^{PIS} - Z_k^{NIS}\right|, k = 1, 2, \ldots, K$

Positive-Ideal Solution (PIS) and Negative-Ideal Solution (NIS) are used to construct the membership functions. Then, we arrive at the following problem:

$\text{Max}_{x \geq 0} \text{Min}_k \mu_k(x)$ or Equivalently

$$\text{Max } \lambda \quad (13)$$

Subject to

$\lambda \leq \mu_k(x); k = 1, 2, \ldots, K$

$Ax \leq b; x \geq 0$

$\lambda \in [0, 1]$

3. Biogeography-based Optimization Algorithm and its Hybridizations

Biogeography-based optimization (BBO) is one of the intelligence algorithm that was developed in 2008 by Simon [8-10]. Algorithm was inspired by the migration of species between habitats. Complex behavior via both exploration and exploitation strategies was simplified for purpose of evolutionary algorithm to solve global optimization problems. The characteristic of biogeography is close to natural selection. The species are fitter; they are better able to survive. When they survive longer, they are better to disperse and adapt. The BBO has distinctive characteristics which consist of the inter-habitat distance on migration; nonlinear migration relationships, mortality and reproduction rates on migration, predator/prey relationships, the different mobility measures of difference species on migration, geographical momentum during migration, habitat land area and habitat clusters on migration [11-13].

Complete candidates are firstly generated and each candidate is composed of features, or independent variables. Good and poor candidates correspond to a biological habitat that is well and poorly suited for life, respectively. It is possible for high-fitness candidates to share features with other solutions. The features of emigration and immigration are for high-fitness and low-fitness candidates, respectively. There are two main BBO operators of migration and mutation. The migration operator is to share search information among individuals. The mutation operator is applied to enhance the candidate diversity. The mutation operators of immigration rate and emigration rate of a habitat can be calculated by the linear migration function. More specifically, in the migration function when the number of species increases, fewer species can survive for immigration and more species tend to emigrate to other habitats, and vice visa [14].

Each candidate's migration rate from the deterministic curve is used to stochastically share features. The immigration rate is used to stochastically decide to immigrate the features. If a decision is made in favor of immigration, then a second random decision via a random number that is uniformly distributed between 0 and 1is generated; the emigrating solution is stochastically selected based on emigration rate. Mutation is a probabilistic function that can modify candidate features. An aim is to increase diversity among the population. BBO algorithm is defined as follows. One generation of the BBO algorithm, where N is the population size, Hk is the k^{th} candidate, H is the entire candidates, Hk(SIV) is the feature of Hk, z is a temporal solution, ub and lb are upper and lower bound of the search space, respectively [15-16]. The BBO and its hybridizations based on the variable neighborhood search or VNS [17-18] and particle swarm optimization or PSO algorithms [19-20] are summarized in Figure 2.

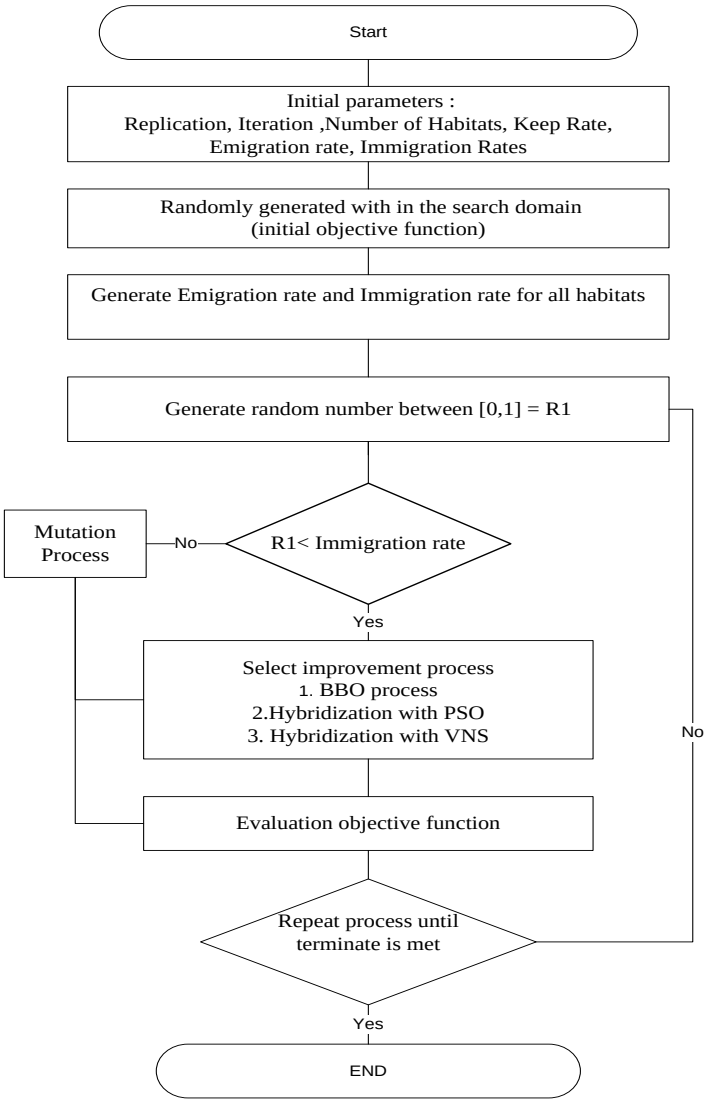


Figure 2. The BBO and its Hybridizations

4. Results and Discussion

In this research, the computational procedures previously described as the fuzzy natures of the dynamic fuzzy maintenance scheduling (DFMS) were performed in a computer simulation implemented in a Visual C#2008 program. The DFMS model presents an exceptional approach to solve the maintenance job scheduling problem with objective to manage breakdown and deterioration due to installation and to sustain performance by preventing unscheduled maintenance and considering uncertainties. The DFMS model was tested with 30 maintenance jobs. Fuzzy natures were used to assess the imprecision in a realistic scenario.

First, a finite interval of maintenance jobs applied the center of gravity defuzzification scheme to generate fuzzy variables and then randomized using the instantaneous probability characteristics of interval time per job. Secondly, the linear and continuous membership function of the objective functions was used to quantify the level of the fuzzy aspiration. The linear membership function was identified according to an analytical definition of membership functions. In BBO, a membership interval is calculated based on all responses in both total maintenance cost and total process time.

The BBO parameters of emigration probability, mutation probability and keep rate are set at 0.2, 0.1 and 0.2, respectively. VNS parameter of the neighbor range (k) equal 5. Both parameters of acceleration factors of c1 and c2 are set at 2. An upper limit on the maximal change of particle velocity (Vmax) is two. The operator balancing the global and the local search (ω) is set at 0.08. Maintenance job interval times were specified by fuzzy numbers and modelled using triangular membership function representations. Fuzzy defuzzification scheme via the center of gravity was used within a finite interval to obtain fuzzy variables. The fuzzy inputs randomized using the instantaneous probability characteristics were used to determine the stochastic measures of the DFMS model input.

Analyses and data visualizations were used to present the feasibility of the model. Minimal total cost and total time scenarios are calculated from all previous data with 100 iterations. A mathematical equation of the proposed model for this DFMS problem can be shown as followed with the corresponding functions of (3)-(8).

Max [λ]

$$\lambda \leq 1 - \left[\frac{Z1 - 733{,}822.31}{118{,}769.19}\right]$$

$$\lambda \leq 1 - \left[\frac{Z2 - 382.31}{68.38}\right]$$

When the performance of the BBO variants of BBO, BBOVNS and BBOPSO was compared, the BBOVNS showed better objectives (Table 3). Based on Z1, the BBOVNS method was the best with the higher level of λ (0.922), lower levels of cost (743,063.02 Baht) and job makespan (400.32 hours). To indicate the stability of the results, each algorithm was repeated 15 times. Moreover, median cost of a BBOVNS is lower than those of BBO and BBOPSO that minimum median value and indicates the stability of the results of this algorithm (Figure 3). The concept is the change of neighborhoods under the specific radius during searching for a better maintennce job scheduling (Table 4). However, there is no statistical significance on the differences of the speed of convergence and the parameter levels on the preliminary results when compared. Moreover, the best so far maintenance job scheduling brings no overtime and sub-contract items.

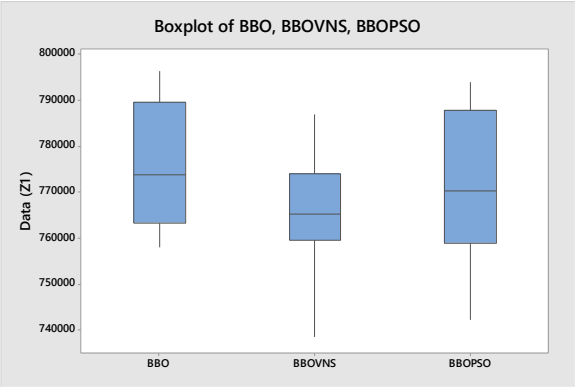


Figure 3. Performance Comparison based on Z1

Table 3. Numerical results of optimal fuzzy value of the objective function

Parameter	BBO	BBOVNS	BBOPSO
Z1 (Baht)	763377.409	743063.02	746322.81
Z2 (Hour)	428.86	400.318	401.531
Max λ	0.751	0.922	0.895

Table 4. Best-so-far maintenance job scheduling via the BBOVNS

T1			T2		
Team			Team		
1	2	3	1	2	3
26	18	5	5	14	10
22	25	21	19	26	7
29	13	19	25	17	4
6	23	28	24	20	23
7	10	30	3	9	1
14	8	15	8	22	28
3	2	27	21	27	12
17	24	4	30	6	29
20	9	16	16	2	18
12	11	1	13	15	11

5. Conclusions

The DFMS model is concerned with the determination of maintenance jobs of a plaza company on a specific time frame. The model aims to reduce the total maintenance cost, a fuzzy maintenance time, and total time. The application of a fuzzy goal programming and fuzzy variable was proposed in this study. The ultimate outcome was the plaza obtains the optimal maintenance plan with the total maximum levels of achieving the goal and uncertainty of job data that can be captured extensively. Recommended future works are exploration of the fuzzy properties of coefficients and related decision parameters for the DFMS problems.

In summary, hybridizations of the BBO can be applied to solve the DFMS problems with fuzzy natures on both variables and objectives. The BBO employs the particle swarm optimization algorithm to develop solution vectors with accuracy and convergence rate of the BBO. The effects of algorithm parameters on the BBO are presented and an approach for tuning these parameters is discussed in this paper. The hybridization via the VNS is an efficient search algorithm which can find admirable solutions when compared to other algorithms. This research provides a benchmarking scenario for managers in maintenance units to minimize cost for scheduling implementation activities. The future research can be the effect of metaheuristic parameters on convergence rate and comparison of this innovative BBOVNS with other metaheuristic methods.

6. Acknowledgements

Authors wishes to thank the Faculty of Engineering, King Mongkut's University of Technology North Bangkok and the Faculty of Engineering, Thammasat University for the financial supports. This research is partly funded by Research Unit in Industrial Statistics and Operational Research, Faculty of Engineering, Thammasat University.

References

[1] Leung, S. and Wu, Y., 2004. A Robust optimization model for Stochastic aggregate production planning. *Production Planning & Control*, 15(5), 502-514.

[2] Aliev, R., Fazlollahi, B., Guirimov, B. and Rashad R.A., 2007. Fuzzy-genetic approach to aggregate production-distribution planning in supply chain management. *Information Sciences*, 177, 4241-4255.

[3] Zimmermann, H.J., 1996. *Fuzzy Set Theory and Its Application*. 3rd ed. Massachusetts: Kluwer Academic Publisher.

[4] Zimmermann, H.J., 1985. Applications of fuzzy sets theory to mathematical programming. *Information Science*, 35, 29-58.

[5] Zimmermann, H.J., 1976. Description and optimisation of fuzzy systems. *International Journal of General Systems*, 2, 209-215.

[6] Zimmermann, H.J., 1978. Fuzzy programming and linear programming with several objective functions. *Fuzzy Sets and Systems*, 1, 45-56.

[7] Aungkulanon, P., Phruksaphanrat B. and Luangpaiboon, P., 2012. Harmony search algorithm with various evolutionary elements for fuzzy aggregate production planning. *Intelligent Control and Innovative Computing*, 189-201.

[8] Simon, D., 2008. Biogeography-based optimization. *IEEE Transactions on Evolutionary Computation*, 12(6), 702-713.

[9] Ma, H. and Simon, D., 2011. Blended biogeography-based optimization for constrained optimization. *Engineering Applications of Artificial Intelligence*, 24(3), 517-525.

[10] Ma, H. and Simon, D., 2017. *Evolutionary Computation with Biogeography-Based Optimization*, John Wiley & Sons, Hoboken, NJ, USA.

[11] Khademi, G., Mohammadi, H. and Simon, D., 2017. Hybrid invasive weed/biogeography-based optimization. *Engineering Applications of Artificial Intelligence*, 64, 213-231.

[12] Ergezer, M., Simon, D. and Du, D., 2009. Oppositional biogeography-based optimization. in *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics*, San Antonio, TX, USA, October 2009, 1009-1014.

[13] Simon, D., Ergezer, M., Du, D.D. and Rarick, R., 2011. Markov models for biogeography-based optimization. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 41(1), 299-306.
[14] Ma, H., Simon, D. and Fei, M., 2016. Statistical mechanics approximation of biogeography-based optimization. *Evolutionary Computation*, 24(3), 427-458.
[15] Ma, H., Simon, D., Siarry, P., Yang, Z. and Fei, M., 2017. Biogeography-based optimization: a 10-year review. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 1(5), 391-407.
[16] Cui, M., Li, L. and Shi, M., 2019. A selective biogeography-based optimizer considering resource allocation for large-scale global optimization. *Computational Intelligence and Neuroscience*, 2019, https://doi.org/10.1155/2019/1240162
[17] Hansen, P. and Mladenovic, N., 2001. Variable neighborhood search: principles and applications. *European Journal of Operational Research*, 130, 449-467.
[18] Hansen, P. and Mladenovic, N., 2001. J-means: A new local search heuristic for minimum sum-of-squares clustering. *Pattern Recognition*, 34, 405-413.
[19] Kennedy, J. and Eberhart, R. 1995. Particle swarm optimization. *Proceedings of IEEE International Conference on Neural Networks*, 4, 1942-1948.
[20] Shi, Y. and Eberhart, R.C., 1999. Empirical study of particle swarm optimization. *In Proceedings of the 1999 Congress of Evolutionary Computation*, 3, 1945-1950.

Efficacy of Acaricides on *Eutetranychus orientalis* (Acari: Tetranychidae) and Its Compatibility with Predatory Mite *Euseius scutalis* (Acarei: Phytoseiidae) under Field Conditions

Sheriehan M. Al-amin[1]*, A.M.A. Ibrahim[2], Ali M. Ali[2], Amira E. Mesbah[1]and N.A. Soliman[1]

[1]Plant Protection Research Institute, Agricultural Research Center, Dokki, Giza, Egypt
[2]Faculty of Science, Assiut University, Assiut, Egypt

Received: 3 October 2019, Revised: 7 February 2020, Accepted: 18 March 2020

Abstract

Efficacy evaluation of seven acaricides, i.e. Acarine (Abamectin 5% EC), Gat Fast (2% Abamectin + 10% Thiamthoxam (12% SC)), Ortis Super (Fenpyroximate 5% EC), Concord (Chlorfenapyr 24% EC), Perfect (2% Abamectin + 10% Chlorfenapyr (12% EW)), Micronet S (Sulfur 80% WP) and Acarots (Fenpyroximate 5% SC) at recommended dose (RD), against the brown spider mite, *Eutetranychus orientalis* (Tetranychidae) and its predatory mite, *Euseius scutalis* (Phytoseiidae), was applied on citrus crop in Assiut Governorate under field conditions. Three assorted exposure eras: three days, one week and two weeks, were achieved in May 2018. It was found that a total reduction rate of these 7 acaricides against *E. orientalis* was 88.26%, 90.40%, 87.99%, 88.91%, 88.78%, 88.41% and 87.82% and against *E. scutalis* was 23.69%, 19.61%, 14.33%, 12.7%, 15.52%, 16.51% and 15.33%, respectively. Abamactin 5% was significantly higher than other acaricides ($p<0.05$) followed by Fenpyroximate 5% EC and Fenpyroximate 5% SC. On the other hand, the rest of acaricides appeared to be insignificant ($p>0.05$). Acaricides can be used against *E. orientalis* without affecting *E. scutalis* where the results showed compatibility between acaricide and predatory mites in the field. For mode of action, Fenpyroximate is safer for human and animal than others because it acts as mitochondrial electron transport inhibitor with contact action. Application of serial concentrations from these compounds is recommended to reduce its toxicity in the environment.

Keywords: efficacy, *E. orientalis*, predatory mites, acaricide, reduction rates
DOI 10.14456/cast.2020.13

*Corresponding author: E-mail: tamersherihan@gmail.com

1. Introduction

The brown spider-mite, *Eutetranychus orientalis* (Tetranychidae) originated in the Middle East and its presence is extended in citric crops of Africa, Southern, Eastern Asia and Australia [1, 2]. The principle host of *E. orientalis* is *Citrus* spp., despite causing it damage to more than 50 plant species [3]. Also, a broad-spectrum of ornamental, medicinal and agricultural plants suffers from *E. orientalis* as a serious pest [4]. It mainly colonizes the upper leaf surfaces and those areas around the midribs. Discoloration of leaves and pale-yellow streaks along the midribs and veins are the main symptoms that appeared [5].

It is difficult to control this mite which may develop in the coming years and to limit biological and behavioral attitudes and its economic effect because of the misuse of a wide variety of agrochemicals that adversely affect on natural enemies [6], its stimulating on fertility and fecundity [7] and these mites are very small and difficult to detect on or inside the plants that are transported throughout the world.

Phytoseiid mites are important natural enemies of several phytophagous mites and other pests on various crops [8, 9]. *Euseius scutalis* was reported to be a common phytoseiid mite in Middle East countries (Lebanon, Iran, Egypt, Jordan) and North Africa on a variety of host plants including *Citrus* spp. [8]. Previous studies have shown that *E. scutalis* is not only a predator of spider mites, but also feeds on eggs and immature stages of whiteflies [10-12].

Control of *E. orientalis* is based on the chemical control and many research studies were focused on this method. For example, Tanigoshi *et al.* [13] evaluated the toxicity of fenbutatin-oxide, cyhexatin, bromopropylate and amitraz on lemon orchards, *Citrus limon* Burmann. Moreover, Márquez *et al.* [3] tested the efficacy of dicofol, propargite, hexitiazox, etoxazol and fenpiroximate against *E. orientalis* on Valencia-late orange crops and Fine lemon. Several phytosanitary treatments per year are usually needed for its control. In Egypt, acaricides are the primary means used to control *E. orientalis*. Abamectin, one of acaricides which applied in Egypt, belongs to the family known as macrocyclic lactones. It is derived from a life form growing in soil. Scientists first became aware of microorganisms in the mid-1970s [14]. It is often used as insecticide, an acaricide and a nematicide and is sold under many commercial names such as Vertimec, Reaper, CAM-MEK 1.8% and Termictine 5% EC. It does not remain in the environment and does not increase in proportion [15]. Thiamethoxam is used for applying on *Tetranychus urticae* (Koch) and its predator *Phytoseiulus persimilis* [16]. Moreover, thiamethoxam, a second generation of neonicotinoid, has excellent systemic characteristics and control of a broad range of commercially important pests, such as aphids, jassids, whiteflies, thrips, rice hoppers, Colorado potato beetle, flea beetles and wireworms, as well as some lepidopteran species. In addition, it can prevent some of virus transmissions [17]. Fenpyroximate is the ISO common name for tert-butyl (E)-alpha-(1,3-dimethyl-5-phenoxypyrazol-4-ylmethyleneamino-oxy)-p-toluate (IUPAC). This pesticide was applied on *T. urticae* [18]. Chlorfenapyr can disrupt ATP production and loss of energy leading to cell dysfunction and subsequent death of the organism [19]. It was applied against various pests such as termites [20], stored product-insects [21] in addition to *T.urticae* [22]. This molecule has low mammalian toxicity and is classified as slightly hazardous insecticide as per WHO criterion [23].

Phytosanitary authorities have adopted emergency safety measures to avoid the spread of the pest but further studies concerning its biology, behavior and control are needed to generate adequate knowledge to develop optimal control policies. This work aimed to evaluate the efficacy of seven chemicals of plant origin to control *E. orientalis* and predatory mite *Euseius scutalis* as natural enemy under field conditions in Assuit Governorate, Egypt.

2. Materials and methods

2.1. Commercial acaricides

Based on Table 1, Acarine (Abamectin 5% EC), Gat Fast (2% Abamectin + 10% Thiamethoxam (12% SC)), Ortis Super (Fenpyroximate 5% EC), Concord (Chlorfenaypyr 24% EC), Perfect (2% Abamectin + 10% Chlorfynapyr (12% EW), Micronet S (80% WP Sulfur) and Acarots (Fenproximate 5% SC) were obtained from agriculture company for pesticides. The recommended doses: (25 ml/100 l water for Acarine, 160 ml/100 l water for Gat Fast, 200 ml/100 l water for Ortis Super, 60 ml/100 l water for Concord, 120 ml/100 l water for Perfect, 500 g/100 l water for Micronet S and 50 ml/100 l water for Acarots), for direct spray were only applied in this work. Physiological effect to each acaricide was indicated.

Table1. Components evaluated in the assay to control *E. orietalis* and *E. scutalis.*

Commercial name	Active component and formula type	Effect	Recommended dose.
Acarine	Abamectin 5% EC	Contact and ingestion [3]	25 ml/100 l water
Gat fast	2% Abamectin + 10% Thiamthoxam (12% SC)	Contact and ingestion [3]	160 ml/100 l water
Ortis super	Fenpyroximate (5% EC)	Mitochondrial electron transport inhibitor with contact action [3,18]	200 ml/100 l water
Concord	Chlorfenpyr (24% EC)	ATP production inhibitor by contact [19]	60 ml/100 l water
Perfect	2% Abamectin + 10% Chlorfenaypr (12 EW %)	Contact	120 ml/100 l water
Micronet S	Sulfur (WP 80%)	Contact	500 gm/100 l water
Acarots	Fenproximate (5% SC)	Mitochondrial electron transport inhibitor with contact action [3, 18]	50 ml/100 l water

2.2. Field assessment

The study was applied during month of May 2018, under field condition. The field experiments were carried out at the experimental research station, Assiut, (Egypt) on *Citrus sinensis* trees infested with *E. orientalis* on their leaves with the observation *E. scutalis* existence feeding on its prey. Twenty-five trees were under study, distributed on two feeden (one tree per each acaricide

and four replicates for each component in addition to control (well water). Ten infested leaves from each tree were marked and packaged as groups for each chemical prior spraying. To avoid interaction among treatments, groups of marked leaves were separated from each other.

To calculate the number of mites that inhabited the leaves prior to spraying, researchers counted the total number of adults on each leaf with the use of a magnifying glass. Afterwards, seven acaricides with recommended dose (RD) were sprayed on the respective trees with the use of sprayer (10 L capacity). In order to determine the count for three days, one week and two weeks after the spraying, the total number of adults on ten marked leaves was calculated by using a magnifying glass and by touching each mite to observe its movement.

2.3. Statistical analysis

Reduction rate was determined by Henderson and Tilton equation [24]:

$$\text{The corrected \% reduction} = \left(1 - \frac{\text{n in control before treatment} \times \text{n in T after treatment}}{\text{n in control after treatment} \times \text{n in T before treatment}}\right) \times 100$$

Where n = number of mites, T= treated mite, Co=control mite.

Microsoft Excel (2010) was applied to determine mean of number of populations. Statistically, all variables were examined with the use of one-way analysis of variance (ANOVA).

3. Results and Discussion

Presented data showed efficacy of seven acaricides at RD against *E. orientalis* and its predatory mites, *E. scutalis*, under field conditions. These compounds are described by non-contaminated pesticides [15], which derived from plants to reduce problems caused by other chemical pesticide groups.

Tables 2 and 3 and Figures 1 and 2 indicated the mean number of individuals of *E. orientalis* and *E. scutalis* per ten leaves pre-spraying and after spraying during assorted exposure periods. For Acarine (Abamectin 5% EC), mean number of populations of *E. orientalis* and *E. scutalis* pre-spraying, were 189.25 and 23.5, respectively. Three days, one week and two weeks after applying, mean number of *E. orintalis* and *E. scutalis* individuals were 24.75 and 22.25; 25.5 and 22.25; 33.75 and 22.25, respectively, with mean of reduction rate of 89% and 11.36%, 89.2% and 26.9% and 86.5% and 32.81%, respectively.

By applying Gat Fast (2% Abamectin + 10% Thiamethoxam), mean number of populations of *E. orientalis* and *E. scutalis*, pre-spray, were 183.3 and 20.25, respectively. After spraying for three days, one week and two weeks, mean number of *E. orientalis* individuals were 18.8, 20.5 and 28, and mean reduction rates were 91.73%, 91.24% and 88.23%, respectively whereas mean number of *E. scutalis* were 18, 20.75 and 22.5 with mean reduction rate of 16.78%, 20.9% and 21.15%, respectively. The effects of Thiamethoxam on *T. urticae* were higher when residual and contaminated food exposures were considered according to Pozebon *et al*. [16]. These researchers reported that the total effect was higher than 90% when contaminated food exposure was involved. On *P. persimilis*, the total effect was higher in residual and contaminated prey exposures compared with topical exposure, and all combinations of routes of exposure attained a total effect higher than 90%. Low use rates, flexible application methods, excellent efficacy, long-lasting residual activity and favorable safety profile make this new insecticide well-suited for modern integrated pest management programs in many cropping systems [17].

Based on Ortis Super (Fenpyroximate 5% EC), mean numbers of individuals of *E. orientalis* and *E. scutalis* at pre-spraying were 166.5 and 18, respectively. After three days, one week and two week spraying, mean numbers of *E. orientalis* population were 23.25, 23.25 and 30 with mean reduction rate of 88.61%, 88.96%, 86.39%, respectively, and for *E. scutlais* were 17, 20 and 21 with mean reduction rate of 11.58%, 14.2% and 17.2%, respectively.

For Concord (Chlorfenapyr 24% EC), mean numbers of *E. orientalis* populations after pre-spraying were 200.75 and then after three days, one week and two weeks spraying, the mean numbers of individuals were 22, 26 and 37.75 with reduction rates of 91%, 89.85% and 85.89%, respectively. Additionally, mean numbers of individual *E. scutalis* were 14, 12.5, 16.25 and 17.5 at pre-spraying, three days, one week and two weeks after exposure with mean reduction rates of 16.41%, 10.4% and 11.29%, respectively.

For Perfect (2% Abamectin+10% Chloefenapyr (12 EW %)), Micronet S (Sulfur 80% WP) and Acarots (Fenpyroximate 5% SC); mean numbers of *E. orientalis* populations at pre-spraying were 168.5, 185.25 and 194.25 and of *E. scutalis* were 18.5, 21 and 19, respectively. Mean numbers of individuals after three days were 16.75, 19.75 and 24.25 with mean reduction rates of 91.76%, 91.21% and 89.83% for *E. orientalis*, and 19.75, 19.25 and 16.75 for *E. scutalis* with reduction rates of 11.44%, 14.18% and 17.47%, respectively. For one-week exposure, mean numbers of individuals (*E. orientalis*) were 21.25, 23.5 and 26 with mean reduction rates of 90.13%, 90.00% and 89.35%, respectively, whereas mean numbers of *E. scutalis* were 19.75, 22.5 and 21.75 with reduction rates of 17.59%, 17.29% and11.63%, respectively. Two-week exposure indicated that mean numbers of *E. orientalis* were 34.75, 39.25 and 40.5 with reduction rates of 84.44%, 84.01% and 84.29%, respectively, and 21.5, 24.25 and 22.25 for *E. scutalis* with reduction rates of 17.52%, 18.05% and 16.89%, respectively.

Kenneth *et al.* [22] reported that *Tetranychus urticae* mortality from Chlorfenapyr residues was significantly greater than the control (1, 3, 7, and 14 d after application). Even after two weeks, Chlorfenapyr residues caused 55% mortality to adult *T. urticae* compared to 6% mortality in the control. Also, *Tetranychus urticae* mortality from Bifenthrin and Abamectin residues was not significantly greater than the control at 1 d after application. However, *T. urticae* mortality for both Bifenthrin and Abamectin residues was significantly greater than the control (3, 7 and 14 d after application) due to its bio-origin. In addition, more researches are needed on the mode of action and compatibility of tested acaricides with biological agent products (e.g. predatory mites and entomopathogenic fungi) and other available acaricides.

Table 2. Mean number of *E. orientalis* population before treatment and after treatment for three days, one-week and two-week exposure.

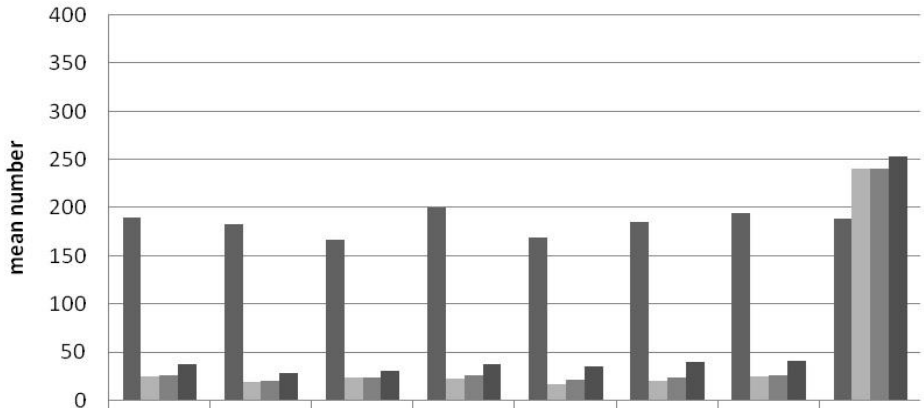
Acaricide	Exposure period			
	Day 0	Three days	One week	Two weeks
	Mean number ±SD	Mean number ± SD	Mean number ± SD	Mean number ± SD
Acarine	189.25 ± 25.81	24.75 ± 4.75^{b}	25.5 ± 6.56^{b}	33.75 ± 9.63^{b}
Gat fast	183.25 ± 29.03	18.75 ± 5.12^{a}	20.5 ± 4.20^{a}	23 ± 2.16^{a}
Ortis	166.25 ± 21.52	23.25 ± 4.65^{b}	23.25 ±4 .65^{b}	30 ± 5.71^{b}
Concord	200.75 ± 20.47	22.00 ± 4.83^{b}	26 ± 5.23^{b}	37.75 ± 5.74^{b,c}
Perfect	168.25 ± 33.01	16.75 ± 5.9^{a}	21.25 ± 4.79^{a}	34.75 ± 5.38^{b}
Micronet S	185.25 ± 31.19	19.75 ± 4.65^{b}	23.5 ± 5.51^{b}	39.25 ± 4.5^{c}
Acarots	194.25 ± 15.59	24.25 ± 4.65^{b}	26 ± 4.97	40.5 ± 4.20^{c}
control	189 ± 17.03	229.8 ± 14.73	240.5 ± 22.69	252.75 ± 16.62

Means followed by different letters in the same column differ significantly ($p<0.05$).

Table 3. The corrected reduction (%) of *E. oreintalis* and its predatory mite, *E. scutalis*, on *C. sinensis* crop treated with assorted acaricides in May 2018.

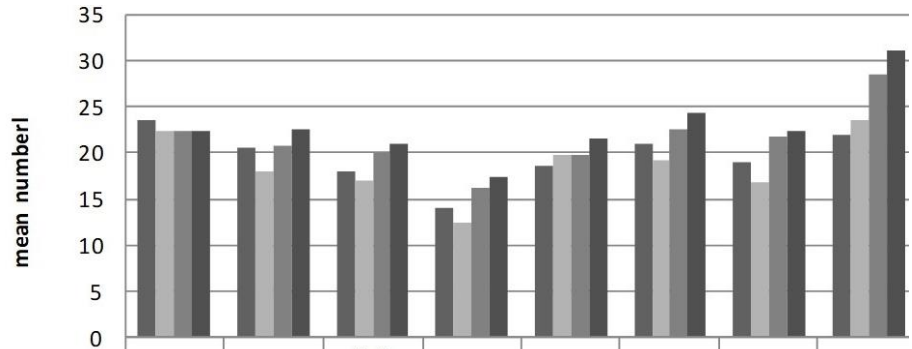
Acaricide	Mean of Red.% after spray					
	Three days		One week		Two weeks	
	E.orientalis	*E.scutalis*	*E.orientalis*	*E.scutalis*	*E.orientalis*	*E.scutalis*
Acarine (Abamectin 5% Ec)	89.00% ± 3.14^{b}	11.36% ± 0.07^{a}	89.22% ± 3.19^{c}	26.91% ± 0.41^{f}	86.57% ± 3.70^{b}	32.81% ± 0.31^{f}
Gat fast (2%Abamectin+ 10% thiamthoxam)	91.73% ± 0.32^{a}	16.78% ± 0.42^{c}	91.24% ± 0.77^{a}	20.9% ± 0.15^{e}	88.23% ± 2.92^{a}	21.15% ± 0.38^{e}
Ortis Super (Fenpyroximate 5% Ec)	88.61% ± 1.10^{b}	11.58% ± 0.52^{a}	88.96% ± 2.01^{b}	14.2% ± 0.07^{c}	86.39% ± 3.02^{b}	17.2% ± 0.97^{c}
Concord (Chlorfenapyr)	91.00% ± 1.61^{b}	16.41% ± 0.85^{b}	89.85% ± 1.50^{b}	10.4% ± 0.97^{a}	85.89% ± 2.27^{c}	11.29% ± 0.62^{a}
Perfect (2% Abamectin + 10% Cholrfenapyr)	91.76% ± 2.43^{a}	11.44% ± 0.39^{a}	90.13% ± 0.61^{b}	17.59% ± 0.69^{d}	84.44% ± 1.76^{c}	17.52% ± 0.69^{c}
Micronet S (Sulfer 80% WP)	91.21% ± 1.63^{a}	14.18% ± 0.33^{b}	90.00% ± 1.76^{b}	17.29% ± 0.50^{d}	84.01% ± 2.02^{c}	18.05% ± 0.82^{d}
Acarots (Fenpyroximate 5% SC)	89.83% ± 0.93^{b}	17.47% ± 0.41^{c}	89.35% ± 2.71^{b}	11.63% ± 0.45^{b}	84.29% ± 2.73^{c}	16.89% ± 0.29^{b}

Means followed by different letters in the same column differ significantly ($p<0.05$).



	Acarine	Gatfact	Ortis Super	Concord	Perfect	Mecronet S	Acrots	Control
prespry	189.25	183.25	166.5	200.75	168.5	185.25	194.25	189
podt three day	24.75	18.75	23.25	22	16.75	19.75	24.25	240.5
post one week	25.5	20.5	23.25	26	21.25	23.5	26	240.5
post two week	37.75	28	30	37.75	34.75	39.25	40.5	252.75

Figure 1. Mean number of *E. orientalis* individuals/10 leaves: pre and post spray on citrus crop in Assiut Governorate



	Acarine	Gat fact	Ortis SUPER	Concord	Perfect	Mecronet S	Acarots	Control
pre spray	23.5	20.5	18	14	18.5	21	19	22
post 3 days spray	22.25	18	17	12.5	19.75	19.25	16.75	23.5
post 7 days	22.25	20.75	20	16.25	19.75	22.5	21.75	28.5
post 14 day spray	22.25	22.5	21	17.5	21.5	24.25	22.25	31

Figure2. Mean number of *Euseius scutalis* individual/10 leaves on citrus crop in Assiut Governorate

The best acaricides that reduced *E. orientalis* population were Perfect (2% Abamectin + 10% Chlorfenapyr, 12% EW) with the mean number of 16.75±5.9, Gat Fast (2% Abamectin + 10% Thiamethoxam, 12% SC) with the mean number of 18.75±5.12 after three-day exposure and continued in superiority after one week and two weeks.

The data in Table 3 show the corrected reduction (%) of *E. oreintalis* and its predatory mite, *E. scutalis*. After three-day exposure, the acaricides that showed high efficiency for controlling *E. oreintalis* were Perfect (2%Abamectin + 10% Chlorphynapyr) with 91.76% for *E. orientalis* and 11.44% for *E. scutalis*. For one-week and two-week exposure, the corrected reduction (%) of *E. orientalis* and *E. scutalis* were Gat Fast (2% Abamectin + 10% Thiamthoxam) with 91.24% for *E. orientalis* and 20.9% for *E. scutalis*, for one week and 88.23% and 21.15%, for two weeks, respectively.

Few studies recorded the effect of acaricides against *E. orientalis*. Márquez *et al*. [3] recorded that Dicofol, Propargite, Hexitiazox, Etoxazol and Fenpyroximate can cause a decrease in the mortality of *E. orientalis* by 100, 98.85, 85.05, 83.92 and 100%, and by 97.82, 85.92, 81.87, 100 and 100% on Valencia-late orange crops and Fine lemon after one week of exposure, respectively. Motoba *et al.* [18] recorded that *T. urticae* sprayed with 0.5 µg/ml Fenpyroximate showed distortions in mitochondria, in peripheral nerve cells such as swelling, irregular cristae arrangement, and lower matrix electron density caused by the compound when observed by transmission electron microscopy. Similar distortions in mitochondria were also obvious in the ovaries and epidermal cells, but not in muscular cells or central nervous mass cells. So, this composition is the most applied in Egypt. On the other hand, EFSA 2013 [25] concluded that Fenpyroximate had low ecotoxic to soil dwelling organisms and LC_{50} inhalation [0.21 - 0.33 mg/l air/4-h (nose-only)] of rat showed that the substance is very toxic.

Alhewairini [15] estimated that adult mortality percentages of *E. orientalis* were 17.40, 20.62, 23.77, 36.70, 41.60, 59.60 and 79.80% after one week of exposure to 500, 1000, 2000, 3000, 4000, 5000 and 6000 ppm of Huwa-San TR50, respectively, under field conditions. The populations of *E. orientalis* reduced to 76.68 and 79.56% and to 78.52 and 80.12% after one-week exposure to the recommended dose (RD) of Abamectin and Bifenthrin under field and laboratory conditions, respectively.

Consequently, all components, in particular Abamectin and fenpyroximate, can be used as pesticide with effectiveness against *E. orientalis*, without affecting on *E. scutalis* population in the field. From this investigation, it indicates that *E. scutalis* did not significantly affect the used compounds from plant-in origin. Natural enemies have received considerable attention in restriction of potential pest populations [26]. Biological control of arthropod pests has been traditionally used for a long time in different crops, therefore, it should be used with other compatible integrated pest management methods [27]. Conservation biological control (CBC) defines as the protection of natural enemies against adverse effects of pesticides and incompatible cultural practices and improving their efficiency by providing food sources [28, 29]. The use of CBC as a component of integrated mite management in agriculture is a strategy that is increasingly important and popular. Concurrent with the increasing use of CBC in agriculture has been a realization that 'generalist' natural enemies (i.e. those who have a broad prey preference) can often play a major role in mite suppression [30]. Thus, CBC as a strategy that enhances guilds or communities of both specialist and generalist natural enemies is now viewed as a mite management strategy, very likely to improve crop protection. Another factor that has encouraged and enhanced the use of CBC in many crop systems is the availability and the use of pesticides that are narrow-spectrum and safe to many beneficial insects and mites [31-33]. CBC research in many crop systems is focused on improving reliability by strengthening the natural enemy community both in terms of population density and species diversity [34].

4. Conclusions

Evaluation of seven acaricides of bio-origin against *E. orientalis* achieved high reduction rates in the field without harming its predatory mites, *E. scutalis*. All applications were under recommended dose for each component. Our results appeared that Gat Fast (2% Abamectin + 10% Thiamthoxam) showed high effective against *E. orientalis* with slight negative impact towards *E. scutalis* within three week exposure by the application of 160cm^3 /100 l water. Mode of action of Abamectin and Thiamthoxam should further be studied.

References

[1] Sewify, G.H. and Mabrouk, A.M., 1991. The susceptibility of different stages of the citrus brown mite *Eutetranychus orientalis* (Klein) (Acarina: Tetranychidae) to the entomopathogenic fungus *Verticillium lecanii* (Zimm) Viegas. *Egyptian Journal of Applied Biology and Control*, 1, 89-92.

[2] Walter, D.E., Halliday, R.B. and Smith, D., 1995. The oriental red mite, *Eutetranychus orientalis* (Klein) (Acarina, Tetranychidae) in Australia. *Journal of the Australian Entomological Society*, 34, 307-308.

[3] Márquez, A., Wong, E., García, E. and Olivero, J., 2006. Efficacy assay of different phytosanitary chemicals for the control of *Eutetranychus orientalis* (Klein) (Oriental Spider Mite) on Fine lemon and Valencia-Late orange crops. *IOBC/WPRS Bulletin*, 29, 305-310.

[4] Rasmy, A.H., 1978. Biology of the citrus brown mite, *Eutetranychus orientalis* as affected by some citrus species. *Acarologia*, 19, 222-224.

[5] Ledesma, C., Vela, J.M., Wong, E., Jacas, J.A. and Boyero, J.R., 2011. Population dynamics of the citrus oriental mite, *Eutetranychus orientalis* (Klein) (Acari: Tetranychidae), and its mite predatory complex in southern Spain. *IOBC/WPRS Bulletin*, 62, 83-92.

[6] McMurtry, J.A., Huffaker, C.B. and van de Vrie, M., 1970. Ecology of tetranychid mites and their natural enemies: a review. I. Tetranychid enemies: Their biological characters and the impact of spray practices. *Hilgardia*, 40, 331-458.

[7] Luckey, T.D .,1968. Insecticide hormoligosis. *J Econ Entomol* 61:7-12.

[8] Bounfour, M. and McMurtry J.A., 1987. Biology and ecology of *Euseius scutalis* (Athias-Henriot) (Acarina: Phytoseiidae). *Hilgardia* ,55(5), 1-23.

[9] McMurtry, J.A. and Croft, B.A., 1997. Life-styles of Phytoseiid mites and their roles in biological control. *Annual Review of Entomology*, 42, 291-321.

[10] Meyerdirk, D.E. and Coudriet, D.L., 1986. Evaluation of two biotypes of *Euseius scutalis* (Acarina: Phytoseiidae) as predators of *Bemisia tabaci* (Homoptera: Aleyrodidae). *Journal of Economic Entomology*, 79(3), 659-663.

[11] Yıldız, S., 1998. *Determination of the Phytoseiidae Species from Vegetable Growing Areas of the East Mediterranean-Turkey*. MSc, Çukurova University, Institute of Natural and Applied Sciences (Turkish with English summary).

[12] Nomikou, M., Janssen, A., Schraag, R. and Sabelis, M.W., 2001. Phytoseiid predators as potential biological control agents for *Bemisia tabaci*. *Experimental & Applied Acarology*, 25, 271-291.

[13] Tanigoshi, L.K., Bahdousheh, M., Babcock,J.M. and Sawaqed, R., 1990. *Euseius scutalis* (Athias-Henriot) a predator of *Eutetranychus orientalis* (Klein) (Acari: Phytoseiidae, Tetranychidae) in Jordan: toxicity of some acaricides to *E. orientalis*. *Arab Journal of Plant Protection*, 8, 114-120.

[14] Lasota, J.A. and Dybas, R.A., 1990. Abamectin as a pesticide for agricultural use. *Acta Leidensia*, 59, 217-225.
[15] Alhewairini, S.S., 2018. Efficacy comparison of HUWA-SAN TR50, Abamactin and Bifenthrin for the control of the oriental spider mite, *Eutetranychus orientalis* (Klein) (Acari: Tetranychidae). *Pakistan Journal of Agricultural Science*, 55 (4), 1003-1007.
[16] Pozebon, A., Duso, C., Tirello, P. and Ortiz, P.B., 2011. Toxicity of thiamethoxam to *Tetranychus urticae* Koch and *Phytoseiulus persimilis* Athias-Henriot (Acari Tetranychidae, Phytoseiidae) through different routes of exposure. *Pest Management Science*, 67 (3), 352-359.
[17] Maienfisch, P., Angst, M., Brandl, F., Fischer, W., Hofer, D., Kayser, H., Kobel, W., Rindlisbacher, A., Senn, R., Steinemann, A. and Widmer, H., 2001. Chemistry and biology of thiamethoxam: a second generation neonicotinoid. *Pest Management Science*, 57 (10), 906-9013.
[18] Motoba, K., Suzuki, T. and Uchida, M., 1992. Effect of a new acaricide, fenpyroximate, on energy metabolism and mitochondrial morphology in adult female *Tetranychus urticae* (two-spotted spider mite). *Pesticide Biochemistry and Physiology*, 43 (1), 37-44.
[19] Raghavendra, K., Barik, T.K., Sharma, P., Bhatt, R.M., Srivastava, H.C., Sreehari, U. and Dash, A.P., 2011. Chlorfenapyr: a new insecticide with novel mode of action can control pyrethroid resistant malaria vectors. *Malaria Journal,* 10(16), https://doi: 10.1186/1475-2875-10-16.
[20] Misbah-Ul-Haq, M., Khan, I.A., Farid, A., Ullah, M., Gouge, D.H. and Baker, P.B., 2016. Efficacy of indoxacarb and chlorfenapyr against Subterranean termite *Heterotermes indicola* (Wasmann) (Isoptera). *Turkiye Entomoloji Dergisi*, 40 (3), 227-241.
[21] Arthur, F.H., 2009. Efficacy of chlorfenapyr against adult *Tribolium castaneum* exposed on concrete: effects of exposure interval, concentration and the presence of a food source after exposure. *Insect Science*, 16, 157-163.
[22] Kenneth, W.C., Edwin, E.L. and Peter, B.S., 2002. Compatibility of acaricide residues with *Phytoseiulus persimilis* and their effects on *Tetranychus urticae*. *American Society for Horticulture Science*, 37 (6), 906-909.
[23] Tomlin, C.D.S., 2000. *The Pesticide Manual: A World Compendium*. 12th ed. London: British Crop Protection Council.
[24] Henderson, C.F. and Tilton, E.W. 1955. Test with acaricides against the brown wheat mite. *Journal of Economic Entomology*, 48, 157-161.
[25] European Food Safety Authority, 2013. Conclusion on the peer review of the pesticide risk assessment of the active substance fenpyroximate, *EFSA Journal*, 11 (12), 3493, https://doi.org/10.2903/j.efsa.2013.3493
[26] Debach, P. and Rosen, D., 1991. *Biological Control by Natural Enemies*. 2nd ed. Cambridge: Cambridge University Press.
[27] Jonsson, M., Maratten, S.D., Landis, D.A. and Gurr, G.M., 2008. Recent advances in conservation biological control of arthropods by arthropods. *Biological Control*, 45, 172-175.
[28] Barbosa, P., 2003. *Conservation Biological Control*. San Diego: Academic Press.
[29] Marino, P.C., Landis, D.A. and Hawkins, B.A., 2006. Conserving parasitoid assemblages of North American pest Lepidoptera: Does biological control by native parasitoids depend on landscape complexity. *Biological Control*, 37, 173-185.
[30] James, D.G., 2001. History and perspectives of biological mite control in Australian horticulture using exotic and native phytsoeiids. *Acarology: Proceedings of the 10th International Congress.* Melbourne, Australia, 436-443.
[31] James, D.G., 2002. Selectivity of the miticide, bifenazate and aphicide,pymetrozine, to spider mite predators in Washington hops. *International Journal of Acarology*, 28, 175-179.

[32] James, D.G., 2003. Pesticide susceptibility of two coccinellids (*Stethorus punctum picipes* (Casey) and *Harmonia axyridis* Pallas) important in biological control of mites and aphids in Washington hops. *Biocontrol Science and Technology*, 13, 253-259.
[33] James, D.G., 2004. Effect of buprofezin on survival of immature stages of *Harmonia axyridis*, *Stethorus punctum picipes* (Coleoptera: Coccinellidae), *Oriu tristicolor* (Hemiptera: Anthocoridae) and *Geocoris* spp. (Hemiptera: Geocoridae). *Journal of Economic Entomology*, 97, 900-904.
[34] Chandler, D., Davidson, G., Pell, J.L., Ball, B.V., Shaw, K. and Sunderlan, K.D., 2000. Fungal biocontrol of Acari. *Biocontrol Science and Technology*, 10 (4), 357-384.

Effects of Compressed Pressure and Speed of the Tandem Mill of Sugar Cane Milling on Milling Performance

Wichian Srichaipanya[1] and Somchai Chuan-Udom[1, 2]*

[1]Department of Agricultural Engineering, Faculty of Engineering, Khon Kaen University, Khon Kaen, Thailand
[2]Applied Engineering for Important Crops of the North East Research Group, Khon Kaen University, Khon Kaen, Thailand

Received: 10 July 2019, Revised: 31 October 2019, Accepted: 20 March 2020

Abstract

Thailand is one of the sugar cane exporters in the world market and it offers a great contribution to the country's national income. It is important for sugar factory to operate efficiently in milling against its competition to achieve a better performance. This study investigated the effects of pressure and speed of the tandem mill in sugar cane crushing based on extraction (EXT), power requirements (POW), specific energy consumption (SEC) and specific extraction per energy consumption (SEE) which are important parameters for sugar factory. Testing was performed using a small milling machine. The compressed pressure was adjusted by using the hydraulic pressure and the inverter was used to change the speed. After the test, the polarization meter was used to measure sugar in juice. It was found that when the compressed pressure increased, the EXT of the tandem mill had a tendency to increase from 2.7 to 6.0 x 10^6 N/m^2 and slightly increased from pressure 6.0 to 10.0 x 10^6 N/m^2. The POW of the tandem mill slightly increased. The SEC of the tandem mill tended to increase its tendency. The SEE slightly increased from pressure 2.7 to 6.0 x 10^6 N/m^2 and severely decreased from pressure 6.0 to 10.0 x 10^6 N/m^2. When the speed of the tandem mill increased, the EXT of the tandem mill had a tendency to decrease, the POW tended to increase their tendency, the SEC slightly decreased from 0.0844 m/s to 0.1 m/s and slightly increased from 0.1 m/s to 0.1631 m/s. The SEE slightly decreased. The maximum extraction (%) is earned at the intermediate compressed pressure and the lowest speed. The minimum power requirement (W) is earned at the lowest compressed pressure and the lowest speed. The minimum specific energy consumption (kWh/t) is earned at the lowest compressed pressure and the intermediate speed. The maximum specific extraction per energy consumption (%.t/kWh) is earned at the intermediate compressed pressure and the lowest speed. The results found in this study can benefit sugar factory to find performance and to make the decision on process operation.

Keywords: Tandem mill, extraction, sugar cane milling
DOI 10.14456/cast.2020.14

*Corresponding author: Tel.: +66 43 00 9700 Fax: +66 43 36 2149
E-mail: somchai.chuan@gmail.com

1. Introduction

Sugar trade averages 56 million tons/year in the world whereas Brazil, Thailand and Australia accounting for 65% of the trade in 2014. The world's largest sugar producer and exporter is Brazil, accounts for 24.01 million tons. Thailand, the second one, accounts for 7.97 million tons [1]. There are several processes in the sugar cane production such as sugar cane growing, sugar cane milling, credit banking, exportation, etc. [2]. The sugar production comprises juice extraction, preheating, evaporation, crystallization, centrifugal and drying [3]. The sugar refinery process is part of sugar cane crushing mills. Furthermore, raw sugar and refined sugar are produced by using sugar cane bagasse as power that take out from milling [4]. Currently, there are 55 mills in Thailand with a sugar cane crushing capacity about 93 million tons per year [5].

A study by Atherton [6] showed the relationships between the pressure and the volume of juice expressed whereas the volume of juice expressed was up to any given pressure increased. The effect of speed, preparation and compression ratio of performance of the experimental sugar mill in the University of Queensland showed that with small mill and no juice grooves, the coarser preparations gave better drainage, resulting in good juice extraction [7]. Compression ratio and performance of experimental sugar mill showed that an increase in roll load led to the uneconomical use of the mill [8].

A small scale sugar cane juice mill was developed for farmers who were involved in the processing of sugar, ethanol and other related products. The extraction efficiency ranged between 40 and 61% at operating speeds of 0.25 and 0.36 m/s and the output capacities were 10.50, 12.00 and 14.25 kg/h at operating speeds of 0.25, 0.3 and 0.36 m/s respectively [9, 10]. Sugar cane milling is one step in sugar process and there are several machines in the sugar process such as sugar cane unloading, sugar cane knives, shredders, sugar cane milling, bagasse conveyor and so on [11]. Normally, there are two ways for juice extraction: 1) extraction by the mills and 2) extraction by the diffusers [12]. The sugar cane milling was simulated by finite element algorithm. Energy dispersing during milling of sugar cane could be presented in terms of four components; juice flow, bulk plasticity, seepage induced plasticity and frictional sliding. In this simulation, constant crushing rate showed that higher roll speeds and thinner blankets reduced power requirements, frictional sliding, roll torque, roll load and increased extraction of juice slightly [13]. The variable speed of the drives and different speeds of the rolls presented almost constant torques on each roll. The results of the tandem mill such as extraction, capacity and power consumption, were very good [14]. The sugar extraction efficiency and energy consumption have been compared with a similar process in a modern fuel alcohol distillery in current sugar extraction process. Distilleries could use more soak water to increase juice extraction during crushing because the fermentation broth must be diluted. An analysis of the substitution process showed that if the steam consumption of evaporator did not rise significantly, the net revenue increased significantly [15]. There was a feed opening which resulted in the maximum throughput at the same speed for a particular mill configuration. However, the feed opening affected mill torque and the forces acting on the mill housing. The maximum throughput could be limited due to insufficient roller roughness causing slippage [16].

It is very important to have some technique available in order to manage the standard milling operation that Thailand sugar factory has never had it. Therefore, the objectives of this study are to investigate the effects of compressed pressure on the speed of the tandem mill in sugar cane milling on production parameters based on Extraction (EXT), Power requirements (POW), Specific energy consumption (SEC) and Specific extraction per energy consumption (SEC).

2. Materials and Methods

2.1 Equipment used in the test

Testing was performed using a method of Shinde [17]. A small milling machine equipped with two tandem mills with a diameter of 21.5 mm and a length of 43 mm (Figure 1) was used in this study. A 5 kW electric motor was used as a power source. The compressed pressure was adjustable by using the hydraulic pressure. During the test, the inverter was used to change the speed and monitor current, voltage for power requirement, specific energy consumption and specific extraction per energy consumption calculation. After the test, the polarization meter was used to measure sugar in juice for extraction calculation. The test was performed entirely in the laboratory of the sugar factory involved in this study.

(a) (b)

Figure 1. (a) A small milling machine and (b) Two tandem mills

2.2 Test method

Testing of each parameter involved three replications, using 'Khon Kaen 3' sugar cane as a sample. Data were collected from the input sugar cane, the discharge of juice, and bagasse outlet weight. The power requirement for crushing was measured using the inverter display. The parameters obtained were used to calculate the power requirement, specific energy consumption and specific extraction per energy consumption. Moreover, the polarization meter (Anton Paar, model MCP 500 Sucromat, Austria) was used to analyze the extraction. A schematic diagram of milling process in this study is shown in Figure 2.

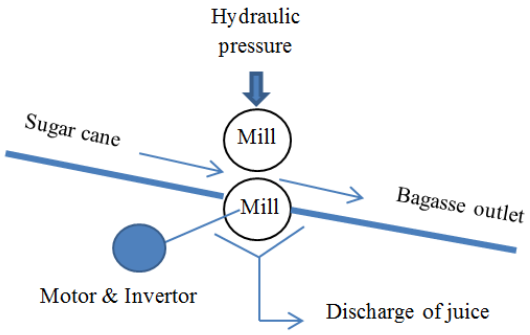


Figure 2. A schematic diagram of milling process

2.3 Indicator parameters

The indicating parameters in the test comprised extraction, power requirements, specific energy consumption and specific extraction per energy consumption.
Extraction was calculated using Eq. (1)

$$EXT = \left(\frac{B}{A+B} \right) \times 100 \quad (1)$$

where, EXT is the extraction from sugar cane milling in percent, A is the weight of sugar from bagasse in grams and B is the weight of sugar from the extracted juice in grams (ICUMSA GS1/2/3-1(1994))[18].
Power requirement was calculated using Eq. (2)

$$POW = 1.732 \times VOLT \times I \times PF \quad (2)$$

where, POW is the power requirement in watts, VOLT is the voltage, I is the current in ampere and PF is the power factor.
Specific energy consumption was determined using Eq. (3)

$$SEC = \left(\frac{POW}{FR} \right) \quad (3)$$

where, SEC is the specific energy consumption in kWh/t, POW is the power requirement in watts and FR is the feed rate of sugar cane in t/h.
Specific extraction per energy consumption was determined using Eq. (4)

$$SEE = \left(\frac{EXT}{SEC} \right) \quad (4)$$

where, SEE is the specific extraction per energy consumption in %.t/kWh, EXT is the extraction from sugar can milling in percent and SEC is the specific energy consumption in kWh/t.

2.4 Model development

Linear regression (backward elimination) is used for model development. All the independent variables (X_1, X_2, ...) are entered into the first equation and each one is deleted one at a time if they do not contribute to the regression equation. Assume the original model as Eq. (5).

$$Y = \beta_0 + \beta_1 X_1 + \Lambda + \beta_{r-1} X_{r-1} + \varepsilon . \quad (5)$$

Step 1:
At the beginning, the original model is set to be Eq. (5).
Then, the following $r-1$ tests are carried out, $H_{0j} : \beta_j = 0, j = 1,2,..., r-1$. The lowest partial F-test value F_l corresponding to $H_{0l} : \beta_l = 0$ or t-test value t_l is compared with the preselected significant values F_0 and t_0. One of two possible steps (step2a and step 2b) can be taken.
Step 2a:
If $F_l < F_0$ or $t_l < t_0$, then X_l can be deleted and the new original model is Eq. (6).

$$Y = \beta_0 + \beta_1 X_1 + \Lambda + \beta_{l-1} X_{l-1} + \beta_{l+1} X_{l+1} + \Lambda + \beta_{r-1} X_{r-1} + \varepsilon \quad (6)$$

Go back to step 1.

Step 2b:
If $F_l > F_0$ or $t_l > t_0$, the original model is the model we should choose [19].

3. Results and Discussion

3.1 Effects of compressed pressure and tandem mill speed on milling performance

Data from Table 1 were used to create analysis of variance and regression equations between the compressed pressure and the tandem mill speed, which affected milling performance such as the extraction (EXT), power requirements (POW), specific energy consumption (SEC) and specific extraction per energy consumption (SEE). The statistical variance analyzed from data in Table 1 revealed that altering compressed pressure and tandem mill speed significantly affected the extraction, the power requirements, the specific energy consumption, and the specific extraction per energy consumption ($P<0.01$), as shown in Table 2.

Table 1. Effects of compressed pressure and tandem mill speed on the extraction, power requirements, specific energy consumption and specific extraction per energy consumption

P (N/m^2)	V (m/s)	EXT (%)	POW (W)	SEC (kWh/t)	SEE (%.t/kWh)
$2.7x10^6$	0.0844	43.09±3.02	954±8.50	4.72±0.11	9.13±0.53
$2.7x10^6$	0.1238	32.44±2.70	1,362±6.93	4.60±0.09	7.05±0.47
$2.7x10^6$	0.1631	25.53±2.63	2,012±17.00	5.15±0.08	4.95±0.44
$6.0x10^6$	0.0844	57.41±2.42	1,093±5.20	5.45±0.02	10.53±0.40
$6.0x10^6$	0.1238	56.06±2.16	1,573±12.00	5.35±0.04	10.48±0.48
$6.0x10^6$	0.1631	53.72±1.80	2,187±40.73	5.64±0.10	9.52±0.31
$10.0x10^6$	0.0844	68.11±2.28	1,358±10.00	6.77±0.05	10.06±0.35
$10.0x10^6$	0.1238	63.02±0.55	1,895±13.5	6.44±0.05	9.78±0.16
$10.0x10^6$	0.1631	57.32±1.44	2,642±10.39	6.82±0.03	8.41±0.22

P = Compressed pressure, V = Tandem mill speed. Values shown as mean±SE

Table 2. Analysis of variance of extraction, power requirements, specific energy consumption and specific extraction per energy consumption affected by compressed pressure and tandem mill speed

Source of Variation	df	EXT	POW	SEC	SEE
Compressed Pressure (P)	2	429.418*	2162.271*	2058.632*	152.948*
Tandem Mill Speed (V)	2	53.194*	10122.845*	95.782*	76.476*
PV	4	7.818*	40.076*	9.204*	13.927*
Block	2	1.235**	0.931**	3.115**	0.785**

* = Highly significant at $P < 0.01$, ** = Not significant

3.2 Effects of compressed pressure and tandem mill speed on the extraction (EXT)

Data from Table 1 were used to create regression equations between the compressed pressure and the tandem mill speed, which affected the extraction as shown in Table 3. The EXT model in Table 3 indicates the relationship between the compressed pressure and the tandem mill speed affecting the extraction. For clarifying in optimization (Figure 3), when the compressed pressure increased, the extraction of the tandem mill had a tendency to increase from 2.7 to 6.0 x 10^6 N/m^2 and slightly increased from pressure 6.0 to 10.0 x 10^6 N/m^2. It can be explained that the total volume of juice expressed remains constant at higher pressure [6]. Moreover, the increased compressed pressure brought about increased compaction in the sugar cane inlet, and the increase in high compaction gave a slightly increase in extraction. Effects of speed, when increased the speed of the tandem mill, the extraction of the tandem mill had a tendency to decrease. This finding correlated with a study by Bullock [7], where the mean speed and mean juice extraction determined the mean point on the graph of extraction versus speed. The co-efficient of regression expressed the slope of the best fit to the available data. This result seems to be negative, indicating a falling-off of extraction as speed increases.

Table 3. Equation from regression analysis of compressed pressure (P) and tandem mill speed (V) on the extraction (EXT), the power requirements (POW), the specific energy consumption (SEC), and the specific extraction per energy consumption (SEE)

Model	Equation	Adj. R2	SE	P-value
EXT	$\mathrm{EXT} = 21.56 + (1.25e^{-5})\mathrm{P} - 135.71\mathrm{V} - (7.72e^{-13})\mathrm{P}^2$	0.950	3.183	0.000
POW	$\mathrm{POW} = -575.29 + (7.21e^{-5})\mathrm{P} + 14552.10\mathrm{V}$	0.977	84.24	0.000
SEC	$\mathrm{SEC} = 6.40 + (2.56e^{-7})\mathrm{P} - 43.62\mathrm{V} + (187.70)\mathrm{V}^2$	0.968	0.150	0.000
SEE	$\mathrm{SEE} = 5.54 + (2.31e^{-6})\mathrm{P} - 28.97\mathrm{V} - (2.56e^{-13})\mathrm{P}^2$	0.813	0.793	0.009

Adj. R2 = Adjusted coefficient of determination.

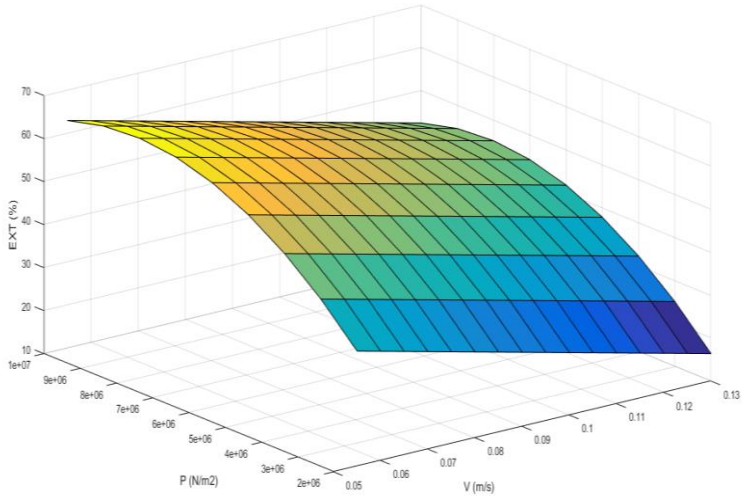


Figure 3. Effects of compressed pressure and tandem mill speed on extraction (EXT)

3.3 Effects of compressed pressure and tandem mill speed on the power requirements (POW)

Data from Table 1 were used to create a regression equation between the compressed pressure and the tandem mill speed, which affected the power requirements as shown in Table 3. The POW model in Table 3 indicates the relationship between the compressed pressure and the tandem mill speed affecting the power requirements. For clarifying in optimization (Figure 4), when the compressed pressure increased, the power requirements of the tandem mill slightly increased from 2.7 to 10.0 x 10^6 N/m^2. This result correlated with a study by Adam and Loughran [13], where the power required increased with by a mill compression ratio. Effects of speed, when the speed of the tandem mill increased from 0.0844 m/s to 0.1631 m/s, the power requirements tended to increase their tendency. Total power consumption in the whole tandem tended to increase when the speed increased [14].

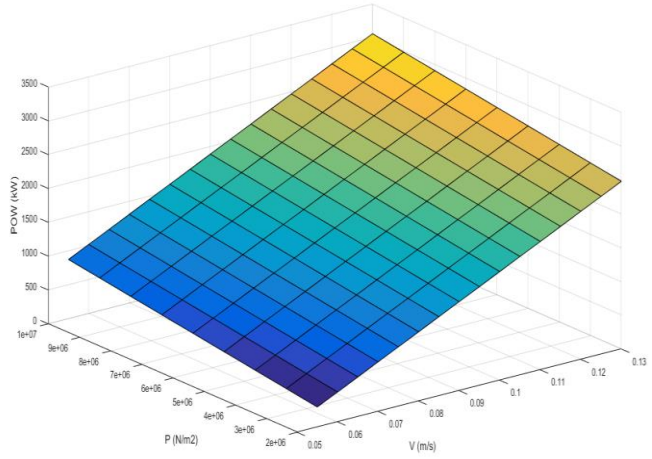


Figure 4. Effects of compressed pressure and tandem mill speed on power requirements (POW)

3.4 Effects of compressed pressure and tandem mill speed on the specific energy consumption (SEC)

Data from Table 1 were used to create a regression equation between the compressed pressure and the tandem mill speed, which affected the specific energy consumption as shown in Table 3. The SEC model in Table 3 indicates the relationship between the compressed pressure and the tandem mill speed affecting the specific energy consumption. For clarifying in optimization (Figure 5), when the compressed pressure increased, the specific energy consumption of the tandem mill tended to increase its tendency from pressure 2.7 to 10.0 x 10^6 N/m^2. Energy dispersing during milling of prepared sugar cane could be presented in terms of four components: bulk plasticity, juice flow, frictional sliding, and seepage induced plasticity. At higher compression rations with normal blanket thickness, frictional sliding on the roll surface could comprise up to 20% of total power consumption [13]. Considering the effects of speed, when the speed of the tandem mill increased, the specific energy consumption slightly decreased from 0.0844 m/s to 0.1 m/s and slightly increased from 0.1 m/s to 0.1631 m/s. According to Adam and Loughran [13], the total specific power consumption increased with increasing roll-surface speed.

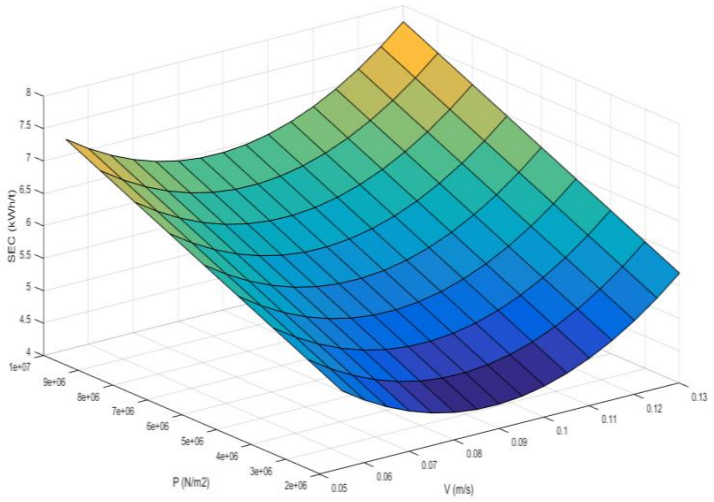


Figure 5. Effects of compressed pressure and tandem mill speed on specific energy consumption (SEC)

3.5 Effects of compressed pressure and tandem mill speed on the specific extraction per energy consumption (SEE)

Data from Table 1 were used to create a regression equation between the compressed pressure and the tandem mill speed, which affected the specific extraction per energy consumption as shown in Table 3. The SEE model in Table 3 indicates the relationship between the compressed pressure and the tandem mill speed affecting the specific extraction per energy consumption. For clarifying in optimization as shown in Figure 6, when the compressed pressure increased, the specific extraction per energy consumption slightly increased from pressure 2.7 to 6.0 x 10^6 N/m^2 and severely decreased from pressure 6.0 to 10.0 x 10^6 N/m^2. When the speed of the tandem mill increased from 0.0844 m/s to 0.1631 m/s, the specific extraction per energy consumption slightly decreased.

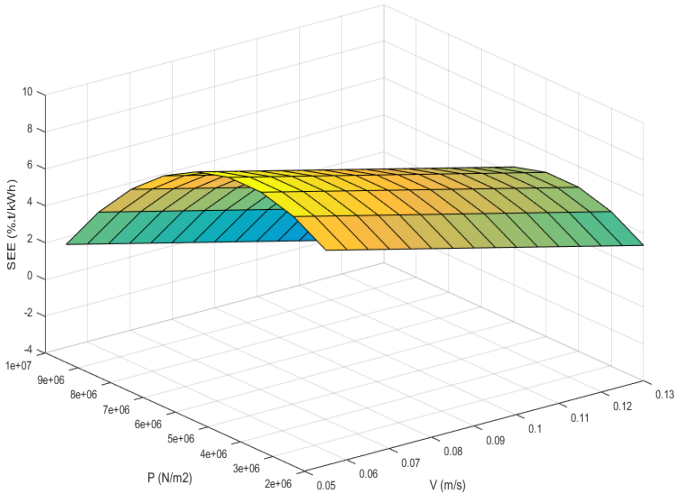


Figure 6. Effects of compressed pressure and tandem mill speed on specific extraction per energy consumption (SEE)

4. Conclusions

This study investigated the optimization of milling performance of a sugar mill. The key objectives were to analyze the effect of compressed pressure and tandem mill speed on 1) the extraction of sugar from sugar cane, 2) power requirements in the mill, 3) specific energy consumption in the mill, 4) specific extraction per energy consumption in the mill. The compressed pressure and speed were adjusted from 2.7 to 10.0 x 10^6 N/m^2 and 0.0844 m/s to 0.1631 m/s, respectively. The results showed that the maximum extraction (%) is earned at the intermediate compressed pressure and the lowest speed. The minimum power requirement (W) is earned at the lowest compressed pressure and the lowest speed. The minimum specific energy consumption (kWh/t) is earned at the lowest compressed pressure and the intermediate speed. The maximum specific extraction per energy consumption (%.t/kWh) is earned at the intermediate compressed pressure and the lowest speed. The method and results of this study can benefit for sugar factory to find performance and to make the decision on process operation

5. Acknowledgements

This research is kindly supported by Applied Engineering for Important Crop of North East Research Group, Khon Kaen University and sugar factory in Udonthani, Thailand. Thanks to my advisors and all staffs.

References

[1] International Sugar Organization, 2016. *The Sugar Market*. [online] Available at : http://www.isosugar.org/sugarsector/sugar

[2] Arjchariyaartong, W., 2006. *The Competitiveness of the Sugar Industry in Thailand*. Unpublished doctoral dissertation. University of Hohenheim, Stuttgart, Germany.
[3] Modesto, M. Ensinas, A.V. and Nebra, S.A., n.d. *Sugar Cane Juice Extraction Systems Comparison-Mill Versus Diffuser*. [online] Available at: https://www.abcm.org.br/anais/cobem/2005/PDF/COBEM2005-0836.pdf
[4] Food and Agriculture Organization of the United Nations, 1997. *Proceedings of the Fiji/FAO 1997 Asia Pacific Sugar Conference*. [online] Available at : http://www.fao.org /docrep/005/ x0513e/x0513e24.htm
[5] Office of The Sugar Cane and Sugar Board, 2017. *Quality Report of Sugar Cane C.C.S in Year 2016/2017*. Bangkok: Office of the Sugar Cane and Sugar Board.
[6] Atherton, P.G., 1954. Bagasse Compression Tests. *Proceedings Queensland Society of Sugar Cane Technologists Twenty-first Conference*, 235-240.
[7] Bullock, K.J., 1955. The effect of speed, preparation and compression ratio on the performance of the experimental sugar mill in the university of Queensland. *Proceedings Queensland Society of Sugar Cane Technologists Twenty-second Conference*, 201-218.
[8] Bullock, K.J., 1957. Compression ratio and performance of the experimental sugar mill. *Proceedings Queensland Society of Sugar Cane Technologists Twenty-fourth Conference*, 43-62.
[9] Olaoye, J.O., 2011. Development of sugar juice extractor for small scale industries. *Journal of Agriculture Technology*, 7(4), 931-944.
[10] Olaoye, J.O., 2009. Partial mechanization of sugarcane juice extraction process. *XXXIII CCIOSTA-CIGR V Conference 2009*. [online] Available at : https://www.researchgate.net/ publication/272649200
[11] Hugot, E., 1986. *Handbook of Cane Sugar Engineering*. 3rd ed. Netherlands: Elsevier Publishing Company.
[12] Oliverio, J.L. Avila, A.C.R.D. Faber, A.N. and Soares P.A. *Juice Extraction Systems: Mills and Diffuser-The Brazilian Experience*. [online] Available at: http://www.codistil.com.br/ index. php?option=com_docman&task= doc_view&gid=163&Itemid=40&lang=pt
[13] Adam, C.J. and Loughran, J.G., 2005. The effect of blanket thickness on extraction energy in sugarcane rolling mills: a finite element investigation, *Biosystems Engineering*, 92(2), 255-263.
[14] Lewinski, J., Grassmann, P. and Kallin, T., 2011. Operation of sugar mills with individual variable speed drive. *Proceedings South Africa Sugar Cane Technologists Association*, 84, 460-471.
[15] Lobo, P.C., Jaguaribe, E.F., Rodrigues, J. and Rocha, F.A.A., 2007. Economics of alternative sugar cane milling options. *Applied Thermal Engineering*, 27, 1405-1413.
[16] Wienese, A., 2003. Mill feeding: Back to basics. *Proceedings South Africa Sugar Cane Technologists Association*, 77, 369-377.
[17] Shinde, V.V., 2015. Weight reduction and analysis of sugar mill roller using FEA techniques, *International Journal of Latest Trends in Engineering and Technology*, 5(1), 346-354.
[18] International Commission for Uniform Methods of Sugar Analysis, 1994. *ICUMSA*. Norwich, England: ICUMSA Publication Department.
[19] Backward Elimination and Stepwise Regression. [online] Available at: https://www.coursehero.com/ file/16140622/ch63

Data Quality Enhancement for Decision Tree Algorithm Using Knowledge-Based Model

Sirichanya Chanmee and Kraisak Kesorn*

Department of Computer Science and Information Technology, Faculty of Science, Naresuan University, Phitsanulok, Thailand

Received: 9 March 2020, Revised: 13 March 2020, Accepted: 19 March 2020

Abstract

Data mining is an approach to discovering knowledge or unrevealed patterns from huge data sets by using several methods, such as statistics, machine learning and other data analysis techniques. However, the main limitation of these conventional techniques is the ignorance of data relationships and semantics. The data are considered as meaningless numbers with statistical methods being used for model building. For example, the decision tree, a classification method of data mining, is produced from a given set of labeled data, and those data are classified without understanding the semantics of the data or the relationships between attributes. To understand the inherent meaning in the data and to take advantage of the relationships between data elements, we introduce a knowledge-based approach to improve data quality. The proposed approach uses the ontology as the background knowledge to assist the decision tree classification in the process of data preparation. The ontology is used to infer the relationships between attributes and concepts in an ontology. This relationship information can assist the system in identifying related attributes which could assist in the classification process. Two datasets in different domains; agriculture and economics, were used to evaluate the generalization of the proposed approach. Accuracy was the standard measure of success, and was tested in the evaluation of the model. The experimental results showed that the proposed approach can efficiently enhance the performance of the data classification process.

Keywords: data analytics, data mining, ontology, semantic, classification, decision tree
DOI 10.14456/cast.2020.15

1. Introduction

Data mining [1] is an analytic approach that applies various conventional technique, such as statistical and machine learning to discover hidden knowledge within a set of data. Such datasets can be enormous in this age of Big Data. Ignoring the semantics inherent but undiscovered in the data and the inability to identify relationships between data elements are the limitation of this approach. Previously, for the purpose of building models, data was considered only as numerical values. This has been a limitation when data can, in fact, be categorical or other. To overcome this limitation, an approach called *Semantic Data Mining* has been proposed.

*Corresponding author: Tel.: +66 81 555 7499
E-mail: kraisakk@nu.ac.th

Semantic Data Mining [2] refers to an approach in which domain knowledge is incorporated into the data mining tasks to assist in the analysis process.

The domain knowledge can be used to constrain the search space and to reveal more visible patterns in the data [3], and also to identify data relationships. To illustrate, Kuo *et al*. [4] applied medical knowledge in an ontology to categorize 85 attributes relevant to cardiovascular disease, into seven groups for identifying the association rules governing cardiovascular disease resulting in death. Their experimental results found that the use of domain knowledge could help to reveal meaningful rules.

An Ontology [5] is an explicit specification of a shared conceptualization. The knowledge in an ontology is presented as a hierarchical structure of entities and relationships between them. The basis of ontology is a generalization/ specialization of concepts. For example, when focusing on aspects related to plant pathology, the terms such as fungal disease, powdery mildew, and downy mildew might be relevant concepts where the first is a general concept of the latter two. For semantic data mining, an ontology is used to assist several tasks of data mining such as association rules [6, 7], classification [8, 9], and clustering [10, 11].

Classification is a common task in data mining for categorizing prior data into the defined class. The class of each test case is considered on the observed patterns of each training data. The performance of classification depends on the quality of data to be classified, such as the number of missing values, the number of irrelevant attributes, and the size of data. For dealing with the size of data, the notion of data abstraction that denotes the general concept of each value can be applied to obtain the smaller and more general data. Also, the use of abstract data can be obtained the more meaningful data, for instance, the grade point average (GPA) attribute which is the numerical values can be generalized by the higher-level concepts for categorizing these values into several levels such as excellent, good and fair which help to present the student's academic performance.

To take advantage of the abstraction, Tang and Fong [12] proposed an approach that used the abstract values of each attribute for building the compact decision tree. The concept hierarchy was used to identify the general concepts related to primitive values. The proposed approach was evaluated by using 8,000 online auction instances. The experimental results found that the tree was simpler, and the accuracy also improved. An ontology is one approach that has been used to acquire the abstract values of each attribute, and these values are employed to improve the decision tree performance. For example, Zhang *et al*. [13] presented an ontology-based decision tree algorithm that used abstraction at multiple levels for tree induction. A customer purchase database was used to evaluate the proposed approach. The results showed that the used of abstract values could guide the decision tree induction process and help to enhance the performance of the decision tree. Vieira and Antunes [14] proposed an algorithm that the decision tree incorporated with knowledge in an ontology. The existing attributes and the abstract values which inferred from an ontology were used for a tree induction. The results showed that the proposed approach produced a compact decision tree, and accuracy also increased. As mentioned before, the use of abstraction can improve the decision tree's performance by handling the variety of attribute's values, and this approach can help to generate a smaller and more accurate decision tree. The ontology was also used to infer the related abstract concepts of each attribute and identify the association between data elements.

In this paper, we present the ontology-based approach for two disparate domains; soybean disease and census data classification. The abstract concepts in the ontology are used for the data preparation process to improve the quality of the dataset which is later used in the decision tree algorithm. Our study makes a contribution to existing research by examining whether the use of the abstract concepts in an ontology for the data preparation processes can improve the performance of the classification algorithms.

2. Materials and Methods

The framework of the knowledge-based approach for the data quality enhancement is presented in Figure 1. The details of the materials and the processes of this approach are described as follows.

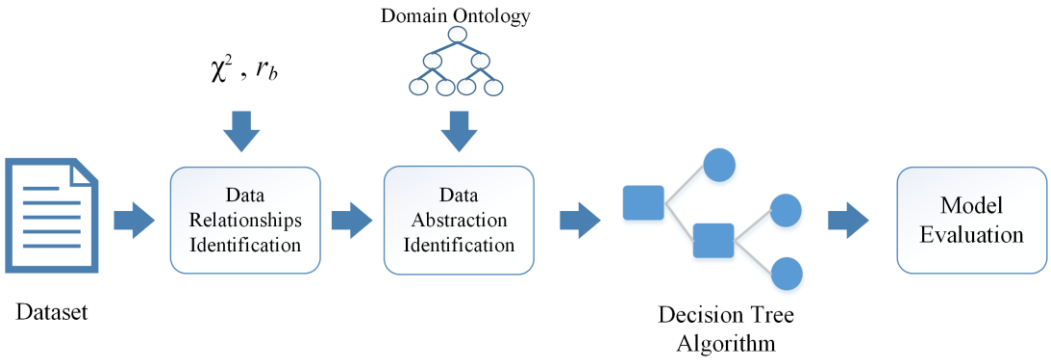


Figure 1. Framework of the data quality enhancement for decision tree algorithm using knowledge-based model

2.1 Dataset gathering and ontology designing

The datasets used for the experiment consisted of the soybean dataset and the census bureau dataset. The soybean cultivation and diseased indication dataset [15] was used in our experiment. This dataset included 35 attributes, 683 records and 15 classes of disease. There are 121 records with missing value being eliminated at the pre-processing phase for improving data quality, and only 562 remained which were used to evaluate the proposed approach.

The other dataset used in the experiment was the census bureau dataset [15]. This dataset included 13 attributes, and it were used to classify the household income in the United States into two classes. The records with missing data were eliminated then this dataset would reduce from 32,560 records to 30,161 records.

Protégé [16] designed and constructed the ontologies used in this experiment. The soybean disease ontology consisted of the concepts related to soybean diseases such as the disease name, the indicators of disease and disease type, and the environmental factors, as shown in Figure 2. The development of the soybean disease ontology adopted some ideas from the relevant existing ontologies such as the soybean ontology [17] and the ontology of rice diseases in Thailand [18]. The knowledge of disease symptoms extracted from the soybean disease diagnostic series [19] published by North Dakota State University and the expert-derived rules in Michalski's research [20]. For the census data processing, the personal information ontology necessary was designed to obtain the related abstract values for the attributes of the census dataset. This ontology was adapted from the ontology called OntoLife [21]. The designed personal information ontology consisted of the sociodemographic of a sampling group such as gender, education, marital status, and hometown

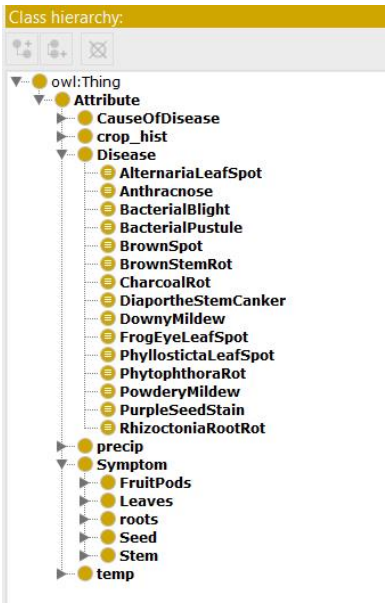


Figure 2. The structure of soybean ontology

2.2 Data relationshionships identification

Data preparation is the process for dealing with the dataset before using that data in the modeling and analysis process. The data preparation process includes data cleansing, data selection, and data transformation. In our research, the Listwise Deletion technique [22] was used for handling the missing values of a dataset by excluding the entire record with any missing variable values. This technique helped to produce the complete dataset for using in the analysis process. After this step, 562 records remained for the soybean dataset and 30,161 records for the census dataset.

For identifying the relationships between attributes and defined classes, biserial correlation (r_b) and Chi-square (χ^2) were used. The biserial correlation is a measure used to estimate the association between a dichotomous nominal variable and an interval variable [23]. The biserial correlation define as in (1) [24].

$$r_b = \frac{r_{pb}\sqrt{P_1 P_2}}{y} \tag{1}$$

where r_{pb} is the point-biserial coefficient, P_1 is the fraction of case in the first category, P_2 is the fraction of case in the second category, and y is the ordinates of the normal distribution at the point of P_1 and P_2.

Also, Chi-square is the statistic used to test the independence between variables when those variables are nominal. The formula for calculating the Chi-square value is shown as (2), derived from [25]

$$\chi^2 = \frac{(O-E)^2}{E} \tag{2}$$

where O is the frequency of the observed values and E is the frequency of the expected values.

The null hypothesis (H_0) for the Chi-square independent test is that two categorical variables are not associated. On the other hand, the alternative hypothesis (H_a) state that the two categorical variables are associated. When the p-value of the Chi-square test is less than 0.05, the null hypothesis is rejected, and the alternative hypothesis is accepted.

Thus we can conclude that the relationship between those two variables exists. When the significant results of the Chi-square test were obtained, the Cramer's V was then used as a measure of their relationships. The value of the Cramer's V is between 0 and 1 without any negative values and the interpretation of these values as shown in Table 1 [26]. After this process, the irrelevant attributes are discarded.

Table 1. Interpretation of Cramer's V

Cramer's V	Interpretation
> 0.25	Very strong
> 0.15	Strong
> 0.10	Moderate
> 0.05	Weak
> 0	No or very weak

2.3 Data abstraction identification

Data transformation is a process to transform the raw data into the appropriate format, such as the numerical values of temperature would be replaced by the defined categories: 'normal', 'greater than normal', and 'lower than normal'. In our study, the designed ontologies were used to identify the association between concepts (so called knowledge in each domain) and values in datasets, and the abstract values derived from the ontologies would replace the primitive values of the datasets in the related attributes for the data transformation process. The use of abstract values allowed the data scientists to view these data more meaningfully.

For example, as shown in Figure 3, the near ground node was the superclass in the soybean ontology, and the below soil node and the above soil node were subclasses. In the 1st process, the values of the stem canker attribute would be mapped with the knowledge in the ontology for identifying the related concepts. In the 2nd process, the related concepts obtained from the ontology would replace the original values of the stem canker attribute, and the revised dataset were used as input of the classification algorithms.

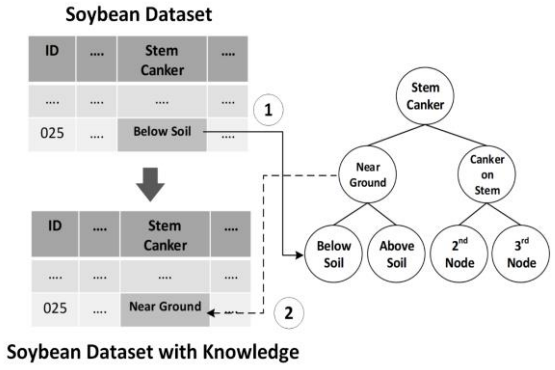


Figure 3. The process of mapping dataset with ontology

2.4 Classification process

The classification algorithm used for identifying soybean disease is a decision tree. The decision tree [27] is a classification algorithm that can handle both numerical data and categorical data, and the results are easy to interpret. This algorithm has a tree structure included nodes and branches. Each node of the tree is a decision node, and the leaf nodes are the classes/results of the classification. The decision tree induction algorithms are based on the recursive partitioning approach, and the impurity measures such as entropy and information gain are used as splitting criteria to select the most informative attribute for growing the tree.

The entropy and the information gain are defined as in (3) and (4) [28]

$$Entropy(S) = \sum_{i=1}^{c} - p_i \log_2 p_i \tag{3}$$

where S is an attribute that used to compute the entropy and p_i is the probability of the instances belonging to the i^{th} class.

$$IG(S, A) = Entropy(S) - \sum_{v \in Value(A)} \frac{|_{Sv}|}{|S|} Entropy(S) \tag{4}$$

where A is an attribute in dataset, $|_{Sv}|$ is the number of attribute A's instances which has value v, and $|S|$ is the total number of instances.

Typically, a data scientist prefers the less complex decision tree because the more complex models (larger size of the decision tree) may lead to insufficient performance [29]. The tree complexity is controlled by the stopping criteria and the pruning method. The metrics that are used to measure the tree complexity consist of the total number of nodes, the total number of tree leaves, the depth of the tree, and the number of attributes used. Also, the stopping criteria of the tree growing state include the following condition:

- All the attribute values belong to a single class.
- The tree was growth to the maximum tree's depth.
- The minimum number of instance in the parent nodes is more than the number of instances in the terminal nodes.
- The number of instances in one or child node is less than the minimum number of instances for child node when the node was split.
- The best splitting criteria is less than a threshold [29].

2.5 Evaluation methodology

The accuracy measure is used to estimate the overall success rate of classification. The accuracy is defined as (5) [30].

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100 \tag{5}$$

where TP is the number of positive instances which are classified as positive, TN is the number of negative instances which are classified as negative, FP is the number of negative instances which are classified as positive, and FN is the number of positive instances which are classified as negative.

3. Results and Discussion

The decision tree algorithm was used to classify both datasets. The results of the classifications using knowledge in the ontology are given in this section.

3.1 Relationships between attributes

The purpose of this experiment was to examine the effect of the data dimensionality and the relationships between data on the classification. One indicator used to estimate the performance of algorithms is the time complexity of the decision tree algorithm which is estimated according to the number of attributes and the size of training set. For instance, $O(m \cdot n^2)$ is the time complexity of C4.5 algorithms [31] and $O(m \cdot n)$ is the time complexity of Su and Jang's algorithm [32], where m denotes the size of the training dataset and n is the number of attributes. Thus, the decrease in the number of attributes used for tree induction can improve the performance of the algorithm.

To test the efficiency of the classification, we prepared the soybean dataset by removing all records with missing data leaving 562 records remaining. Chi-square tests were used to identify the relationships between attributes and soybean diseases by considering the p-values. If the *p-value* of each attribute is greater than 0.05, we concluded that this attribute is not associated with the diseases. Cramer's V value was then used to measure the strength of the association between the attributes and the diseases. The results of the Chi-square test are shown in Table 2. There were four unrelated attributes, indicated by the *bold-italic* letters, including hail, crop-hist, germination, and roots. These attributes were excluded, and then the number of remaining attributes of the soybean dataset was 31.

Table 2. The measurement of the association between attributes using Chi-square and Cramer's V of soybean dataset

Attribute	χ^2	*p*-value	Cramer's V	Association Level	Attribute	χ^2	*p*-value	Cramer's V	Association Level
date	188.18**	0.00	0.334	●	stem	227.02**	0.00	0.636	●
plant-stand	70.76**	0.00	0.355	●	lodging	94.14**	0.00	0.409	●
precip	149.78**	0.00	0.365	●	stem-cankers	594.05**	0.00	0.727	●
temp	176.80**	0.00	0.397	●	canker-lesion	374.16**	0.00	0.471	●
hail	***3.02***	***0.39***	***0.073***	□	fruiting body	148.06**	0.00	0.513	●
crop-hist	***4.39***	***0.88***	***0.051***	□	external decay	89.03**	0.00	0.398	●
area-damage	121.63**	0.00	0.269	●	mycelium	79.15**	0.00	0.375	●
severity	198.04**	0.00	0.42	●	int-discolor	274.74**	0.00	0.699	●
seed-tmt	12.66**	0.04	0.106	◆	sclerotia	118.01**	0.00	0.458	●
germination	***3.85***	***0.70***	***0.059***	□	fruit-pods	595.80**	0.00	0.728	●
plant-growth	256.60**	0.00	0.676	●	fruit-spots	568.06**	0.00	0.58	●
leaves	149.96**	0.00	0.516	●	seed	41.65**	0.00	0.272	●
leafspots-halo	416.37**	0.00	0.609	●	mold-growth	39.98**	0.00	0.267	●
leafspots-marg	415.72**	0.00	0.608	●	seed-discolor	49.24**	0.00	0.296	●
leafspot-size	431.18**	0.00	0.619	●	seed-size	85.72**	0.00	0.391	●
leaf-shread	58.19**	0.00	0.322	●	shriveling	122.89**	0.00	0.468	●
leaf-malf	10.97**	0.02	0.14	◆	***roots***	***4.02***	***0.21***	***0.085***	□
leaf-mild	21.65**	0.00	0.196	○					

Note: ** p-value < 0.05 , ● Very Strong, ○ Strong, ◆ Moderate, □ Weak

The records with missing values in the census bureau dataset were also removed. Then, biserial correlation and Chi-square tests were used to measure the association between each attribute and the defined classes of this dataset. The results of the biserial correlation are shown in Table 3 and the Chi-square tests are shown in Table 4. The result indicated that all attributes correlated with the assigned classes, so all attributes of the census bureau dataset were used for the analysis process.

Table 3. The measurement of the association between attributes using biserial correlation of census bureau dataset

Attribute	r_{pb}	*p*-value	r_b
age	0.24**	0.00	0.33
capital-gain	0.22**	0.00	0.30
capital-loss	0.15**	0.00	0.20
education-num	0.34**	0.00	0.45
hour-per-week	0.23**	0.00	0.31

Note: ** p-value < 0.05

Table 4. The measurement of the association between attributes using Chi-square and Cramer's V of census bureau dataset

Attribute	χ^2	*p*-value	Cramer's V	Association Level
workclass	804.20**	0.00	0.16	○
education	4070.91**	0.00	0.37	●
marital-status	6061.30**	0.00	0.45	●
occupation	3687.30**	0.00	0.35	●
relationship	6233.43**	0.00	0.46	●
race	304.28**	0.00	0.10	◆
sex	1416.52**	0.00	0.22	○
native-country	317.74**	0.00	0.10	◆

Note: ** p-value < 0.05 , ● Very Strong, ○ Strong, ◆ Moderate, □ Weak

The results of the classification of soybean diseases that used different numbers of attributes are shown in Table 5. The accuracy of the model using only some related attributes was higher than the accuracy of the model that used all attributes in the soybean dataset, increasing by 1.78% from 89.94% to 91.72%. The processing time of the classifications also reduced when only related attributes were used. The processing time for analysis of all attributes was 0.0217 seconds, which decreased by 0.0012 seconds to 0.0205 seconds for the analysis of related attributes only.

Table 5. The results of decision tree classification that used different numbers of attributes

Results	All attributes	Only some related attributes
Numbers of attribute	35	31
Accuracy	89.94%	91.72%
Processing time (Seconds)	0.0217	0.0205

The results in Table 5 illustrate that knowing the relationships between data can help to improve the classification performance by identifying the irrelevant attributes which these attributes will be eliminated from the dataset. This lowers the processing time of data mining when analyzing the reduced dataset.

We compared our approach to the Recursive Feature Elimination (RFE) [33] in terms of the elimination of irrelevant attributes. RFE is a well-known wrapper approach for feature selection. Even though Principal Component Analysis (PCA) [34] is a familiar method to reduce the dimension of a dataset, it was not used to compare with our approach because it does not eliminate the attributes in the dataset. On the other hand, PCA will create a new feature set that consists of several inter-correlated input features for dimensionality reduction. The comparative result of our approach and RFE method is shown in Table 6. The results showed that our approach and RFE perform insignificantly different on irrelevant attributes identification and accuracy. The number of related attributes when using the RFE was slightly lower (29 attributes) than that when using our proposed approach (31 attributes). The classification accuracies of both datasets when using the attributes derived from the RFE method were not significantly different (about 0.59% and 0.66% for the soybean and census bureau dataset respectively) compared to the accuracies of our approach. However, the major drawbacks of the wrapper methods are that they required computations to acquire the feature subset, and the classifier used to obtain the feature subset will tend to overfit [35]. To avoid these problems, the use of biserial correlation and Chi-square as a method to identify the unrelated attributes could help to obtain the number of related attributes and accuracy that close to those obtained from the existing feature selection methods.

Table 6. The comparative result of two different feature selection method

Dataset	Our approach		RFE	
	Related attributes	**Accuracy**	**Related attributes**	**Accuracy**
Soybean	31	91.72%	29	92.31%
Census Bureau	13	80.72%	12	81.38%

3.2 Using knowledge in an ontology to assist the decision tree algorithm

The purpose of this experiment was to examine the use of abstract values (inferred data) from an ontology in the classification process of the decision tree algorithm. The data in the datasets are mapped with an ontology for inferring the related concepts/abstraction of each value.

As shown in Table 7, the inferred concepts from the soybean disease ontology were the values of three attributes including stem-canker, fruit-pods, and fruit-spots. Furthermore, the inferred concepts from the personal information ontology were the values of four attributes of the census bureau dataset including age, education, marital-status, and native-country. These inferred concepts will replace the primitive values of each related attribute in the dataset. Next, the datasets with the abstract values are used as the input to the decision tree for identifying the soybean's disease and the class of the U.S income. The accuracy was used to measure the performance of the classification. Also, the depth of the tree was used to estimate the classification efficiency because the depth of the tree is one metric to measure the tree complexity.

Table 7. The abstract values inferred from the ontologies

Dataset	Attribute	Primitive Data	Abstract Data
Soybean	stem-canker	below-soil, above-soil	near ground
		above-sec-nde	canker on stem
	fruit-pods	diseased, few-present	presented symptom on fruit pod
	fruit spots	colored, brown-w/blk-specks	colored fruit spots
Census Bureau	age	Numerical values	< 16, 16-19, 20-24, 25-34, 35-44, 45-54, 55-64, > 64
	education	Preschool	Preschool
		1st-4th, 5th-6th	Primary Education
		7th-8th, 9th, 10th, 11th, 12th,HS-grad	Secondary Education
		Bachelors, Some-college, Prof-school, Assoc-acdm, Assoc-voc, Masters, Doctorate	Post-Secondary Education
	marital-status	Never-married, Widowed, Divorced	Single
		Married-civ-spouse, Separated, Married-spouse-absent, Married-AF-spouse	Married
	native-country	Cambodia, India, Japan, China, Iran, Philippines, Vietnam, Laos, Taiwan, Thailand, Hong Kong	Asia
		England, Germany, Greece, Italy, Poland, Portugal, Ireland, France, Hungary, Scotland, Yugoslavia, Holand-Netherlands.	Europe
		United-States, Puerto-Rico, Canada, Outlying-US(Guam-USVI-etc), Cuba, Honduras, Jamaica, Mexico, Dominican-Republic, Haiti, Guatemala, Nicaragua, El-Salvador, Trinadad&Tobago,	North America
		Ecuador, Columbia, Peru	South America

The results of the classification are indicated in Table 8. When the dataset with the abstract values was used, the accuracy of the classification increased from 91.72% on original dataset of the soybean dataset to 92.31% on the revised dataset. The maximum depth of the tree of classification on both the primitive soybean dataset and the soybean dataset with abstract values was 14. The depth of tree is equivalent on both soybean datasets because there were few differences in the attribute values in the original dataset and the revised dataset. To illustrate, as shown in Table 7, the values of the attribute stem-canker were changed from three values (below-soil, above-soil, and above-sec-nde) into two values (near ground, and canker on stem), so the frequency of the observed values that were used as criteria for tree growth of the revised dataset was similar to the original dataset, and it might not be enough to effect the tree's depth. For the classification on the census bureau data, the accuracy of the original census dataset was 80.72% which rose to 82.31% when the related abstract values were used. On the contrary, the maximum depth of the revised dataset was lower than the original dataset because there are more differences in the attribute values of both datasets. As presented in Table 7, the variety of primitive values could be categorized into small groups, for instance, the attribute named native-country showed 41 countries, and these countries could be categorized into four groups: Asia, Europe, North America, and South America. By categorizing the attribute values in this way reduces the variety of attributes values resulting in, for the tree construction process, the tree growing phase terminating

faster because of the fewer attributes values to consider, and identifying which of these values belongs to which class [29]. Therefore, the size of the decision tree was smaller when using the dataset with abstract values. The maximum tree depth for the primitive value was 42, while when the census dataset with the abstract values was used, the highest tree's depth reduced by 4 to 38.

As the results have shown, the use of abstract values in an ontology for the decision tree algorithm affects the classification performance. The inference engine of an ontology was used to infer the related concepts of each value, and these related concepts were used to transform the primitive data into the general concepts that helped to reduce the wide variety of attribute values. The variety of attribute values affected the depth of the tree because one stopping criterion of the tree growing phase was when it was considered that all attribute values belong to an appropriate class. If the values of the attributes were different, the tree depth was deeper and had a higher time complexity. This would tend to produce poor classification performance.

Table 8. The classification accuracies and tree's depth

Dataset	Primitive Data Classification		Abstract Data Classification	
	Accuracy	Maximum Depth	Accuracy	Maximum Depth
Soybean	91.72 %	14	92.31 %	14
Census Bureau	80.72%	42	82.31%	38

3.3 Parameter tuning for the optimal results

The purpose of this experiment was to identify the optimal classification results when using the abstract values. The grid search technique [36], which is a method to find the best parameters for the classification algorithm, was used to identify the optimal tree depth for obtaining the best performance of the classification. The important parameter of a decision tree is the maximum depth of the tree. We varied the maximum depth parameter from 1, 2, 3, ...,n where n is the maximum number of tree depth the decision tree algorithm can construct. The results of the classification with parameter tuning are presented in Table 9.

Table 9. The accuracies and tree's depth when applied parameter tuning

Dataset	Primitive Data Classification		Abstract Data Classification	
	Accuracy	Optimal Depth	Accuracy	Optimal Depth
Soybean	91.72%	12	92.31%	12
Census Bureau	84.80 %	10	84.83 %	8

The value of the optimal tree depth derived from the grid search technique of the original soybean dataset and the soybean dataset with abstract values, was 12, and these parameters were used to determine the optimal performance of the model. The accuracy when using the optimal tree's depth as a parameter in the decision tree algorithm was 91.72% for classifying the original soybean dataset and was 92.31% for analysis on the soybean dataset with abstract values.

The grid search technique was also used to estimate the optimal depth of the tree to acquire the optimal performance of the household income classification. The grid search parameters were set from 1, 2, 3, ...,n where n is the maximum number of tree depth the decision tree algorithm could be constructed. The optimal tree depth for analysis of the original census dataset was 10 and for the revised census dataset, 8. The accuracy of the classification with parameter tuning was 84.80% for the classification of the original dataset and 84.83% for the classification of dataset with related abstraction.

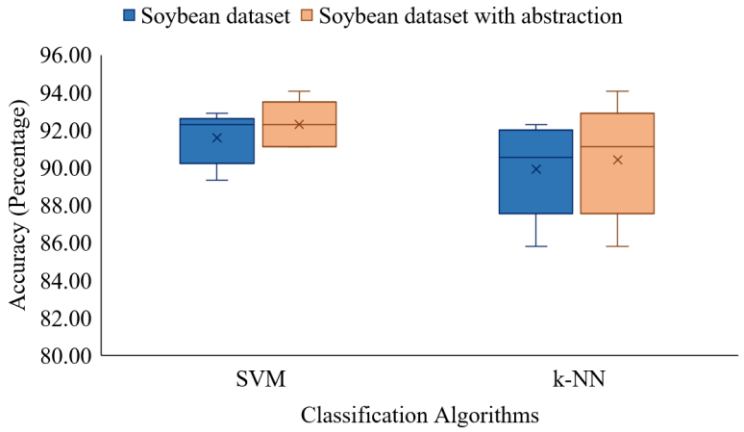
The results in this experiment showed that the use of abstract values did help to obtain efficient classification performance when parameter tuning was applied. Lowering the depth of the tree when using the abstract values was helpful to data scientists when considering the decision rules.

3.4 Using abstract values with various classification algorithms

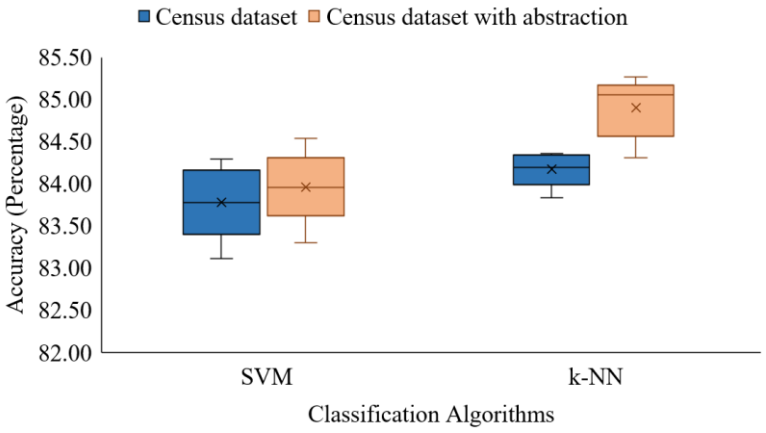
The purpose of this experiment was to examine the use of abstraction values from an ontology in various classification algorithms. Initially, Support Vector Machine (SVM) [27], a supervised learning algorithm that uses the hyperplane, was used to perform the classification. This algorithm requires a small number of samples for training. It provides high accuracy, but the performance depends on parameter selection. In this experiment, SVM with a linear kernel was used for the classification process. Also, k-Nearest Neighbor (k-NN) [27] was applied in this experiment. The classification process of the k-NN algorithm is based on the similarity between sample data. The similarity between data is measured by computing the distance between them, and then the data would be grouped into the nearest class. The performance of k-NN is dependent on the size of the data and the number of 'k,' which does not have a principle of selection. Also, in this experiment, both the soybean dataset and the census dataset were used to evaluate the performance of various algorithms. For the classification of the soybean dataset, the parameter *k* of *k*-NN was set to 2, and the parameter *C* of SVM was set to 1. For the analysis of the census bureau dataset, the parameter *k* of *k*-NN was set to 10, and the parameter *C* of SVM was also set to 1.

The results of the soybean classification process using the different algorithms are presented in Figure 4(a). The accuracy of all algorithms that used the data from an ontology was higher than those used in the original dataset. The SVM's average accuracy increased from 91.60% to 92.31% when analyzing the dataset using the ontology knowledge. Also, the average accuracy of *k*-NN increased from 89.94% to 90.41% when analyzing the dataset with abstraction values.

The results of the household income classification with various classification algorithms are shown as Figure 4(b). The accuracy of the classification of both the primitive census dataset and the revised census dataset also increased. When using the SVM algorithm, the average accuracy of the original census dataset classification was 83.78%, and the average accuracy improved minimally to 83.96% for the classification of the revised census dataset. Furthermore, the average accuracy of the *k*-NN algorithm on the analysis of the primitive census dataset was 84.18%, and the accuracy also climbed marginally to 84.91% when the abstraction data were used. The results of this experiment showed that the abstract values derived from an ontology could improve the classification results of various classification algorithms [13, 14]. For the soybean disease classification, the accuracy of the *k*-NN was lower than the accuracy of SVM. In contrast, the classification accuracy of *k*-NN on the census dataset was greater than the accuracy of the SVM algorithm. Since the performance of the SVM depend on the parameter selection, and the *k*-NN's performance depend on the good value of '*k*' and the size of data [27], so the accuracies of SVM and *k*-NN on both dataset could be in different patterns.



(a)



(b)

Figure 4. The Accuracy of various classification algorithms that used the primitive dataset and the dataset with abstract values

3.5 Algorithm complexity

In this section, we provide the details of the process to identify the abstract value of each attribute in the dataset using a domain ontology. The algorithm to define the ontology's concept related to the attribute's value is shown as Algorithm 1, and the process of replacing the primitive values with the abstract values is shown as Algorithm 2.

In Algorithm 1, the hierarchical structure of an ontology was utilized to identify the related abstract value of each attribute. The concepts (class) in the ontology were mapped with the attribute's name. If the attribute's name matches the parent class, the attribute's value is used to identify the child class of which this value is a member. Then, this identified class is used as the abstract value for replacing the primitive values in the dataset.

Algorithm 1: Abstract value identification.

Input: A list of classes in an ontology ($\{C\}$), an attribute in dataset (a_i), and a value of attribute (v_{ai})
Output: an abstract value

```
1  FOR each class c_i where c_i ∈ C
2      IF attribute a_i match with parent class of c_i
3          IF list of instance of c_i ({I}) exist
4              FOR each instance ins_i where ins_i ∈ I
5                  IF value of attribute v_ai match with instance ins_i
6                      RETURN c_i
7                  ENDIF
8              ENDFOR
9          ENDIF
10     ENDIF
11 ENDFOR
```

Algorithm 2 shows that all attributes of a dataset, their values and the classes of an ontology, were loaded to Algorithm 1 for determining the abstract data. The attributes values that could be matched with the concept in the ontology were duplicated. Then, the inferred data obtained from Algorithm 1 were substituted for the primitive value in each duplicate attribute. Finally, this transformed dataset was used as input for various classification algorithms.

Algorithm 2 : Replacing the primitive data

```
     Input : sample dataset (S), an ontology (O)
     Output : a dataset with abstract data
1    Initial the empty set {Ab} for abstract data
2    Find the list of class {C} of ontology
3    // Identify the abstract value of each attribute's value
4    FOR each attribute a_i where a_i ∈ S
5        FOR all unique values v_ai of attribute a_i
6            InferClass = call Algorithm1({C}, a_i, v_ai)
7            update {Ab} with InferClass
8        ENDFOR
9    ENDFOR
10   //Update the dataset with the abstract values
11   FOR each attribute a_i where a_i ∈ S
12       Count_Infer = count the number of the abstract values of attribute a_i
13       IF Count_Infer is greater than zero
14           Duplicate attribute a_i as new attribute a_infer
15       ENDIF
16       FOR each row of dataset
17           Update attribute a_infer with {Ab}
18       ENDFOR
19   ENDFOR
```

For defining the computational complexity of our algorithm, we determine the worst-case execution time of the algorithm for any input of size n. As shown in Algorithm 1, the outer FOR loop will execute 2n times and the inner FOR loop will execute n times. Therefore, Algorithm 1 would require time to run as shown in (6).

$$T_1 = 2n \times n = 2n^2 \tag{6}$$

As shown in Algorithm 2, lines 1 and 2 are the simple statements which executed at once. The outer FOR loop (lines 4 to 9) would execute *n* times. In the inner FOR loop (lines 5 to 8), Algorithm 1 is executed *n* times, so the total number of times that two loops execute are defined as in (7).

$$T_2 = 2 + n \times (n \times T_1) = 2 + 2n^4 \tag{7}$$

For Algorithm 2, lines 10 to 19, the nested FOR loop will require time to execute as shown in (8).

$$T_3 = n \times (2 + n) = 2n + n^2 \tag{8}$$

Therefore, the total time of our algorithm is shown as (9).

$$T_{total} = T_2 + T_3 = 2n^4 + n^2 + 2n + 2 \tag{9}$$

We conclude that the complexity of our algorithm is $O(n^4)$ where n^4 is the highest order of growth of a function. This could be considered as our main limitation compared to other approaches. For example, the classical decision tree has $O(m \cdot n^2)$ while SVM [37] and *k*-NN [38] have $O(n^2)$. However, this is a trade-off between performance and complexity and the approach should be carefully selected to fit to the work.

4. Conclusions

To enhance the quality of data for the classification task, we proposed an approach that uses an ontology as the background knowledge to assist the data preparation process. For data quality improvement, the records with missing values were eliminated, and biserial correlation, Chi-square and Cramer's V were used to measure the relationships between attributes and the defined classes. The use of the associated attributes in the classification process improved the classification accuracy by reducing the number of attributes used for the decision tree construction. The reduction of the number of used attributes means the time complexity of the algorithm can be reduced because the number of attributes are one metric that has been used to compute the time complexity with Big O notation [32].

After measuring the association between attributes, the soybean ontology was used to infer the related concepts between data. The values in the soybean dataset were substituted by the related concepts from the ontology. The used of abstraction from the ontology improvs the performance of the decision tree algorithm by narrowing the variety of attribute values. The consideration of all attribute values into a single class is one stopping criteria of the decision tree growing phase, so if there are quite different numbers of attributes, the model might be complex. Furthermore, when the abstract values are used in the classification process, the depth of the tree is reduced. Since the depth of the tree is one metric to measure the tree complexity, the lowering of the tree's depth can affect the performance of the decision tree [39, 40]. Also, the notion of abstraction could be applied to the other algorithms such as SVM, *k*-NN because the classification accuracy obtained from each algorithm could improve when the dataset with abstract values were used.

In conclusion, the use of abstract values in the classification task enables better performance and allows the data scientist to view the data in more meaningful ways. However, the level of concepts in then hierarchy used for generalizing primitive data is an important aspect to consider when using this technique. If concepts used are high in the hierarchy, the data may be more general and lead to missing significant information.

There are still opportunities for the researcher to apply an ontology in the data mining process to improve the classification efficiency, such as the use of an ontology to adjust the classification algorithms and to assist the post-processing process. In the future, our work will be continued to improve the performance of the decision tree by using the knowledge in ontology as the criteria to assist the node selection of the decision tree induction process. Also, when the new incoming data are loaded to an existing decision tree, the knowledge in ontology will be used to consider which node of the current decision tree could be modified to obtain better performance.

5. Acknowledgements

This work was supported in part by a grant from Department of Computer Science and Information Technology and Faculty of Science, Naresuan University (Grant no: R2562E027). We also thank Mr. Roy I. Morien of the Naresuan University Graduate School for his assistance in editing the English grammar and expression in this paper.

References

[1] Hand, D.J., 2007. Principles of data mining. *Drug-Safety*, 30(7), 621-622.

[2] Dou, D., Wang, H. and Liu, H., 2015. Semantic data mining: A survey of ontology-based approaches. *Proceedings of the 2015 IEEE 9th International Conference on Semantic Computing*. Anaheim, CA, USA, February 7-9, 2015, 244-251.

[3] Anand, S.S., Bell, D.A. and Hughes, J.G., 1995, The Role of Domain Knowledge in Data Mining, *Proceedings of the 4th International Conference on Information and Knowledge Management*, Baltimore, Maryland, USA, November, 1995, 37-43.

[4] Kuo, Y.-T., Lonie, A., Sonenberg, L. and Paizis, K., 2007. Domain ontology driven data mining: A medical case study. *Proceedings of the 2007 International Workshop on Domain Driven Data Mining*, San Jose, California, USA, August 12, 2007, 11-17.

[5] Staab, S. and Studer, R., 2009. *Handbook on Ontologies*. Heidelberg: Springer Science & Business Media.

[6] Marinica, C. and Guillet, F., 2010. Knowledge-based interactive postmining of association rules using ontologies. *IEEE Transactions on Knowledge and Data Engineering*, 22(6), 784-797.

[7] Asadifar, S. and Kahani, M., 2017. Semantic association rule mining: A new approach for stock market prediction. *Proceedings of the 2nd Conference on Swarm Intelligence and Evolutionary Computation*, Kerman, Iran, March 7-9, 2017, 106-111.

[8] Benites, F. and Sapozhnikova, E., 2014. Using semantic data mining for classification improvement and knowledge extraction. *Proceedings of the LWA 2014 Workshops*, Aachen, Germany, September 8-10, 2014, 8-10.

[9] Effati, M. and Sadeghi-Niaraki, A., 2015, A Semantic-based classification and regression tree approach for modelling complex spatial rules in motor vehicle crashes domain. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 5(4), 181-194.

[10] Wang, H., Azuaje, F. and Bodenreider, O., 2005. An ontology-driven clustering method for supporting gene expression analysis. *Proceedings of the 18th IEEE Symposium on Computer-Based Medical Systems*, Dublin, Ireland, June 23-24, 2005, 389-394.

[11] Trappey, A.J.C., Trappey, C.V., Hsu, F. and Hsiao, D.W., 2009. A fuzzy ontological knowledge document clustering methodology. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39(3), 806-814.

[12] Tang, A. and Fong, S., 2010. A taxonomy-based classification model by using abstraction and aggregation. *Proceedings of the 6th International Conference on Advanced Information Management and Service*, Seoul, South Korea, November 30-December 2, 2010, 448-454.

[13] Zhang, J., Silvescu, A. and Honavar, V., 2002. Ontology-driven induction of decision trees at multiple levels of abstraction. *Proceeding of International Symposium on Abstraction, Reformulation, and Approximation*, Kananaskis, AB, Canada, August 2-4, 2002, 316-323.

[14] Vieira, J. and Antunes, C., 2014. Decision tree learner in the presence of domain knowledge. *Proceedings of Chinese Semantic Web and Web Science Conference*, Wuhan, China, August 8-12, 2014, 42-55.

[15] Dua, D. and Karra Taniskidou, E., 2017. *UCI Machine Learning Repository*. [online] Available at: http://archive.ics.uci.edu/ml
[16] Knublauch, H., Fergerson, R.W., Noy, N.F. and Musen, M.A., 2004. The Protégé OWL Plugin: An open development environment for semantic web applications. *Proceedings of the Semantic Web*, Hiroshima, Japan, November 7-11, 2004, 229-243.
[17] Crop Ontology Curation Tool, 2011. *Soybean Ontology*. [online] Available at: http://www.cropontology.org/ontology/CO_336/Soybean.
[18] Jearanaiwongkul, W., Anutariya, C., and Andres, F., 2018. An ontology-based approach to plant disease identification system. *Proceedings of the 10th International Conference on Advances in Information Technology*, Bangkok, Thailand, December 10-13, 2018, 1-8.
[19] Markell, S. and Malvick, D., 2018. *Soybean Disease Diagnostic Series-Publications*. [online] Available at: https://www.ag.ndsu.edu/publications/crops/soybean-disease-diagnostic-series.
[20] Michalski, R. S., 1980. Learning by being told and learning from examples: An experimental comparison of the two methods of knowledge acquisition in the context of development. An expert system for soybean disease diagnosis. *International Journal of Policy Analysis and Information Systems*, 4(2), 125-161.
[21] Kargioti, E., Kontopoulos, E. and Bassiliades, N., 2009. OntoLife: An ontology for semantically managing personal information. *Proceedings of Artificial Intelligence Applications and Innovations III*, Thessaloniki, Greece, April 23-25, 2009, 127-133.
[22] Baraldi, A.N. and Enders, C.K., 2010. An introduction to modern missing data analyses. *Journal of School Psychology*, 48(1), 5-37.
[23] Bedrick, E.J., 2005. Biserial Correlation. In *Encyclopedia of Biostatistics.*
[24] Andy, F., 2000. *Discovering Statistics Using Spss for Windows: Advanced Techniques for the Beginner*. CA.: Sage Publications.
[25] McHugh, M.L., 2013. The Chi-Square test of independence. *Biochemia Medica*, 23(2), 143-149.
[26] Akoglu, H., 2018. User's Guide to Correlation Coefficients. *Turkish Journal of Emergency Medicine*, 18(3), 91-93.
[27] Singh, A., Thakur, N. and Sharma, A., 2016. A review of supervised machine learning algorithms. *Proceedings of the 3rd International Conference on Computing for Sustainable Global Development*, New Delhi, India, March 16-18, 2016, 1310-1315.
[28] Cios, K.J., Pedrycz, W., Swiniarski, R.W. and Kurgan, L.A., 2007. *Data Mining: A Knowledge Discovery Approach*. New York: Springer.
[29] Kotsiantis, S.B., 2013. Decision trees: a recent overview. *Artificial Intelligence Review*, 39(4), 261-283.
[30] EMC Education Services, 2015. *Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data*. Indianapolis: Wiley.
[31] Quinlan, J.R., 1993. *C4.5: Programs for Machine Learning*. San Francisco: Morgan Kaufmann Publishers Inc.
[32] Su, J. and Zhang, H., 2006. A fast decision tree learning algorithm. *Proceedings of the 21st National Conference on Artificial Intelligence*, Boston, Massachusetts, July 16-20, 2006, 500-505.
[33] Guyon, I., Weston, J., Barnhill, S. and Vapnik, V., 2002. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46(1), 389-422.
[34] Abdi, H. and Williams, L.J., 2010. Principal component analysis. *WIREs Computational Statistics*, 2(4), 433-459.
[35] Chandrashekar, G. and Sahin, F., 2014. A survey on feature selection methods. *Computers & Electrical Engineering*, 40(1), 16-28.

[36] Syarif, I., Prugel-Bennett, A. and Wills, G., 2016. SVM parameter optimization using grid search and genetic algorithm to improve classification performance. *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, 14(4), 1502-1509.
[37] Chapelle, O., 2007. Training a support vector machine in the primal. *Neural Computation*, 19(5), 1155-1178.
[38] Cai, Y. and Wang, X., 2011. The analysis and optimization of KNN algorithm space-time efficiency for Chinese text categorization. *Proceedings of Advances in Computer Science, Environment, Ecoinformatics, and Education*, Wuhan, China, August 21-22, 2011, 542-550.
[39] Breiman, L., Friedman, J., Stone, C.J. and Olshen, R., 1984. *Classification and Regression Trees*, Wardsworth, Belmount: Chapman and Hall.
[40] Rokach, L. and Maimon, O., 2005. Top-down induction of decision trees classifiers - a survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 35(4), 476-487.

Current Applied Science and Technology Vol. 20 No. 2 (May-August 2020)

Opinion Mining for Laptop Reviews Using Naïve Bayes

Pakawan Pugsee1* and Thanapat Chatchaithanawat[2]

[1]Innovative Network and Software Engineering Technology Laboratory, Department of Mathematics and Computer Science, Faculty of Science, Chulalongkorn University, Bangkok, Thailand
[2]Stream I.T. Consulting Co., Ltd., Bangkok, Thailand

Received: 6 March 2020, Revised: 9 March 2020, Accepted: 23 March 2020

Abstract

This research is to develop an opinion mining application which allows users to clarify what the reviews on the laptop mentioned. The aim of the research is to analyze user's opinions from laptop reviews on popular online communities. The proposed methodology is composed of four essential processes: preparing data for analysis, detecting subjective text paragraphs, identifying the aspects and classifying the sentiments of text paragraphs. The subjective textual contents are determined by detecting subjective words occurred in the sentences of text paragraphs. Then, only the subjective paragraphs might be classified into specific aspects using comparisons with the vocabularies of aspect domains. Finally, the paragraph sentiments will be categorized into positive or negative opinions using the Naïve Bayes classifier. The experimental results with the performance evaluation showed that the accuracy and precision of the subjective detection of text paragraphs are greater than 90%. In addition, the accuracy and precision of sentiment classification are more than 70%. Therefore, this tool can help consumers in categorizing laptop review paragraphs into aspects and sentiment groups for making selections before purchasing laptops.

Keywords: opinion mining, review analysis, laptop reviews, Naïve Bayes
DOI 10.14456/cast.2020.16

1. Introduction

Currently, the market of portable computer or laptops has become more competitive marketing with tablets and mobile devices. There are several whole manufactures within the portable computer business and that they frequently produce many new laptop models to contend one another. For this reason, consumers have many choices in making decision for buying laptops. Although there are many laptop-review forums on the Internet such as online communities and blogs, customers must take time to scan and explore for too much data. It is very useful if there is a tool that facilitates customers to choose the laptops that they want, gather review information from varied review forums and analyze the helpful data for them.

*Corresponding author: Tel.: +66 22 18 5170 Fax: +66 22 55 2287
E-mail: pakawan.p@chula.ac.th

Opinion mining or sentiment analysis is the methodology that tries to recognize people's mind or opinions by analyzing information from text data, e.g. user comments, blogs, or reviews. One objective of opinion mining is to differentiate the opinion of a supply text into the positive or negative opinions. The opinions or sentiments stated intentions, emotions, decisions, evaluations, needs or desires [1]. Moreover, opinion mining is often used to analyze customer reviews to examine consumers' satisfaction. Opinion mining tools, or sentiment analysis systems can assist users who are customers or consumers to get useful information about interesting products and services. Furthermore, these tools or systems can be used for investigating market trends and surveying customer desires to improve the product qualities and the potency of consumer service.

Most reviews on online communities regarding laptops, e.g. notebookcheck.net, notebookreview.com, cnet.com and laptopmag.com consist of information about laptops in performance, design (style) and features (options). Therefore, this article studied on opinion analysis about laptops' reviews in 3 aspects that are the performance, the design and the features of a product. In addition, this opinion mining tool is the implementation of the framework [2] with a few modifications for improvement.

2. Materials and Methods

2.1 Background knowledge and related works

The goal of opinion mining or sentiment analysis is to distinguish comments or the attitude on various topics in the natural language, so that this analysis can classify the emotional aspects of communication. The research in this field is about grouping of words or messages as the positive attitude or the negative attitude. Some sentences or phrases can express opinions or attitudes, positive or negative. These sentences or phrases also help identify the groups of reviews or comments more easily. Therefore, Pang *et al*. [3] and Turney [4] developed two approaches in the sentiment analysis to identify comment or opinion messages on a social network into the positive or the negative groups.

Sentiment analysis of text statements needs some techniques of natural language processing. Sentiment analysis with natural language processing of product reviews has been utilized in widespread applications to enhance consumer retention and business processes [5]. The natural language processing is the study of computer science, artificial intelligence and linguistics in term of the interaction between humans and computers. It is composed of standard methods to make computers understand natural language or human language involving natural language comprehension and making computers understand human or natural language input. There are three main processes which are syntactic analysis, semantic analysis and pragmatic analysis. First, syntactic analysis will check the grammatical structures and the position of various groups of words that make up the sentence. Secondly, semantic analysis is the accuracy verification in term of the meaning of the sentence. The grammatical sentences normally have the exact meaning. However, some grammatical sentences considered in this field might have ambiguous meaning or no meaning at all. Lastly, pragmatic analysis is the situation needed to be considered to interpret these sentences because sometimes the sentences might not be able to interpret directly. In this case, the sender, the receiver and the content must be in the same situation in order to have the same comprehension. In addition, there are some lexicons containing only sentiment words that are used to classify the sentiments of words in semantic analysis, such as the MPQA (Multi-Perspective Question Answering) subjectivity lexicon [6] and SentiWordNet [7]. The MPQA subjectivity lexicon and SentiWordNet are a publicly available lexical for opinion mining. The MPQA Subjectivity Lexicon can be used to score words or phrases of words to determine whether

they are positive or negative. For every entry, the lexicon creates a result to indicate if an entry is positive, neutral or negative in its opinion. SentiWordNet assigns to each synset of WordNet with three sentiment scores: positivity, negativity, objectivity [8].

Moreover, many machine learning techniques are applied to classify or cluster the sentiments or opinions of text statements. Machine learning is a type of artificial intelligence that makes computers have the self-learning ability [9]. It can be categorized into 2 main types: supervised learning and unsupervised learning. Supervised learning is a learning of the input data in which the answers are already given, such as predicting the sentiment of a sentence by training examples of sentences with their opinion meaning, or the stock price at a particular time. This type of machine learning is prepared for the data prediction involving the problems like regression and classification. Unsupervised learning is a learning of the input data in which the answers are still unknown. The type of machine learning helps us get closer to the answers or understand more problems by arranging the data structure. The model will be prepared to use in the data structure in order to reduce duplication and categorize data into the same group, for example, the problem about clustering.

Furthermore, there are many researches about opinion mining with machine learning techniques in the last few years. For examples, the proposed method in Govindaraj and Gopalakrishnan [5] used acoustic and textual features to analyze opinions on customer product reviews from Amazon product reviews and YouTube. Customer feedback in the form of audio clips (.wave file) was proceeded by speech synthesis tool, speech recognizer and voice-to-text converter before feature selection using hand-coded rules. Acoustic and textual features were calculated and extracted to generate the training data sets for building three classification models with three different feature sets by the support vector machine (SVM) classifier. The opinions of a customer chatting were categorized into 5 levels of sentiment score: extremely positive, positive, neutral, negative and extremely negative. Valdivia *et al.* [10] anticipated an analysis for matching between users' sentiments and automatic sentiment-detection algorithms using TripAdvisor as a resource for sentiment analysis, including the challenges of sentiment analysis and TripAdvisor. The best-known sentiment analysis task aims to observe the sentiments at intervals documents, sentences, or words. This work is often separated into 3 steps: polarity detection (positive, negative, or neutral), aspect extraction (features for organizing the text) and classification (machine learning or lexicon approaches). There are various forms of texts, such as tweets, blog, and reviews. In addition, human language is complicated because of different grammatical rules, cultural variations and jargon in statements. This study obviously expressed the requirement of mining opinions beyond user ratings. Therefore, the implementation of sentiment analysis techniques to extract opinions is crucial to understanding the mind of a traveler and can influence quality improvement in tourism.

Pugsee *et al*. [11] implemented the sentiment analysis application to mine opinion on Twitter messages. Tweets about skin care (with "#skincare") was analyzed by combining word information with the machine learning techniques. SentiWordNet [8] was modified to improve the performance of application for skin care products and two machine learning techniques, i.e. Naïve Bayes and SVMs were implemented to identify the sentiments of messages. The user opinions on tweets were categorized into 5 levels of sentiments: very positive, positive, neutral, negative and very negative. The performance of result classification was evaluated by the accuracy, the precision, and the recall rate that all of their values are more than 75%.

Therefore, our research designed to use SentiWordNet with basic machine learning techniques like a decision tree and Naïve Bayes to implement the opinion mining tool for the laptop reviews. The reasons are that both methods are simple machine techniques that can be implemented and embedded in the software tools easier, including a not too long processing time. Moreover, the sentiment classification with Naïve Bayes in Pugsee *et al.* [11-12] has sufficient performance.

2.2 Opinion mining methodology

The proposed opinion mining tool can provide an organized summary of the product reviews for customers and assist them with the decision making when they want to buy laptop products. The overview of the proposed opinion mining tool following the framework for analyzing laptop reviews [2] with some modification is shown in Figure 1. This tool consists of four main processes: data preparation, subjective detection, aspect identification and sentiment classification.

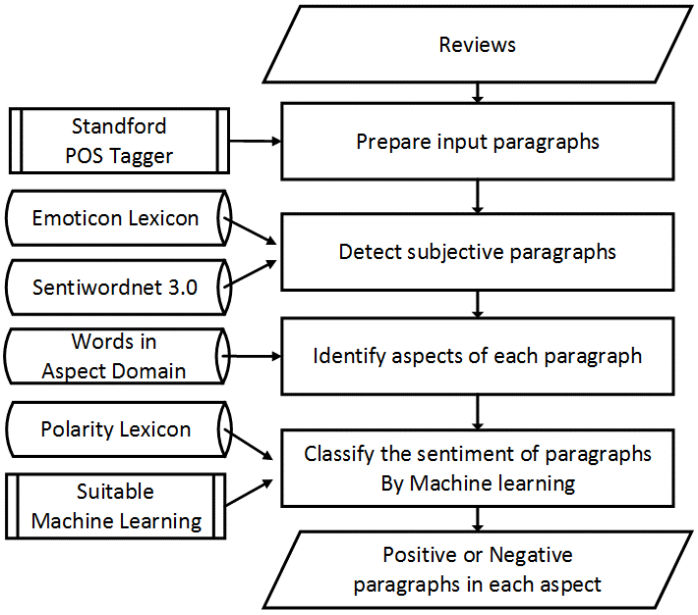


Figure 1. The overview of proposed opinion mining tool

According to Figure 1, the proposed tool is composed of four processes which are to prepare data for analysis, to detect subjective text paragraphs, to identify the aspect of each text paragraph and to classify the sentiments of each text paragraph. The objectives of the opinion mining tool are to categorize content paragraphs in subjective or objective paragraphs, to identify paragraphs' aspects into four aspects and to classify the sentiments of paragraph reviews. The inputs of this tool are the laptop reviews from online communities and the outputs are both groups of text paragraphs that are positive or negative text paragraphs in each aspect domain.

2.2.1 Prepare input paragraphs

This process is implemented based on the technique in Chatchaithanawat and Pugsee [2]. The first step is to delete special characters and symbols in text paragraphs. There are more than 200 special characters and symbols were added from Chatchaithanawat and Pugsee [2], such as other characters not in English alphabets and symbols. In addition, photo and URL links will be deleted from the input reviews. When the photos from reviews in community website are saved in text format it will be saved as [IMG] tag. This process will delete [IMG] tag from the original reviews. Moreover, normal URL links will also be deleted from the reviews by detecting "http" and

"www". Furthermore, picture links will be deleted from reviews by detecting the ".jpg", ".gif" and ".png".

The second step is to tag words with their parts of speech, after separating paragraphs and sentences by a tab character, a full stop, and a newline. Stanford POS Tagger [13] demonstrated to identify words' parts of speech, such as adjectives, adverbs, verbs and nouns. In the next step, the tagged words are changed into the basic forms using WordnetStemmer [14] to manage stemming method. Stemming is to transform into the base form of the focused words by removing the prefix and suffix of the words. The focused words are words in the adjective group, the adverb group, the verb group and the noun group. Figures 2 and 3 present an example of a review paragraph and a prepared paragraph, which are the input and output of this process.

The touchpad is able to recognise even the complex 3-finger gestures with great precision. During about 2 hours of use I've only had 3 times when the mouse didn't do what I was expecting, mostly when trying to select text (which is tricky business on touchpads anyway). I didn't have trouble with palm rejection either, though it might be because my hands don't touch the touchpad while typing :p .

Figure 2. An example of a review paragraph

The_DT touchpad_NN is_VBZ able_JJ to_TO recognise_VB even_RB the_DT complex_NN 3-finger_NN gestures_NNS with_IN great_JJ precision_NN ._.
During_IN about_RB 2_CD hours_NNS of_IN use_NN I_PRP 've_VBP only_RB had_VBN 3_CD times_NNS when_WRB the_DT mouse_NN did_VBD n't_RB do_VB what_WP I_PRP was_VBD expecting_VBG ,_, mostly_RB when_WRB trying_VBG to_TO select_VB text_NN -LRB-_-LRB- which_WDT is_VBZ tricky_JJ business_NN on_IN touchpads_NNS anyway_RB -RRB-_-RRB- ._.
I_PRP did_VBD n't_RB have_VB trouble_NN with_IN palm_NN rejection_NN either_CC ,_, though_IN it_PRP might_MD be_VB because_IN my_PRP$ hands_NNS do_VBP n't_RB touch_VB the_DT touchpad_NN while_IN typing_NN :p_NN ._.

Figure 3. A prepared paragraph

2.2.2 Detect subjective paragraphs

The word information from SentiWordNet [8], which is categorized into the adjective group, and the adverb group, will be useful for identifying whether those words are subjective or objective. Both groups are interesting words in this research and the best information for analyzing subjective statements because most subjective words are in the adjective and the adverb groups. This process detects subjective paragraphs like algorithm described in Chatchaithanawat and Pugsee [2]. Every text paragraph, which has at least one subjective word or emoticon text, will be considered as a subjective paragraph. Additionally, emoticon texts will also be identified by comparing emoticon texts in the paragraph with data from an emoticon lexicon [15] including some emoticons found in experimental data. The steps of this process are shown in Figure 4.

The inputs of this process are prepared paragraphs from the previous process. Then, the emoticons in prepared paragraphs will be detected by comparing found emoticons of text paragraph to the emoticons in the lexicon. If emoticons found match with data in the emoticon lexicon at least one emoticon, those paragraphs will be collected as subjective paragraphs. If there are no detected emoticons in the prepared paragraphs, the subjective words will be detected in this process by comparing with words in SentiWordNet. If the subjective words are found at least one word, the paragraphs will be kept as subjective paragraphs. The emoticons in the lexicon and the subjective words in SentiWordNet are compared to sequential words in the paragraph using unicode matching and string matching, respectively. Consequently, only the subjective paragraphs

are the outputs of this process. The detected subjective words and emoticon texts in the paragraph are exposed in Figure 5.

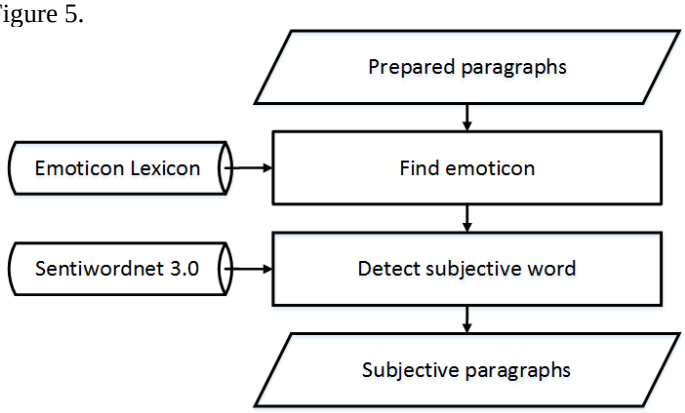


Figure 4. The steps of detecting subjective paragraphs process

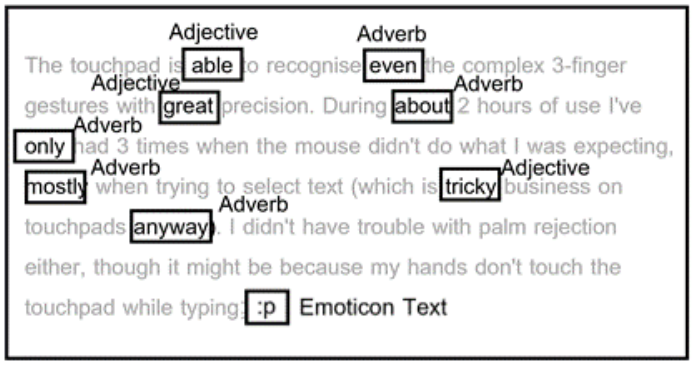


Figure 5. The detected subjective words and emoticon texts in the paragraph

2.2.3 Identify aspects of each paragraph

In this process, the subjective paragraphs from the previous process will be divided into four different aspects ("Performance", "Design", "Feature" and "Other") by comparing words in review paragraphs with words in word lists of three aspect domains. Individual subjective paragraphs can match more than one aspect. Otherwise, some subjective paragraphs that cannot be recognized in previous groups will be identified into "Other" aspect. The words of each aspect domain are listed by analyzing the popular words found in laptop reviews. The steps of this process are shown in Figure 3. The challenge of this process is creating the word lists in each aspect that are useful to categorize the aspect of paragraph correctly. Finding the frequency of all content words in laptop reviews and determining the threshold of the word frequency to count as words in each aspect were proceeded to generate the word lists.

According to Figure 6, this process will detect words in aspect domains for identifying types of review aspects. The examples of words in each aspect domain are shown in Table 1, and the detected words of a paragraph are shown in Figure 7. These words will be collected from all review paragraphs by using AntConc [16], which helps to find the frequency of words in each

paragraph. Then, the aspect words with high frequency from all reviews will be categorized into each aspect domain by the researcher's judgment to classify the aspect of paragraphs.

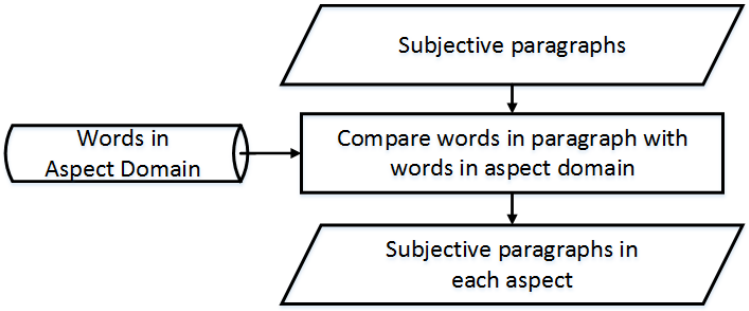


Figure 6. The step of identifying aspects of each paragraph process

Table 1. Examples of words in three aspects

Performance	Design	Feature
battery	display	Bluetooth
CPU	height	camera
GPU	materials	DVI
memory	screen	HDMI
processor	size	touchpad
ram	weight	USB
resolution	width	wireless

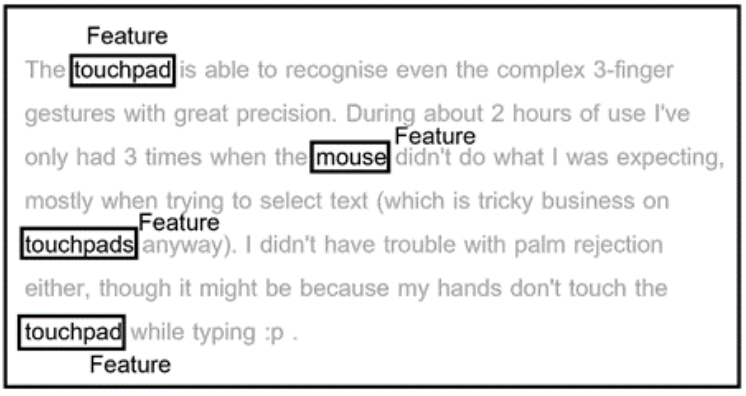


Figure 7. The detected words in the feature aspect

2.2.4 Classify the sentiments of paragraphs

The subjective paragraphs in individual aspect will be classified into the sentiment types of paragraphs by using the selected machine learning technique. The results of this process are two groups of text paragraphs (positive or negative paragraphs). There are 6,234 text paragraphs in experiments and these texts are categorized into 2,534 positive paragraphs and 3,700 negative paragraphs. The selected features of the classification model are all adjectives, adverbs, their parts

of speech, and their polarity score will be learned and classified by WEKA (Waikato Environment for Knowledge Analysis) [17] to choose the suitable feature set and a classifier. WEKA is one of the popular machine learning software implemented in JAVA programming language by the University of Waikato, New Zealand. This software tool is free to use under the General Public License (GPL). It is a collection of machine learning algorithms for data mining tasks, such as data pre-processing, classification, regression, clustering, association rules, and visualization. Our research tried to test on three different feature sets with two machine learning techniques (Naïve Bayes and Decision Tree). Then, Decision Tree (J48) and Naïve Bayes classifier of WEKA are executed to generate the classification models of sentiment analysis. To test the performance of classification, the confusion matrix is applied with labeled the positive and negative paragraphs by manual annotation in order to assess the performance of the classification model with evaluation values. The confusion matrix and three evaluation values are displayed in Table 2.

Table 2. A confusion matrix and evaluation values

Actual class	**Predicted class**		**Accuracy**	**Precision**	**Recall**
	Positive	Negative			
Positive	True positive (TP)	False negative (FN)	TP+TN/ (TP+FN+FP+TN)	TP/ (TP+FP)	TP/ (TP+FN)
Negative	False positive (FP)	True negative (TN)		TN/ (TN+FN	TN/ (TN+FP)

According to Table 2, there are three evaluation values of the performance of classification that are accuracy, precision and recall. The accuracy is calculated from the number of data with the correct prediction comparing to the total number of data. The precision is counted using the number of data with the correct prediction comparing to the number of predicted data in each class, while the recall is calculated by comparing to the number of actual data in each class.

In the first experiment, all adjective and adverb words with their parts of speech will be used in the training data. Figure 8 displays the adjective and adverb words in the paragraph. The percent of accuracy, precision and recall rates will be calculated by confusion matrices. The confusion matrices of the first experiment and the percentage of accuracy, precision, and recall rates are shown in Table 3 and Table 4, respectively.

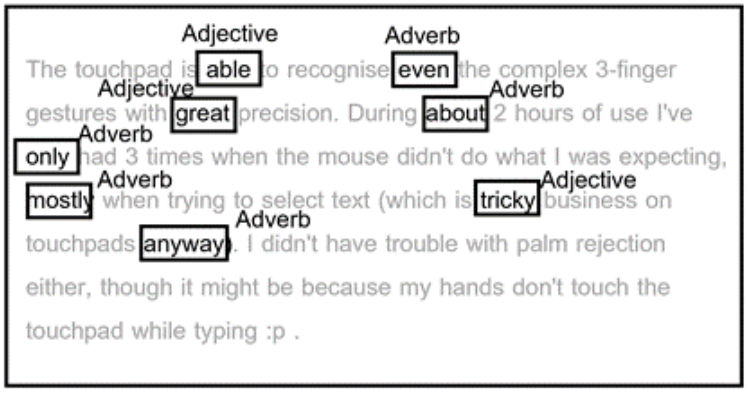


Figure 8. The adjective and adverb words in the paragraph

Table 3. The confusion matrices of results in Experiment I

Actual class		Predicted class			
		Naïve Bayes		J48	
		Positive	Negative	Positive	Negative
Positive	2,534	1,499	1,035	1,695	839
Negative	3,700	922	2,778	821	2,879
Total	6,234	2,421	3,813	2,516	3,718

Table 4. The percentage of accuracy, precision and recall of Experiment I

Classifier		Accuracy	Precision	Recall
Naïve Bayes	Positive	68.61%	61.91%	59.16%
	Negative		72.86%	75.08%
J48	Positive	73.37%	67.37%	66.89%
	Negative		77.43%	77.81%

In the second experiment, the word information from SentiWordNet [8] was modified to create the polarity lexicon which consisted of words and their polarity levels ("strong positive", "positive", "neutral", "negative" and "strong negative"). Some concatenated adjective words with their polarity level are added into the polarity lexicon to enhance the tagged polarity, e.g., high-end, full-colored, and industry-standard. The polarity levels of all adjectives and adverbs are included into the training data with words and their parts of speech from the first experiment. Figure 9 presents the polarity levels of adjectives and adverbs in the paragraph. The confusion matrices of the second experiment and the percentage of accuracy, precision and recall rates are shown in Tables 5 and 6, respectively.

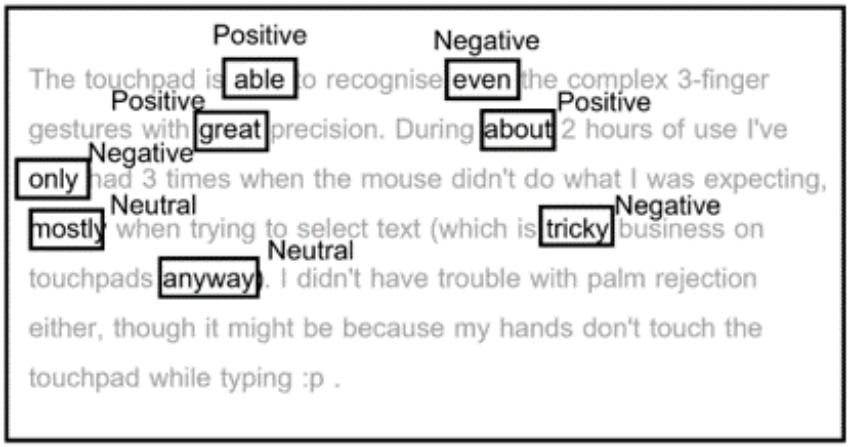


Figure 9. The polarity levels of adjectives and adverbs in the paragraph

Table 5. The confusion matrices of results in Experiment II

Actual class		Predicted class			
		Naïve Bayes		J48	
		Positive	Negative	Positive	Negative
Positive	2,534	1,676	858	958	1,576
Negative	3,700	860	2,840	368	3,332
Total	6,234	2,536	3,698	1,326	4,908

Table 6. The percentage of accuracy, precision and recall of Experiment II

Classifier		Accuracy	Precision	Recall
Naïve Bayes	Positive	72.44%	66.09%	66.14%
	Negative		76.80%	76.76%
J48	Positive	68.82%	72.25%	37.81%
	Negative		67.89%	90.05%

The classification results in Experiment I and Experiment II are different. The performance of the classification model of decision tree technique is higher than those by the Naive Bayes classifier in the first experiment. The reason is that there are various adjective and adverb words found in reviews, so only the probability of words is not sufficient to classify the sentiment, while the decision tree has bias in the majority of data. On the other hand, the classification model of Naive Bayes classifier has capacity more than the classification model of decision tree technique. It is found that the polarity level of words can help improve the sentiment classification performance, but the decision tree is overfitted to the data with bias in the majority class. Therefore, there is a test on the third feature set that there is only the polarity level.

In the third experiment, only polarity levels of words in adjective and adverbs which are strong positive, very positive, positive, neutral, negative, very negative and strong negative will be used in the training data. The confusion matrices and the percentage of accuracy, precision and recall are shown in Tables 7 and 8.

Table 7. The confusion matrices of results in Experiment III

Actual class		Predicted class			
		Naïve Bayes		J48	
		Positive	Negative	Positive	Negative
Positive	2,534	1,907	627	1,811	723
Negative	3,700	781	2,919	830	2,870
Total	6,234	2,688	3,546	2,641	3,593

Table 8. The percentage of accuracy, precision and recall of Experiment III

Classifier		Accuracy	Precision	Recall
Naïve Bayes	Positive	77.41%	70.94%	75.26%
	Negative		82.32%	78.89%
J48	Positive	75.09%	68.57%	71.47%
	Negative		79.88%	77.57%

From three experiments, the results showed that the performance of Naïve Bayes classification (accuracy, precision and recall) is higher than those of J48 classification for the

second and the third experiment. Moreover, the performance of Naïve Bayes classification with the feature set in the third experiment is the highest performance. Furthermore, all evaluation values of the Naïve Bayes classification are higher than those of decision tree technique. Therefore, this research selects the feature set in the third experiment and Naïve Bayes classification to generate classification models and to implement the opinion mining tool.

3. Results and Discussion

3.1 Implementation

This implemented software tool is an easy way to apply the opinion mining methodology for analyzing laptop reviews. The developers have designed the layout of the user interface for this tool as one page to make the software easier to use. The main screen consists of three areas: the menu bar, a middle text area and four bottom text areas as shown in Figure 10. The menu bar includes "Single Review" button for analyzing a review, "Multiple Review" button for analyzing reviews, text box for inputting a filtered word and drop-down list for selecting output types. The middle text area shows only subjective paragraphs in the review and four bottom text areas show the subjective paragraphs in each aspect domain as displayed in Figures 11 and 12.

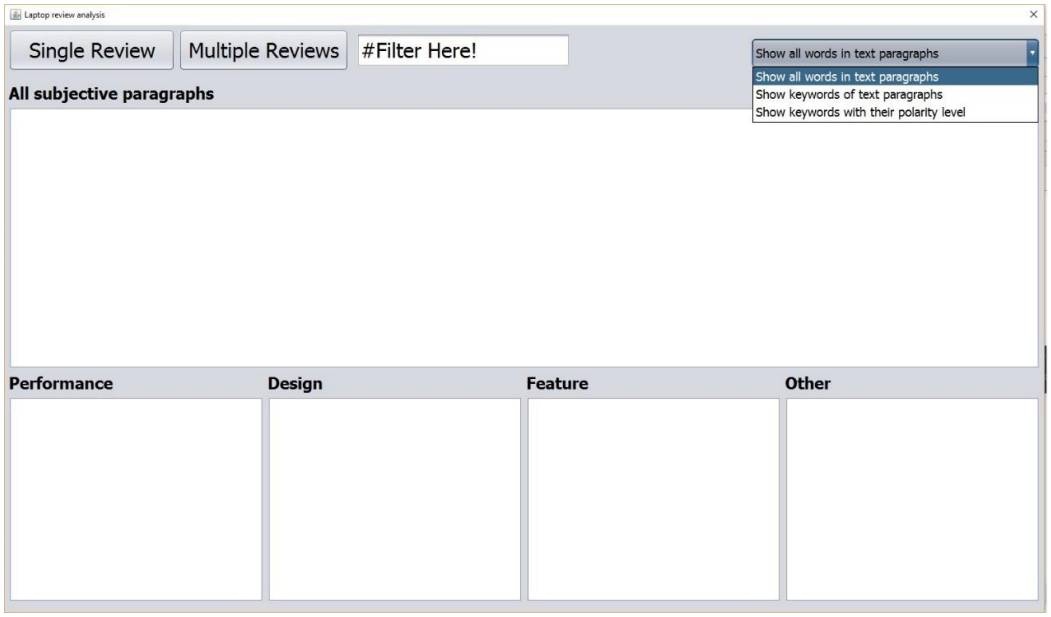


Figure 10. The main user interface of the proposed opinion mining tool

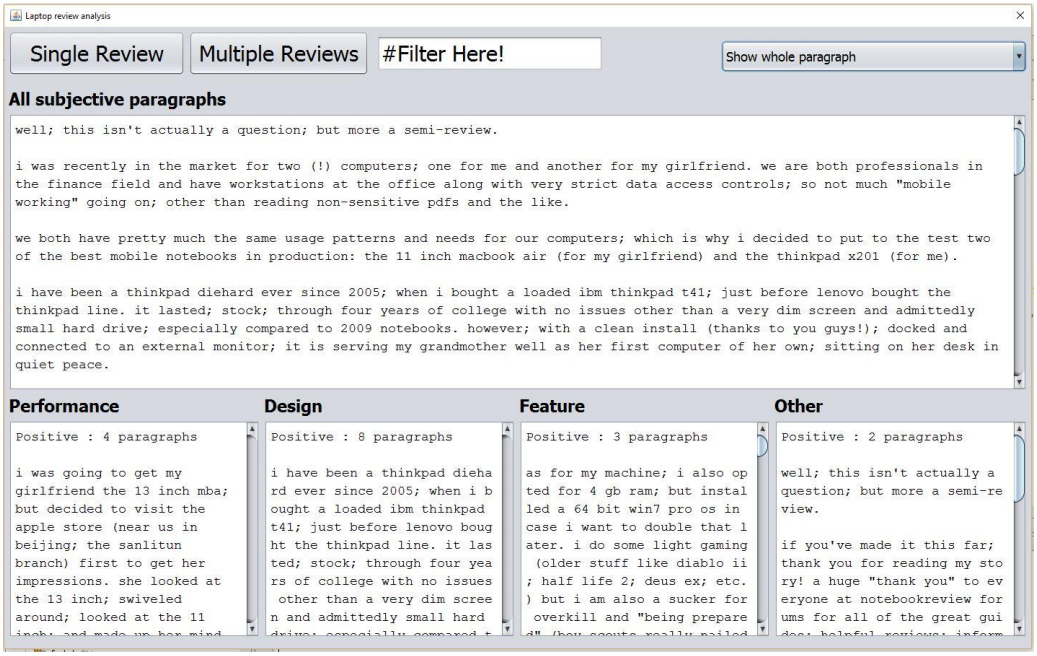


Figure 11. The output interface of the single review

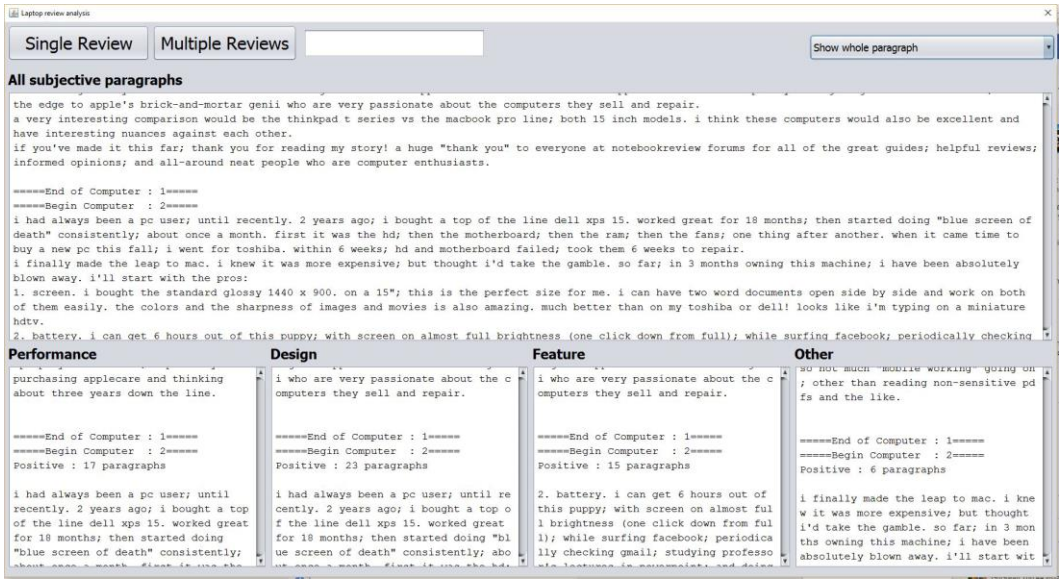


Figure 12. The output interface of the multiple reviews

The main user interface is the first page that the users will see when they start the opinion mining tool. This page can be divided into 2 main functional parts. Single review, the single review button, is used to analyze only one text review in one text file (the output result is shown in Figure 11). Multiple reviews, the multiple reviews button is used to analyze more than one text reviews in one text file (the output result is shown in Figure 12).

Additionally, the output on the screen of this software implementing the opinion mining methodology can be divided into 3 main types: show all words in text paragraphs, show only the keywords of text paragraphs and show only the keywords with their polarity level of text paragraphs. The default output results show all words in paragraphs (Figures 11 and 12). The bottom text area shows all positive paragraphs and all negative paragraphs in the review separated by the aspect domains. For each aspect of text area, there are the total number of positive paragraphs, all positive text paragraphs, the total number of negative paragraphs and all negative text paragraphs, respectively.

The next output results (as shown in Figure 13) display the keywords of the text paragraphs. All text areas show only adjective words, adverb words and some words displaying the aspect of paragraphs instead of all words in paragraphs. These words are the keywords for detecting aspect and the keywords to generate feature sets for classifying sentiments. The last output results (as shown in Figure 14) demonstrate the keywords with their polarity levels. All adjective and adverb words with their polarity levels generated from their polarity scores are displayed, including words in each aspect domain. The keywords' polarity levels are the feature set of the sentiment classification by the machine learning.

Figure 13. The output interface showing only keywords

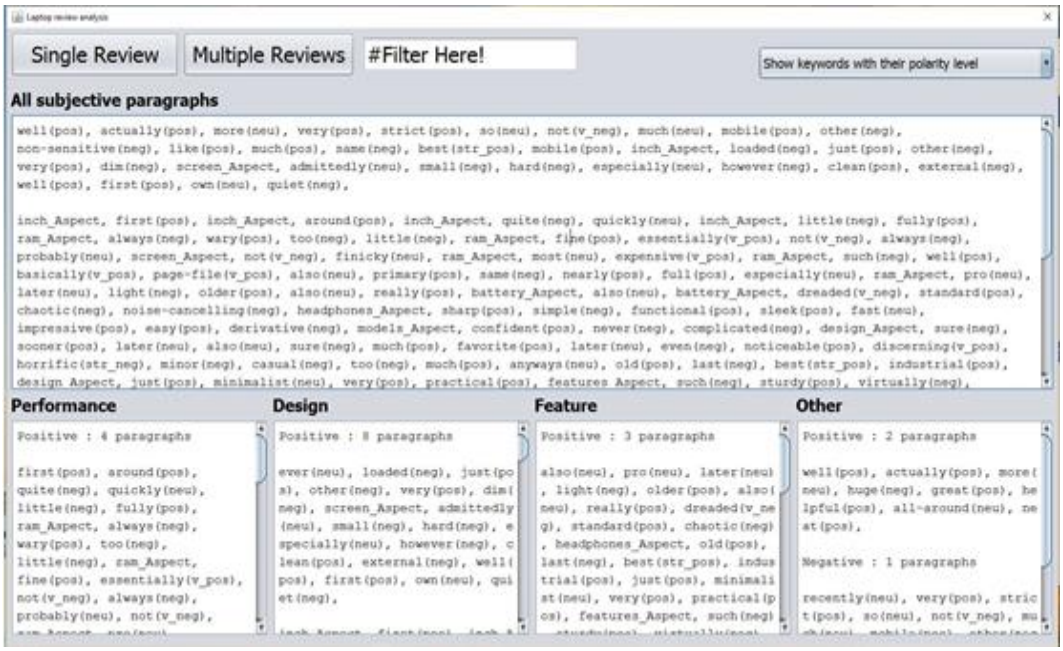


Figure 14. The output interface showing only keywords and their polarity level

3.2 Results and evaluation

In the results and evaluation section, there are 3 parts of the experiment in this research which are subjective detection, aspect identification, and sentiment classification. The results are explained in the form of confusion matrices, and all evaluations of three parts will be measured by calculating accuracy, precision and recall values from the confusion matrices.

3.2.1 Subjective detection

All selected review topics were separated into 15,384 paragraphs by detecting a new line. They can be separated into 6,399 subjective paragraphs and 8,985 objective paragraphs. Words in text paragraphs of the experimental data were compared with words in SentiWordNet and the modified emoticon lexicon to detect subjectivity. The result of the subjective detection is classified into 6,992 subjective paragraphs and 8,392 objective paragraphs. The confusion matrix of the result in the subjective detection is shown in Table 9 and the evaluation values are expressed in Table 10.

Table 9. The confusion matrix of subjective detection

Actual class		**Predicted class**	
		Subjective	Objective
Subjective	6,399	6,367	32
Objective	8,985	625	8,360
Total	15,384	6,992	8,392

Table 10. The percentage of accuracy, precision and recall of subjective detection

Actual class	Accuracy	Precision	Recall
Subjective	95.73%	91.06%	99.50%
Objective		99.62%	93.04%

Referring to Table 10, the accuracy, precision and recall rates are more than 90% for subjective paragraph detection. This means that the opinion mining tool can detect subjective review paragraphs effectively. The reason is that if there is at least one emoticon text or one subjective word in the paragraphs, these paragraphs will correctly be detected subjective paragraphs.

3.2.2 Aspect identification

Subjective paragraphs were identified into Performance, Design, Feature and Other aspects. The subjective paragraphs (6,992 paragraphs) can be divided into 1,347 performance paragraphs, 1,796 design paragraphs and 1,658 feature paragraphs by researchers reading categorizing manually. The confusion matrix of the result in the aspect identification is shown in Table 11.

Table 11. The confusion matrix of aspect identification

Aspect	Actual class	Predicted class	True positive	False positive	True negative	False negative
Performance	1,347	1,489	1,334	155	5,490	13
Design	1,796	2,150	1,737	413	4,783	59
Feature	1,658	1,811	1,614	197	5,137	44

The percentage of three evaluation values for the aspect identification is also shown in Table 12. The percentage of accuracy and recall rate of all aspect domains is greater than 90%, and the precision rate is about 80% or more. As a result, the aspect identification has high accuracy to identify the aspect of collected reviews in this research. This means that the generated aspect word lists are very useful for aspect identification.

Table 12. The percentage of accuracy, precision and recall of aspect identification

Actual class	Accuracy	Precision	Recall
Performance	97.60%	89.59%	99.03%
Design	93.25%	80.79%	96.71%
Feature	96.55%	89.12%	97.35%

3.2.3 Sentiment classification

All paragraphs in each aspect were classified the sentiments of contents by using the Naïve Bayes classifier. Only polarity levels of adjective and adverb words are training data in the sentiment classification. In this case, the objective paragraphs (625 of 6,992 paragraphs) and the neutral paragraphs (133 of 6,992 paragraphs) are removed from the training data, so there are only 6,234 paragraphs in the experiment. The confusion matrix and the percent of accuracy, precision and recall rate of the sentiment classification are displayed in Tables 13 and 14, respectively.

Table 13. The confusion matrix of sentiment classification

Actual class		Predicted class	
		Positive	Negative
Positive	2,534	1,907	627
Negative	3,700	781	2,919
Total	6,234	2,688	3,546

Table 14. The percentage of accuracy, precision and recall of sentiment classification

Actual class	Accuracy	Precision	Recall
Positive	77.41%	70.94%	75.26%
Negative		82.32%	78.89%

Referring to Table 14, the accuracy, precision and recall rates are more than 70% for sentiment classification. This means that the proposed methodology can classify the sentiment of subjective review paragraphs acceptably on collected reviews.

A major limitation of this tool is unseen words which are not included in the polarity lexicon, but may be found on laptop reviews. The reason is that these review analysis methods focus on words and the polarity of words to identify the aspect and to classify the sentiment of review paragraphs. If there are some missing input words from our lexicons, such as words with wrong spelling or technical words, the polarity level finding in the sentiment classification process cannot give the correct polarity level of words. Therefore, the performance of the sentiment classification process will be reduced by this error.

4. Conclusions

Nowadays, many laptops are manufactured with various features. When consumers decide to purchase a laptop, they normally search for laptop reviews in order to get the information first. Moreover, many reviews are created to let the consumers know more about each laptop. For those reasons, this research developed an opinion mining tool which helps users to know what is mentioned in the laptop reviews. The tool is separated into four main processes: to prepare data for analysis, to detect subjective text paragraphs, to identify the aspect of each text paragraph and to classify the sentiments of each text paragraph. The results of performance evaluation show that the subjective detection and the aspect identification has high accuracy and precision, including acceptably accurate and precise sentiment classification.

In conclusion, this opinion mining tool is useful for developing the review analysis system of laptops in order to help consumers gain information before purchasing a laptop. However, the user interface and feature of this tool will be improved in the future, such as data visualization and selected aspect comparison.

5. Acknowledgements

The first author was supported in part by a grant from the Thailand Research Fund and Office of the Higher Education Commission (Grant No. MRG6180250).

References

[1] Karamibekr, M., 2015. *A Sentiment Analysis Framework for Social Issues*. Ph.D., University of New Brunswick, Canada.

[2] Chatchaithanawat, T. and Pugsee, P., 2015. A framework for laptop review analysis. *Proceeding of the 2nd International Conference on Advanced Informatics: Concepts, Theory and Applications*, Chonburi, Thailand, August 19-22, 2015, 1-5.

[3] Pang, B., Lee, L. and Vaithyanathan, S., 2002. Thumbs up? Sentiment classification using machine learning techniques. *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*, Philadelphia Pennsylvania, USA, July 6-7, 2002, 79-86.

[4] Turney, P.D., 2002. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, Philadelphia, Pennsylvania, USA, July 6-12, 2002, 417-424.

[5] Govindaraj, S. and Gopalakrishnan, K., 2016. Intensified sentiment analysis of customer product reviews using acoustic and textual features. *ETRI Journal*, 38(3), 494-501.

[6] Multi-perspective Question Answering, University of Pittsburgh, 2005. *The MPQA Opinion Corpus*. [online] Available at: http://mpqa.cs.pitt.edu/corpora/mpqa_corpus/

[7] Text Learning Group, 2010. *The SentiWordNet SentimentLlexicon*. [online] Available at: http://sentiwordnet.isti.cnr.it

[8] Baccianella, S., Esuli, A. and Sebastian F., 2010. SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. *Proceedings of the* 7th *International Conference on Language Resources and Evaluation*, Valletta, Malta, May 19-21, 2010, 2200-2204.

[9] Mitchell, T., 1997. *Machine Learning*, USA: McGraw-Hill Education -Europe.

[10] Valdivia, A., Luzon, M.V. and Herrera, F., 2017. Sentiment analysis in TripAdvisor. *IEEE Intelligent System*, 32(4), 72-77.

[11] Pugsee, P., Nussiri, V. and Kittirungruang, W., 2019. Opinion mining for skin care products on Twitter. *Communications in Computer and Information Science*, 937, 261-271.

[12] Pugsee, P., Sombatsri, P. and Juntiwakul, R., 2017. Satisfactory analysis for cosmetic product review comments. *Proceeding of 2017 International Conference on Data Mining, Communications and Information Technology*, Phuket, Thailand, May 25-27, 2017, 1-6.

[13] Stanford Natural Language Processing Group, Stanford University, 2008. *Basic English Stanford Tagger*. [online] Available at: https://nlp.stanford.edu/software/tagger.shtml

[14] Massachusetts Institute of Technology and Finlayson, M.A., 2013. *WordnetStemmer*. [online] Available at: https://projects.csail.mit.edu/jwi/api/edu/mit/jwi/morph/WordnetStemmer.html

[15] Yamamoto, Y., Kumamoto, T. and Nadamoto, A., 2014. Role of emoticons for multidimensional sentiment analysis of twitter. *Proceeding of the 16th International Conference on Information Integration and Web-based Applications & Services*, Hanoi, Viet Nam, December 4-6, 2014, 107-115.

[16] Anthony, L., 2019. *AntConc: A Freeware Corpus Analysis Toolkit for Concordance and Text Analysis*. [online] Available at: https://www.laurenceanthony.net/software

[17] Machine Learning Group, University of Waikato, 2019. *Weka: The Workbench for Machine Learning*. [online] Available at: https://www.cs.waikato.ac.nz/ml/weka/

Two Novel Spectrophotometric Methods for Determination of Naproxen via a Modulation to Hydroxy Analog

Hana Sh. Mahmood* and Thura Z. Al-Sarraj

Department of Chemistry, College of Science, University of Mosul, Mosul, Iraq

Received: 2 October 2019, Revised: 7 January 2020, Accepted: 24 March 2020

Abstract

Potassium permanganate was used for oxidation of the synthesized hydroxy analog of Naproxen by two methods. In the first method, the oxidation occurred in acidic medium, the excess of permanganate was followed at 545 nm. Beer's law is obeyed between concentrations of 10 µg/10ml-80 µg/10ml (1-8 ppm) with good sensitivity (molar absorptivity of $9.7x10^3$ l.mol^{-1}.cm^{-1}), good precision (RSD better than ±3.2 %) and high accuracy (relative error better than 1.6%). Sandell's sensitivity index is 0.0237 µg.cm^{-2}. The detection limit (LOD) is 0.075 µg·ml^{-1} and the quantitation limit (LOQ) is 0.25 µg·ml^{-1}. For the second method, potassium permanganate was used to oxidize the hydroxy analog of Naproxen in basic medium, and the manganate produced was followed at 610 nm. The linearity range was from 20 to 70 µg/10 ml (2-7 ppm) with good sensitivity (molar absorptivity of 6.8 x 10^3 l.mol^{-1}.cm^{-1}), good precision (RSD better than ±2.5%) and high accuracy (relative error better than 1.51%). Sandell's sensitivity index is 0.0338 µg.cm^{-2} while the detection limit (LOD) is 0.0098 µg·ml^{-1} and quantitation limit (LOQ) is 0.03296 µg·ml^{-1}. The two methods were applied successfully for the determination of Naproxen after extraction of the active ingredient by ethyl acetate.

Keywords: Naproxen, modulation, oxidation, potassium permanganate
DOI 10.14456/cast.2020.17

1. Introduction

Naproxen has an antipyretic and analgesic activity that targets the nucleoprotein to inhabit RNA - Naproxen association required for Naproxen function resulting in a novel antiviral against influenza type A virus [1-3]. Naproxen is a propionic acid derivative with the following chemical structure (Figure1).

Mahmood and Al-Sarraj [5] reported that Naproxen could be determined after a modification to the hydroxyl analog before coupled to the diazotized p-aminobenzoic acid in alkaline medium for forming an orange azo dye to be measured at 500 nm. Beer's law was then followed over the range from 0.5 to 32.5 µg·ml^{-1}. This method was found to be sensitive, accurate, precise and it was applied for the assay of Naproxen in tablets.

*Corresponding author: Tel.: 07 701 604 350
E-mail: hnsheker@yahoo.com

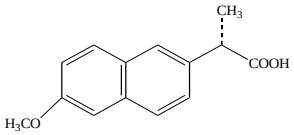


Figure 1. Chemical structure of Naproxen [4]

Naproxen sodium in combined dosage form with Sumatriptan was determined using two UV spectrophotometric methods at 272 and 284 nm as an absorptive point in the first method and at 298 nm and 335 nm as a zero-crossing point in the second method. These two methods have been applied in quality control of dosage forms [6].

A simultaneous determination of Naproxen and Esomeprazole mixture in a laboratory was reported. The method involves area under the curve in the ranges of 227-237 nm and 296.5-306.5 nm, the formation of the simultaneous equation at 232 nm and 301.5 nm, absorption correction at 232 nm λmax of Naproxen, 239.2 nm iso-absorptive point of Naproxen and Esomeprazole, 301.5 nm for absorption ratio method. The linearity for Naproxen and Esomeprazoleis 1-5 $\mu g \cdot ml^{-1}$ and 4-12 $\mu g \cdot ml^{-1}$, respectively. The method is accurate and precise for the simultaneous estimation of Naproxen and Esomeprazole in bulk formulations [7].

A simultaneous estimation of Naproxen and Pantoprazole in combined capsule dosage form has been developed. One method employs solving of simultaneous equations using 262 nm and 289 nm. The other method is Q- value analysis based on measurement of absorptivity at 262 nm and at iso-absorptive point 310 nm which shows linearity in the concentration range of 10.0-50.0 $\mu g \cdot ml^{-1}$ for Naproxen and 8.0- 18.0 $\mu g \cdot ml^{-1}$ for Pantoprazole [8].

A validated univariate and multivariate regression was developed for the simultaneous determination of a quaternary mixture of Imatinib, Gemifloxacin, Nalbuphineand Naproxen. The univariate method depends on measuring every drug in the quaternary mixture by using a ternary mixture of the other three drugs as a divisor. Peak amplitudes were measured at 294 nm, 250 nm, 283 nm and 239 nm within linear concentration ranges of 4.0-17.0, 3.0-15.0, 4.0-80.0 and 1.0-6.0 $\mu g \cdot ml^{-1}$ for Imatinib, Gemifloxacin, Nalbuphine and Naproxen, respectively. Multivariate methods adopted are partial least squares (PLS) in original and derivative mode [9].

Two spectrophotometric methods for determination of Naproxen was developed by the formation of ion-pair complex with bromocresol green at 424nm in one method and bromothymol blue at 422 nm in another method [10]. A bromophenol blue was used in the same procedure for determination of Naproxen in human serum samples with good accuracy and reproducibility [11]. There were some chromatographic methods for the estimation of Naproxen based on high-performance liquid chromatography by usage of C_{18} column were reported [12-14].

Potassium permanganate was used for determination of Famotidine in both pure form and in its dosage forms via oxidation of the drug in acid and alkaline media. Beer's law was obeyed over 0.75-7.5 and 2.5-20 ml in alkaline and acid media with a molar absorptivity value of $2.79x10^4$ and $1.62x10^4$ l $mol^{-1} \cdot cm^{-1}$, respectively [15]. Potassium permanganate was used to determine the usage of Raloxifene hydrochloride in dosage form by the formation of yellowish-brown product measured at 430 nm in acetic acid, while the other method is based on measuring the excess permanganate at 550 nm using H_2SO_4 to acidify the media [16].

In the present study, Naproxen was determined in commercially available tablet after a modulation reaction to the hydroxyl analog followed by oxidation using potassium permanganate as an oxidizing agent in an acidic medium. The excess of permanganate was followed at 545 nm as a decrease with the increase in the modulate Naproxen and in basic medium, whereas the manganate produced was followed at 610 nm as an increase with the increase in the modulate Naproxen.

2. Materials and Methods

All spectral and absorbance measurements were performed on a double-beam Jasco V-630 spectrophotometer with 1.0 cm matched quartz cells. The measurements of pH were performed using HANNA 301 pH meter, whereas BEL balance was used for weight measurements, reflux was utilized by electrothermal heater and stirring was utilized by Wisd stirrer. Chemicals used were of analytical grade.

Modification reaction of Naproxen: 0.04mol of pure Naproxen (9.2 g) was mixed with 25 ml hydrobromic acid (48%) and 25 ml acetic acid, the mixture was refluxed for 1.5 h and then was cooled, diluted with 25 ml distilled water, filtrated, dried and finally recrystallized using ethanol to produce a pink solid crystal with melting point at 190-191°C [17].

Modified Naproxen(100 µg/ml): this solution was prepared by dissolving 0.0100 g of mNaproxen in minimum amount of ethanol (2 ml), then the volume was diluted to 100 ml with distilled water in a volumetric flask. The solution was kept into a dark bottle and used for at least one month.

Pharmaceutical preparation (Naproxentablet):10 tablets of Naproxen (SDI) was ground into fine particles and a weight equivalent to one tablet was dissolved in 3 ml ethyl acetate and 1 ml HCl (3M). When two layers were separated, the organic layer was transferred to another tube and extraction was repeated 3 times, 1 ml of saturated solution of NaCl was added to organic layer and a sufficient amount of sodium sulphate was added after separation. The layer was left on-air for drying [18]; then the pure dried extract of Naproxen was modified to the hydroxyl analog as mentioned in the above step which produce pink solid crystal with melting point of 190-191°C [17] (extraction is necessary because tablets was burning during the modification reaction).

3. Results and Discussion

3.1 Modification reaction of Naproxen

The step of modification reaction of Naproxen involves a conversion of Naproxen into a hydroxy analog (mNaproxen) using hydrobromic acid in the presence of acetic acid at certain amounts as shown in Figure 2.

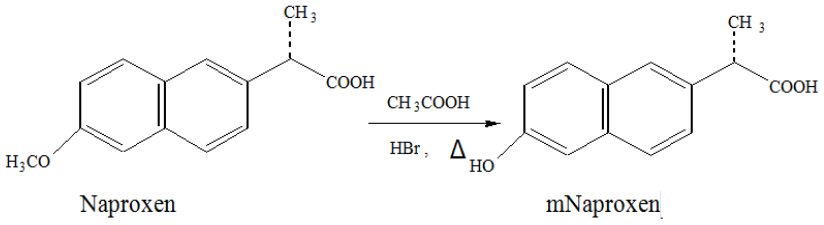


Figure 2. Modification reaction of Naproxen [17]

3.2 Procedure and calibration graph (basic medium)

Between 0.1-0.8ml of 100 µg.ml^{-1}standard of mNaproxen solution, reagents have been added in the following order: 1 ml of potassium permanganate (0.02 N), and 2 ml of H_2SO_4 (5M) has been

finally added, the volumes were completed to 10 ml, and the absorbance has been followed at 545 nm (blank against sample).

The calibration graph is linear for the concentrations from 10 to 80 µg of mNaproxen in 10 ml (1-8 ppm) with a molar absorptivity of 9.7 x 10^3 l·mol^{-1}·cm^{-1} with sensitivity index (Sandell's) equal to 0.0237 µg.cm^{-2} (Figure 3).

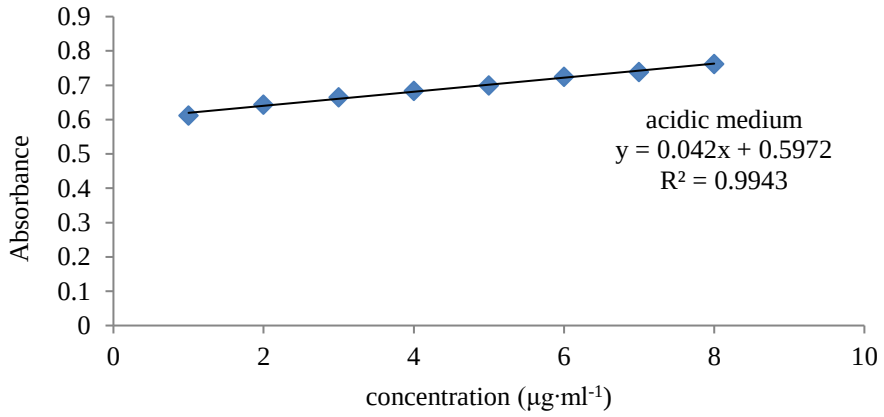


Figure 3. Calibration graph of Naproxenin acidic medium (blank against sample)

3.3 Procedure and calibration graph (basic medium)

To increase volume (0.2-0.7 ml) of 100 µg·ml^{-1} standard of mNaproxen solution, reagents have been added as follows: 3.5 ml of 0.02N potassium permanganate, and 2.5 ml of 0.1M NaOH have been added, and the volumes were completed to 10 ml in volumetric flasks. The absorbance has been measured at 610 nm against blank. The linearity range of the calibration graph is between 20 to 70 µg of mNaproxen in 10 ml (2-7 ppm) with a molar absorptivity of 6.8 x 10^3 l.mol^{-1}.cm^{-1} and the observed Sandell's index is equal to 0.0338 µg.cm^{-2} (Figure 4).

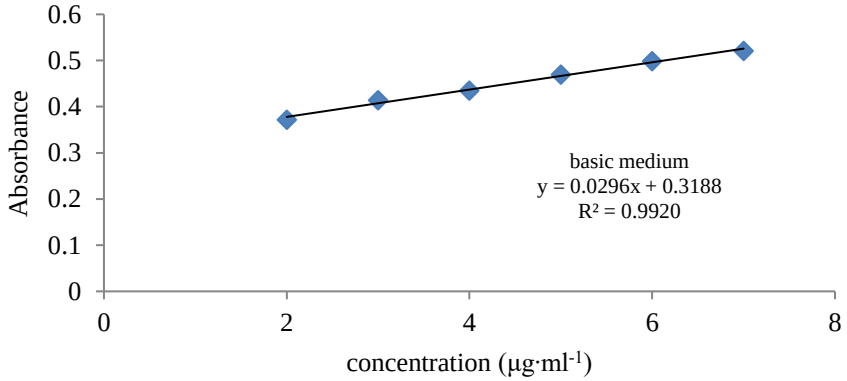


Figure 4. Calibration graph of Naproxenin basic medium (sample against blank)

3.4 Study of conditions (acidic medium)

The chemical reaction can be expressed as follows.
General reaction of MnO_4^- in acidic medium: $MnO_4^- + 8H^+ + 5e^- = Mn^{2+} + 4H_2O$
The oxidation reaction of mNaproxen (Figure 5):

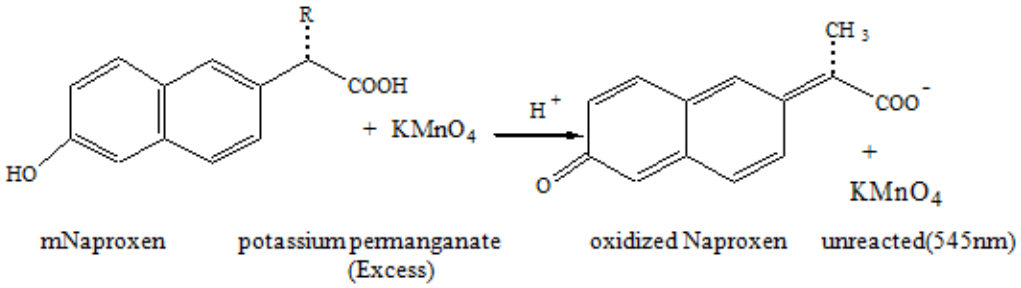


Figure 5. The oxidation reaction of mNaproxen in acidic medium

3.4.1 Effect of $KMnO_4$

The effect of (0.5-1.5) ml of potassium permanganate (0.02 N) has been followed against 20-80 µg of mNaproxen/10 ml and 1 ml H_2SO_4 (5M). The determination coefficient has been evaluated. Table 1 shows that 1 ml of $KMnO_4$ solution gives the best sensitivity.

Table 1. Effect of $KMnO_4$ amount on the sensitivity

Volume (ml) of 0.02N $KMnO_4$	Absorbance/µg of mNaproxen				R^2
	20	40	60	80	
0.5	0.207	0.212	0.220	0.229	0.9849
1.0	0.445	0.478	0.503	0.537	0.9968
1.5	0.569	0.618	0.646	0.689	0.9903

3.4.2 Selection of acid and its amount

Potassium permanganate act as a stronger oxidizing agent in acidic medium, therefore effect of different amount of three acids on the absorption intensity of the coloryield has been studied (Table 2).The results showed that the intensity of the maximum absorptions of the colored product was resulted from adding 2 ml of H_2SO_4, therefore, it is kept for further use.

Table 2. Selection of acid and its amount for acidifying the medium

Acid used (5M)	Absorbance/ml of acid				
	0.5	**1.0**	**1.5**	**2.0**	**2.5**
H_2SO_4	0.462	0.478	0.602	0.673	0.664
H_3PO_4	0.033	0.037	0.040	0.039	0.033
CH_3COOH	0.030	0.031	0.037	0.038	0.037

3.4.3 Study of order of addition

A study of the influence of the orders D+OX+A and D+A+OX (Drug- D, Oxidant- OX, Acid- A) shows no significant difference between them. The order D+OX+A was used in the previous and subsequent steps.

3.4.4 Effect of surfactant

In some reactions the presence of surfactants increases the absorption intensities of the color complex, this may be due to the correlation between chemical structures and surface properties [19], while there is no detailed discussion of surfactant behavior because of the wide range of possible environments for surfactant molecules [20].

In the present work, 1ml of (1×10^{-3}M) of cationic [(cetylpyridinium chloride, CPC), (cetyltrimethylammonium bromide, CTAB)] and 1 ml of anionic [sodium dodecyl sulphate, (SDS)] surfactants were added to the reaction mixture in different order. Table 3 indicates that there are no enhancements in the absorption intensity.

Table 3. Effect of surfactants on the absorption intensity

Surfactant solution (1×10^{-3}M)	**Absorbance/order of addition***		
	I	**II**	**III**
SDS	0.620	0.619	0.606
CTAB	Turbid	Turbid	Turbid
CPC	Turbid	Turbid	Turbid

Note: Absorbance without surfactant = 0.681
* I. Drug (D) + surfactant (S) + $KMnO_4$ (OX) + H_2SO_4 (A),
II. D+OX+S+A,III. D+OX+A+S

3.4.5 Effect of time of oxidation

An oxidation time of the reaction solution has been allowed for 12 min before dilution. It was found that 3 min was sufficiently enough for oxidation, therefore it is followed in subsequent steps.

3.4.6 Stability of reaction

A stability time (60 min) of the colored product after preparation of the sample completely (after dilution) has been followed. The reaction solution was still stable with an absorbance of 0.67.

3.4.7 Absorption spectrum

The absorption spectrum of the colored product against blank is shown in Figure 6. The maximum absorption intensity is exhibitedat 545 nm. This wavelength has been used in subsequent investigations.

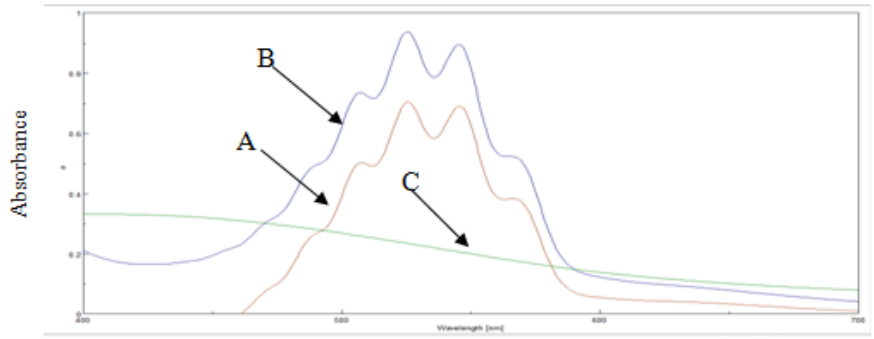


Figure 6. Measurement of absorption spectrum of 100 µg/10 ml of m Naproxen
A = blank against sample, B = blank against distilled water and C = sample against distilled water

3.4.8 Limit of detection and limit of quantification (LOD and LOQ)

In order to calculate the limit of detection (LOD) and limit of quantification (LOQ), 10 solutions of thelowest concentration (within the calibration) of Naproxen have been prepared according to the optimum reaction conditions and measured at 545 nm. The results in Table 4 show that the lowest content of mNaproxen that can be distinguished from background noise and measured with reasonable statistical certainty (LOD) is 0.075 $\mu g \cdot ml^{-1}$ and the lowest concentration on the calibration curve that can be measured with an acceptable level of accuracy and precision (LOQ) is 0.25 $\mu g \cdot ml^{-1}$.

3.4.9 Accuracy and precision of the calibration graph

Three different concentrations of mNaproxen is prepared for determination. The results are listed in Table 5, which indicates good accuracy and precision.

3.4.10 Application of the method for the determination of Naproxen in tablets

To test the applicability of the present method, it has been applied for determiningNaproxen after extraction from pharmaceutical preparations. The results in Table 6 exhibit good applicability of the method.

Table 4. Calculation of LOD and LOQ of the method

The absorbance of C_{low} (Xi)	$X_i - \bar{X}$	$(X_i - \bar{X})^2$
0.598	$-1.3x10^{-3}$	$16.9x10^{-7}$
0.600	$7x10^{-4}$	$4.9x10^{-7}$
0.599	$-3x10^{-4}$	$0.9x10^{-7}$
0.601	$1.7x10^{-3}$	$28.9x10^{-7}$
0.599	$-3x10^{-3}$	$0.9x10^{-7}$
0.599	$-3x10^{-4}$	$0.9x10^{-7}$
0.599	$-3x10^{-4}$	$0.9x10^{-7}$
0.598	$-1.3x10^{-3}$	$16.9x10^{-7}$
0.599	$-3x10^{-4}$	$0.9x10^{-7}$
0.601	$1.7x10^{-3}$	$28.9x10^{-7}$
$\bar{X} = 0.5993$		$\sum(X_i - \bar{X})^2 = 1.01x10^{-5*}$

Note: $^{*}\sigma = \sqrt{\frac{10.1\times10^{-6}}{10-1}} = 1.05 \times 10^{-3}$

Limit of Detection LOD = 3 σC(lowconcentration) / Slope =$(3\times 1.05 \times 10^{-3} \times 1)/0.042 = 0.075$ µg.ml^{-1}

Limit of QuantificationLOQ = 10 σC(low concentration)/slope = $(10\times 1.05 \times 10^{-3} \times 1)/0.042 = 0.25$ µg.ml^{-1}

Table 5. Accuracy and precisionof calibration graph

Amount of mNaproxen (µg/10 ml)	Amount mNaproxen found (µg/10 ml)	Relative error %*	Relative standard deviation %*
20	19.58	-2.1	± 3.2
40	40.52	+1.3	± 2.7
60	60.96	+1.6	± 1.9

*Average of five determinations

Table 6. Application of the method for the determination of Naproxen in Tablets

Amount of mNaproxen (µg /10 ml)	Recovery(%) of Naproxen*		
	Naproxen (tablet 500 mg) - Inaprolfort,Bilim, Turkey	Naproxen (tablet 250 mg)-S.D.I,Iraq	Naprox(tablet 500mg) - Medical Bahri Company, Damascus, Syria
20	98.6	97.5	98.2
40	99.3	97.9	99.4
60	98.8	97.4	98.6

*Average of three determinations

3.4.11 Comparison of the method with standard method and t-testcalculation

Both the present method and British Pharmacopeia method [21] have been applied at the same time for t-test calculation [22] and the value compared with statistical tables for four degrees of freedom at a 95% validation level. The result in Table 7 indicates no real difference between the two methods.

Table 7. Comparison of the method and T-test calculation

Drug	Recovery * (%)		t-exp
	Present method	British Pharmacopeia method**	
Naproxen (tablet 500 mg)-INaproxenrolfort, Bilim, Turkey	98.9	99.9	-2.14
Naprox (tablet 500mg) -Medical Bahri Company, Damascus, Syria	98.7	97.4	1.06

* Average of three determinations

**500 mg of Naproxen was dissolved in 70ml of absolute methanol and diluted after 30minto 100ml in a calibrated flask, further dilutions with absolute methanol were used to prepare 1000 ppm of Naproxen and measured at 331nm.

3.5 Study of conditions (basic medium)

The chemical reaction can be expressed as follows.
General reaction of MnO_4^- in basic medium: $4MnO_4^- + 4OH^- + e^- \rightarrow 4MnO_4^{2-} + 2H_2O + O_2$
The oxidation reaction of mNaproxen (Figure 7):

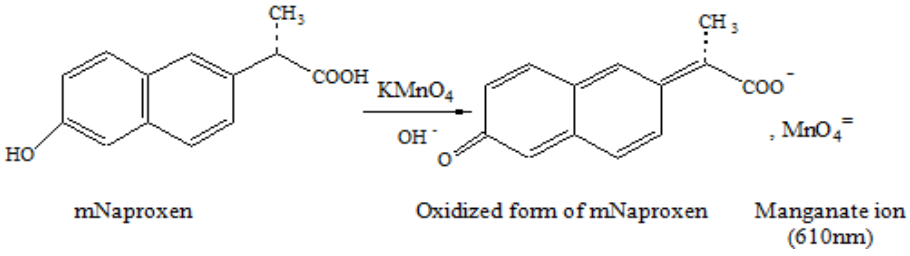


Figure 7. The oxidation reaction of mNaproxen in basic medium

3.5.1 Study ofthe effect of $KMnO_4$

The effect of 0.02N potassium permanganatehas been studied against 10-70 µg of mNaproxen/10 ml and the determination coefficient has been evaluated. Table 8 shows that 3.5 ml of $KMnO_4$ solution gives the best sensitivity.

Table 8. Effect of $KMnO_4$ amounton thesensitivity

Volume of 0.02N $KMnO_4$ (ml)	Absorbance/µg of mNaproxen				R^2
	10	30	50	70	
2.5	0.136	0.220	0.271	0.293	0.9341
3.0	0.139	0.232	0.308	0.368	0.9907
3.5	0.143	0.244	0.334	0.399	0.9910
4.0	0.091	0.176	0.252	0.306	0.9905

3.5.2 Selection of base and its amount

Different amounts of different bases on the absorption intensity of the resulted complex has been studied. The results in Table 9 show that 2.5 ml of NaOH gives the intensity of maximum absorptions of the colored product, therefore, it is used for the following steps.

Table 9. Selection of base and its amount to alkaline reaction medium

Base used (0.1M)	Absorbance/ ml of base					
	0.5	1.0	1.5	2.0	2.5	3.0
NaOH	0.168	0.269	0.334	0.379	0.413	0.370
KOH	0.088	0.131	0.161	0.162	0.174	0.176
Na_2CO_3	0.038	0.053	0.058	0.058	0.054	0.054
$NaHCO_3$	0.020	0.020	0.020	0.014	0.007	0.001

3.5.3 Order of addition

Different order of addition has been checked for predicting the best absorbance value.The order D+ OX+B gives the best results.

3.5.4 Effect of surfactant

One ml of cationic [cetylpyridinium chloride (CPC),cetyltrimethylammonium bromide (CTAB)] and anionic [sodium dodecyl sulphate (SDS)] surfactants with different order of additions has been followed (as shown in Table 10) in order to examine the effect on absorbance values.The results in Table 10 show that the presence of surfactant decreases the absorbance intensity of manganate or cause in turbidity; thiscan be explained by the decrease in oxidizing power of permanganate or formation of insoluble species respectively.

Table 10. Effect of surfactants on the absorbance intensity

Surfactant solution (1×10^{-3}M)	Absorbance/order* of addition		
	I	II	III
SDS	0.402	0.399	0.400
CTAB	Turbid	Turbid	Turbid
CPC	Turbid	Turbid	Turbid

*Absorbance without surfactant = 0.412,
I. =mNaproxen (D) + surfactant (S) + MnO_4 (OX) + NaOH(B), II. = D+OX+S+B, III. = D+ OX+B+S

3.5.5 Effect of time of oxidation

The reaction mixture needs 10 min as an oxidation time, which is sufficient enough for oxidation. These results have been concluded by allowing the reaction mixture before dilution to 15 min.

3.5.6 Absorption spectrum

The absorption spectrum of the colored product against blank is shown inFigure 8.The maximum absorption intensity is at 610 nm, andthis wavelength has been used in subsequent investigations.

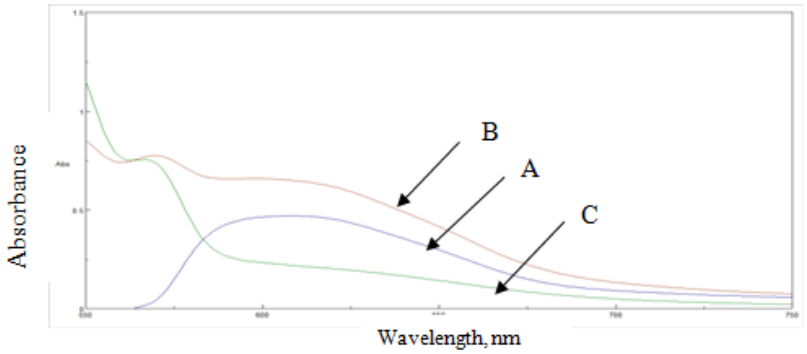


Figure 8. Measurement of absorption spectrum of 40 µg/10 ml of mNaproxen
A = Sample against blank, B = sample against distilled water and C = blank against distilled water

3.5.7 Limit of detection and limit of quantification (LOD and LOQ)

The blank solution exhibits a certain value of absorbance as shown in the absorptionspectrum, therefore to calculate the limit of detection (LOD) and limit of quantification (LOQ), 10 solutions of blank have been prepared according to the optimum reaction conditions and measured at 610 nm. The results in Table 11 show that the lowest content of analyte that can be distinguished from background noise and measured with reasonable statistical certainty (LOD) is 0.00988 $\mu g \cdot ml^{-1}$ and the lowest concentration on the calibration curve that can be measured with an acceptable level of accuracy and precision (LOQ) is 0.0329 $\mu g \cdot ml^{-1}$.

Table 11. Calculation of limit of detection and limit of quantification of the method

Absorbance of Blank (Xi)	$X_i - \bar{X}$	$(X_i - \bar{X})^2$
0.215	$5x10^{-4}$	$2.5x10^{-7}$
0.214	$-5x10^{-4}$	$2.5x10^{-7}$
0.214	$-5x10^{-4}$	$2.5x10^{-7}$
0.215	$5x10^{-4}$	$2.5x10^{-7}$
0.215	$5x10^{-4}$	$2.5x10^{-7}$
0.215	$5x10^{-4}$	$2.5x10^{-7}$
0.215	$5x10^{-4}$	$2.5x10^{-7}$
0.215	$5x10^{-4}$	$2.5x10^{-7}$
0.214	$-5x10^{-4}$	$2.5x10^{-7}$
0.213	$1.5x10^{-3}$-	$22.5x10^{-7}$
$\overline{X}$ =0.2145		$\sum(X_i - \bar{X})^2$=$4.5x10^{-6}$*

*$\sigma = \sqrt{\frac{4.5\times10^{-6}}{10\text{-}1}} = 7.07 \times 10^{-4}$

Limit of Detection (LOD) = 3 σ /0.2145 = 0.00988 µg·ml^{-1}

Limit of Quantification (LOQ) = 10 σ/ slope = 0.03296 µg·ml^{-1}

3.5.8 Accuracy and precision of the calibration graph of mNaproxen

To evaluate the accuracy and precision of the calibration graph, determining mNaproxen at three different concentrations was measured; the results are shown in Table 12, which exhibits good accuracy and precision.

3.5.9 Application of the method for the determination of Naproxen in tablets

The application of the method is shown in Table 13. The results show a good applicability range of the method.

3.5.10 Validation of the method

A comparison between the present method and British Pharmacopeia method [21] have been followed for determination of 500mg Naproxen tablet of two pharmaceutical company by application of the two methods at the same time followed by t-test calculation [22] and the value has been compared with statistical tables for four degrees of freedom at a 95% validation level. The results are shown in Table 14.

The result in Table 14 indicates no real difference between the suggested method and the method adopted in British pharmacopeia.

Table 12. Accuracy and precision of the calibration graph of mNaproxen

Amount mNaproxen taken (µg/10 ml)	Amount mNaproxen found (µg/10 ml)	Relative error %*	Relative standard deviation %*
20	20.30	+1.51	± 2.5
40	40.54	+1.37	± 1.45
60	60.59	+0.99	± 1.11

*Average of five determinations

Table 13. Application of the method for the determination of Naproxen in Tablets

AmountofmNaproxen (µg /10 ml)	Recovery(%) of Naproxen*		
	Naproxen (tablet 500 mg) - Inaprolfort,Bilim, Turkey	Naproxen (tablet 250 mg)- S.D.I,Iraq	Naprox (tablet 500mg) - Medical Bahri Company, Damascus, Syria
20	98.7	100.4	98.3
40	98.2	99.3	100
60	98.1	99.5	98.6

*Average of three determinations

Table 14. Comparison of the method and T-test calculation

Drug	Recovery * (%)		t-exp
	Present method	British Pharmacopeia method	
Naproxen (tablet 500 mg)- INaprolfort,Bilim, Turkey	98.3	99.9	-3.48
Naprox (tablet 500mg) - Medical Bahri Company, Damascus, Syria	98.9	97.4	1.25

* Average of three determinations

4. Conclusions

A modulate Naproxen has been oxidized by potassium permanganate in acidic medium (method 1), the excess of permanganate was followed at 545 nm as a decrease with the increase in the modulate Naproxen. Beer's law was obeyedbetween concentration valuesof 1-8 ppm. In a basic medium (method 2), the manganate produced was followed at 610 nm as an increase with the increase in the modulate Naproxen. The linearity range was between 2-7 ppm. The two methods areprecise, accurate, sensitive and applicable for pharmaceuticals. It does not need any organic reagent, does not need an extensive technique but it requires an extraction step.

5. Acknowledgements

The authors are grateful to Professor Dr. Nabeel S. Othman for the advice and Dr. Abdalrahman Basi lfor providing facilities in laboratory to carry out the present work during the reconstruction of Mosul and the University.

References

[1] Venkataharsha, P., Maheshwara, E., Raju, Y.P., Reddy, V.A., Rayadu, B.S. and Karisetty, B., 2015. Liposomal *Aloe vera* trans-emulgel drug deliveryof Naproxen and nimesulide: A study. *International Journal of Pharmaceutical Investigation*, 15(1), 28-34.

[2] Brashier, D.B.S., Khadka, A., Mishra, P.,Sharma, A.K., Dahiya, N. and Gupta, A.K., 2014. Evaluation of the protectiveeffects of Naproxen and celecoxib on Naproxenhthalene-induced cataract in albino rats. *Nigerian Journal of Experimental and Clinical Bioscience*, 2 (1), 28-32.

[3] Nair, A.S., 2019. Cardiovascular safety of Naproxen for treating cancer and noncancer chronic pain. *Indian Journal of Palliative Care*, 25 (1), 165-166.

[4] Ammar, Y.A., Salem, M.A., Fayed, E.A., Helal, M.H., El-Gaby, M.S.A. and Thabet, H.Kh., 2019. Naproxenderivative: Synthesis, reactions and biological applications. *Synthetic Communications*, 47 (15), 1341-1367.

[5] Mahmood, H.Sh. and Al-Sarraj, T.Z., 2019. A creative indirect spectophotometric determination of Naproxen. *Rafidain Journal of Science*, 28 (1), 51-60.

[6] Trinath, M., Banerjee, S.K., Teja, H.H. and Bonde, C.G., 2010. Development and validation of spectrophotometric method for simultaneous estimation of Sumatriptan and Naproxen sodium in tablet dosage form. *Der Pharmacia Sinica*, 1, 36- 41.

[7] Jain, N.A., Lohiya, R.T. and Umekar, M.J., 2011. Spectrophotometric determination of Naproxen and esomeprazole in a laboratory mixture by simultaneous equation, absorption correction, absorption ratio and area under curve methods. *International Journal of Pharmaceutical Science Research*, 2(5), 130-134.

[8] Sloka, S.N., Gurupadayya, B.M. and Kumar, Ch.A., 2011. Simultaneous spectrophotometric determination of Naproxen and pantoprazole in pharmaceuticals dosage form. *Journal of Applied Chemical Research*, 17, 65-74.

[9] Belal, F., Ibrahim, F., Sheribah, Z.A. and Alaa, H., 2018. New spectrophotometric/chemometric assisted methods for the simultaneous determination of imatinib, gemifloxacin, nalbuphine and naproxen in pharmaceutical formulations and human urine. *Spectrochimica Acta. Part A: Molecular and Biomolecular Spectroscopy*, 198, 51-60.

[10] Alizadeh, N. and Keyhanian, F., 2015 Simple, sensitive and selective spectrophotometric assay of naproxen in pure, pharmaceutical preparation and human serum samples. *Acta Poloniae Pharmaceutica-Drug Research*, 72(5), 867-875.
[11] Keyhanian, F. Alizadeh, N. and Shojaie, A.F., 2014. Spectrophotometric determination of Naproxen as ion-pair with bromophenol blue in bulk, pharmaceutical human serum samples. *Current Chemistry Letters*, 3, 15-22.
[12] Muneer, S., Muhammad, I.N., Abrar, M.A., Munir, I., Kaukab, I., Sagheer, A., Zafar, H. and Sultana, K., 2017. High performance liquid chromatographic determination of Naproxen in prepared pharmaceutical dosage form and human plasma and its application to pharmacokinetic study. *Journal of Chromatography & Separation Techniques*, 8(3), 1-5.
[13] Pansara, V.H., Kakadiya, M. and Saishivam, 2013. Development and validation of RP-HPLC method for simultaneous estimation of Naproxen and paracetamol in their combined solid dosage form. *An International Journal*, 1(7), 633-636.
[14] Veeragoni, A.K., Sindgi, V.M. and Satla, S.R., 2016. Bioanalytical validated LC-MS method for determination of Naproxen in human plasma. *International Journal for Modern Trends in Science and Technology*, 2(9), 96-99.
[15] Basavaiah, K. and Devi, O.Z., 2010. Application of oxidizing properties of permanganate to the determination of famotidine in pharmaceutical formulation. *Journal of Mexican Chemical Society*, 54(4), 182-191.
[16] Basavaiah, K., Tharpa, K., Anil Kumar, U.R., Rajedraprasad, N., Hiriyanna, Sg. and Vinay, K.B., 2009. Optimized and validated spectrophotometric methods for the determination of raloxifene in pharmaceuticals using permanganate. *Archive of Pharmacal Research*, 32(9), 1271-1279.
[17] Al-Hamdany, F.K., 2009. *Synthesis of Some (s)-2-(6'-methoxy-2'-Naproxenhthyl) Propanoic Acid Derivatives (Naproxen)*. Ph.D. University of Mosul.
[18] Ka Lei Senior Project, 2009. *Chem 319: Isolation of Naproxen*. [online] Available at: https://www.cpp.edu/~psbeauchamp/pdf/424_naproxen_procedure.pdf
[19] Mahmood, H.Sh. and Al-Sarraj, T.Z., 2019. A novel spectrophotometric determination of Naproxen via a modification to the hydroxy analog followed by oxidative-coupling reaction. *Colloidal Education Journal*, 1 (1), 261-272.
[20] Myers, D., 2006. *Surfactant Science and Technology*. 3rded. New Jersey: John Wiley and Sons, Inc., p.126.
[21] British Pharmacopeia, 2013. 7th ed. London: System Simulation Ltd the Stationary Office (CD ROM).
[22] Christian, G.D., 2004. *Analytical Chemistry*. 6th ed. New York: John Wiley and Sons, Inc.

Time Series Analysis and Forecast of Influenza Cases for Different Age Groups in Phitsanulok Province, Northern Thailand

Sasithan Maairkien[1], Darin Areechokchai[2], Sayambhu Saita[3] and Tassanee Silawan[1]*

[1]Department of Community Health, Faculty of Public Health, Mahidol University, Bangkok, Thailand
[2]Bureau of Vector Borne Disease, Department of Disease Control, Ministry of Public Health, Nonthaburi, Thailand
[3]Faculty of Public Health, Thammasat University, Lampang, Thailand

Received: 21 December 2019, Revised: 13 March 2020, Accepted: 3 April 2020

Abstract

Influenza is still a major problem in Thailand where the incidence varies among age groups. This descriptive research using retrospective data collection aimed to describe the distribution and synchrony, and to forecast Influenza cases in different age groups in Phitsanulok province, northern Thailand. Influenza cases from January 2009 to December 2016 were obtained from R506, Bureau of Epidemiology. Temporal distribution was visually interpreted from line and decomposed graphs. The synchrony between all pairs of age groups was analyzed using Pearson correlation. The 2017 Influenza cases were forecasted using the seasonal ARIMA model in RStudio version 1.1.419. The results showed that trend of Influenza cases for the three age groups: less than 25 years, 25-64 years, and 65 years and older, slightly decreased from 2011 to 2015 and dramatically increased in 2016. The two peaks were observed, i.e. major peak in September and minor peak in February. The cyclic pattern likely observed major peak in two consecutive years for every five years. All pairs of data series co-varied over time. The best models to forecast Influenza cases were seasonal ARIMA $(1,0,1)(0,1,1)^{12}$, seasonal ARIMA $(1,0,0)(0,1,1)^{12}$ and seasonal ARIMA $(1,0,1)(1,1,1)^{12}$ for age less than 25 years, age 25-64 years and age 65 years and older with the MAPE 15.54, 17. 27 and 15.61 respectively. There were 1,698 forecasted cases in age less than 25 years, followed by 1,478 cases in age 65 years and older and 471 cases in age 25-64 years. The major peak in February and minor peak in September were observed in all age groups. In 2017, the forecasted cases were lower than the reported cases in all data series, except for age 65 years and older.

Keywords: Influenza, time series, forecast, seasonal ARIMA, surveillance
DOI 10.14456/cast.2020.18

*Corresponding author: Tel.: +66 23 54 85 43 Fax: +66 23 54 85 43 ext. 4777
E-mail: tsilawan@gmail.com

1. Introduction

Influenza is an acute viral infection caused by an Influenza virus, which is one of the prioritized problems in Thailand. There are three types of Influenza virus, i.e. A, B and C. The incubation period is about two days. The virus can spread via droplet, hand and contaminated objects and disperse in schools, nursing homes, businesses or towns [1, 2]. World Health Organization (WHO) reported that epidemics occur mainly during winter in the temperate zone meanwhile Influenza may occur throughout the year, causing outbreaks more irregularly, in tropical regions. An estimated annual rate of Influenza was 5-10% in adults and 20-30% in children. There were about three to five million cases of severe illness, and about 250,000 to 500,000 deaths [2]. In 2014, Thailand had Influenza morbidity rate of 114.4 per 100,000 population in which the northern region had the highest morbidity rate from 2010 to 2014 and Phitsanulok province reported the highest Influenza cases among provinces in the region [3, 4].

Although Influenza occurrence varied in different age groups, most of previous researches regarding the distribution and forecast of Influenza focused on the overall cases or morbidity rates [5-7], which was not specific to age group and the synchrony between different age groups to distinguish the contributed factors were not analyzed. Therefore, the current research aimed to describe the distribution of Influenza in different age groups based on surveillance data (R506) managed by the Bureau of Epidemiology (BOE), Ministry of Public Health, Thailand. The research also aimed to analyze the disease synchrony among age groups to indicate the contribution of endogenous or exogenous factors [8]. Forecasting was also performed for each age group. The finding would be useful for control activities in the optimal time, place and target group, as well as an example for the utilization of data routinely collected.

2. Materials and Methods

2.1 Design and data

The descriptive research using retrospective data routinely collected was carried out in Phitsanulok province, northern Thailand. There were 9 districts with approximately 866,891 population. The reported new Influenza cases by age group from January 2009 to December 2016 were obtained from the national epidemiological surveillance database (R506), Bureau of Epidemiology, Department of Disease Control, Ministry of Public Health, Thailand. R506 is a reporting form in the national surveillance system of Thailand where all governmental hospitals and some private hospitals throughout the country have to report new cases of diseases under surveillance which cover more than 70 items. The monthly Influenza cases from January 2009 to December 2016 of the three age groups; less than 25 years, 25-64 years, and 65 years and older; were processed and analyzed using RStudio version 1.1.419.

2.2 Temporal distribution and synchronicity

Line graphs and the Seasonal Trend Decomposition using Loess (STL) were plotted to identify trend, cyclic and seasonal patterns. The synchrony between all pairs of data series between different age groups was analyzed using Pearson correlation to explore how two continuous signals covary over time. A number (-1) indicates negatively correlated meanwhile (0) indicates not correlated and (1) indicates positively correlated.

2.3 Seasonal ARIMA model development and forecast

Monthly Influenza cases from January 2009 to December 2015 were the training set and the data from January to December 2016 were the validating set for identifying the seasonal ARIMA model. The data were checked for stationarity where non-stationary data were transformed by natural log and seasonal differences. The sample autocorrelation function (ACF) and sample partial autocorrelation function (PACF) were generated to identify the possible ARIMA (p, d, q) $(P, D, Q)^s$ models where: 'p', 'd', and 'q' represented autoregressive (AR) order, differencing order, and moving average (MA) of regular terms and 'P', 'D', and 'Q' referred to AR, differencing and MA of seasonal terms, and 's' represented the span of the periodic seasonal behavior (12 month per year) [6]. Model diagnostic testing to identify the fit models was analyzed using the standardized residuals, residual ACF, p values for Ljung-Box statistic, the cumulative periodogram of residuals, and the Box-Pierce statistic. The model with the lowest Akaike Information Criterion (AIC) and the lowest mean absolute percentage error (MAPE) was the best model. Forecast of the Influenza cases from January to December 2017 was performed using the best model for each series fitted with reported cases from January 2009 to December 2016. The research project was approved as a human ethics exemption by the Ethics Review Committee of Human Research, Faculty of Public Health, Mahidol University, Thailand.

3. Results and Discussion

3.1 Temporal distribution of Influenza

The trend of Influenza cases for the three age groups; less than 25 years, 25-64 years and 65 years and older; had increased from 2009 to 2011, thereafter the trend was slightly decreased and increased in 2016. The major peak occurred in September and minor peak most likely occurred in February. The cyclic pattern likely showed major peak in two consecutive years for every five years. After adjusted seasonal variation, the trend of all three age groups declined from 2011 to 2015 and increased again in 2016 with slightly higher than 2009 (Figures 1 and 2).

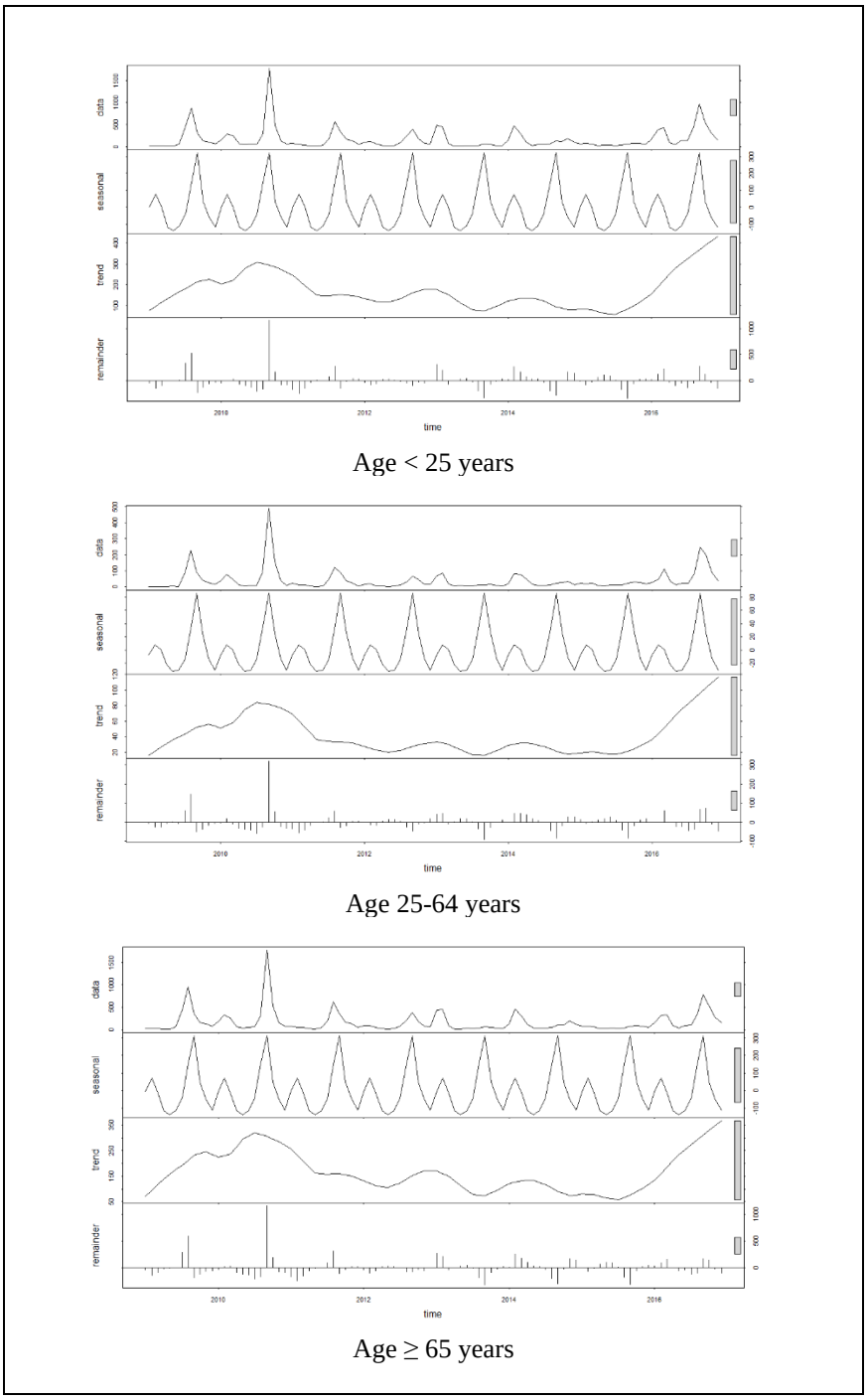


Figure 1. Influenza cases, seasonality and trend of age < 25 year (top), age 25-64 years (middle) and age ≥ 65 years (bottom) in Phitsanulok province, Northern Thailand

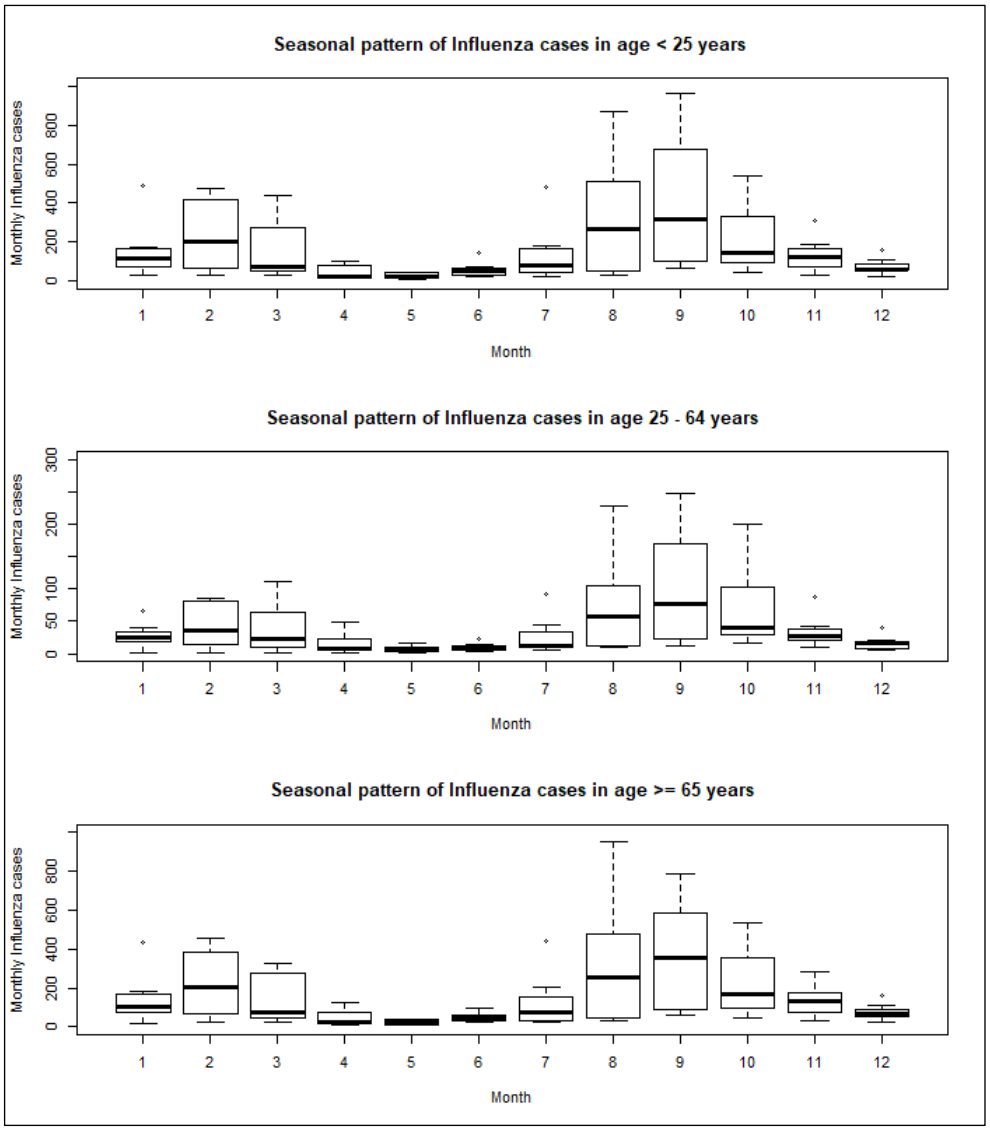


Figure 2. Box plot of Influenza cases (maximum, average, minimum) by month from 2009 to 2016 in age < 25 year (top), age 25-64 years (middle), and age ≥ 65 years (bottom) in Phitsanulok province, Northern Thailand

3.2 Synchrony of Influenza for different age groups

Measuring synchrony of monthly Influenza cases in Phitsanulok province for all pairs of data series by age group, the results showed strong correlation (r = 0.97 to 0.99) in all pairs of the data series. Considering season, the strong correlation was also observed in all pairs of data series by age group during summer, rainy and winter seasons. This indicated that Influenza cases in the three age groups co-varied or consistently fluctuated over time (Table 1).

Table 1. Pearson correlation of monthly Influenza cases for all pairs data series by age group and seasons in Phitsanulok province, northern Thailand

	Correlation Coefficient (r)											
	All year round January - December			**Summer season March – May**			**Rainy season June – October**			**Winter season November- February**		
Pairs of data series	1	2	3	1	2	3	1	2	3	1	2	3
1	1.000	0.971	0.992	1.000	0.899	0.899	1.000	0.959	0.993	1.000	0.965	0.966
2	0.971	1.000	0.985	0.899	1.000	0.985	0.959	1.000	0.934	0.965	1.000	0.969
3	0.992	0.985	1.000	0.899	0.985	1.000	0.993	0.934	1.000	0.966	0.969	1.000

Note: 1=Data series of Influenza cases in age < 25 years, 2=Data series of Influenza cases in age 25 - 64 years and 3=Data series of Influenza cases in age ≥ 65 years

3.3 Identifying the best seasonal ARIMA model for each data series

The best model to forecast Influenza cases for age <25 years, age 25-64 years and age ≥ 65 years was seasonal ARIMA (1,0,1)(0,1,1)12, seasonal ARIMA (1,0,0)(0,1,1)12 and seasonal ARIMA (1,0,1)(1,1,1)12, respectively. The MAPE, which calculated by comparing the predicted cases with reported cases during January to December 2016, were 15.54, 17.27 and 15.61 (Table 2).

Table 2. The best model, AIC and MAPE for each data series

Data series	**Seasonal ARIMA model $(p,d,q)(P,D,Q)^s$**	**AIC**	**MAPE**
Age < 25 years	$(1,0,1)(0,1,1)^{12}$	155.50	15.54
Age 25 - 64 years	$(1,0,0)(0,1,1)^{12}$	209.12	17.27
Age ≥ 65 years	$(1,0,1)(1,1,1)^{12}$	148.20	15.61

3.4 Forecast of Influenza cases from January to December 2017

The best model for each data series was fitted to monthly Influenza cases for January 2009 to December 2016 to forecast Influenza cases from January to December 2017. There were 1,698 forecasted cases in age < 25 years, followed by 1478 cases in age ≥ 65 years and 471 cases in age 25-64 years. The major peak in September and minor peak in February were observed in all age groups (Figure 3). Comparison between forecasted cases and reports cases in 2017, the forecasted cases were lower than the reported cases in all data series, except for age ≥ 65 years, which the forecasted cases were higher than the reported cases (Figure 4).

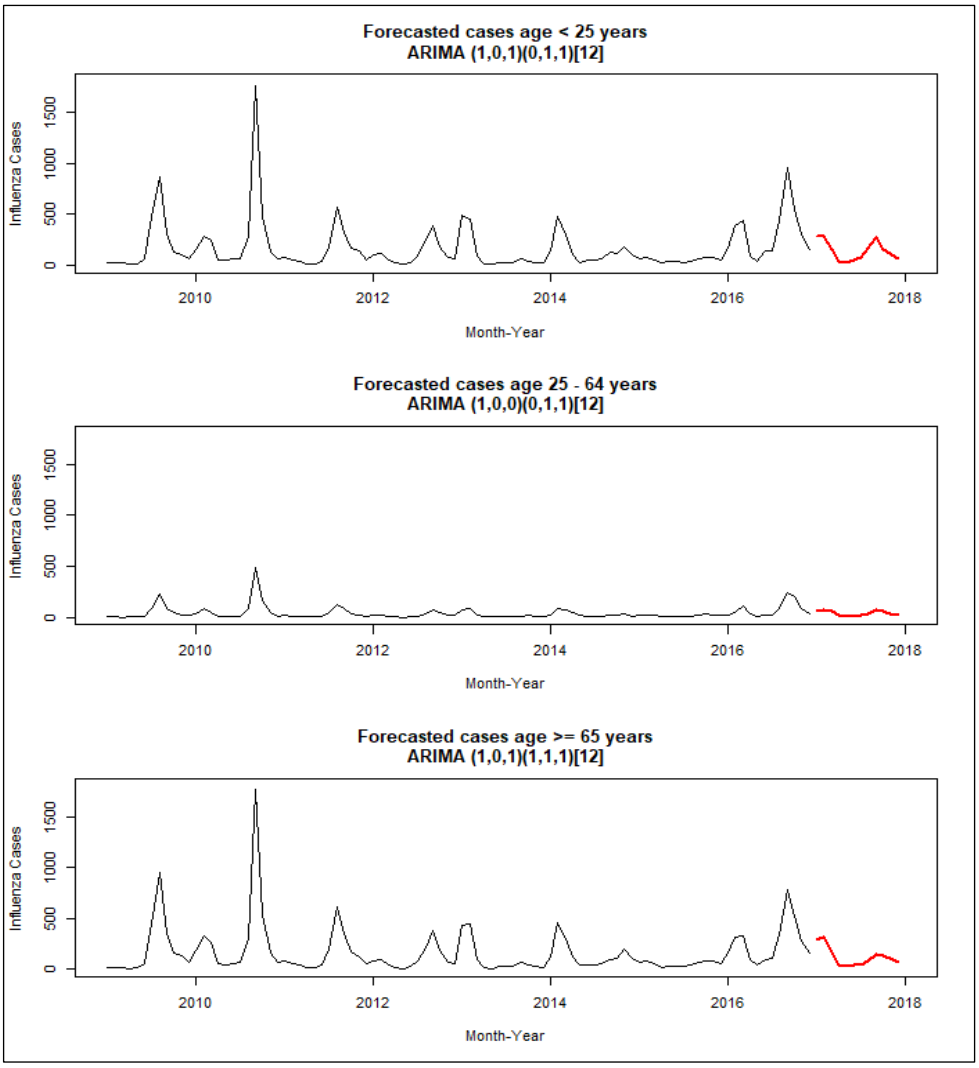


Figure 3. Monthly forecasted Influenza cases from January to December 2017 for each data series: age < 25 years (top), age 25-64 years (middle), and age ≥ 65 years (bottom) in Phitsanulok province, Northern Thailand

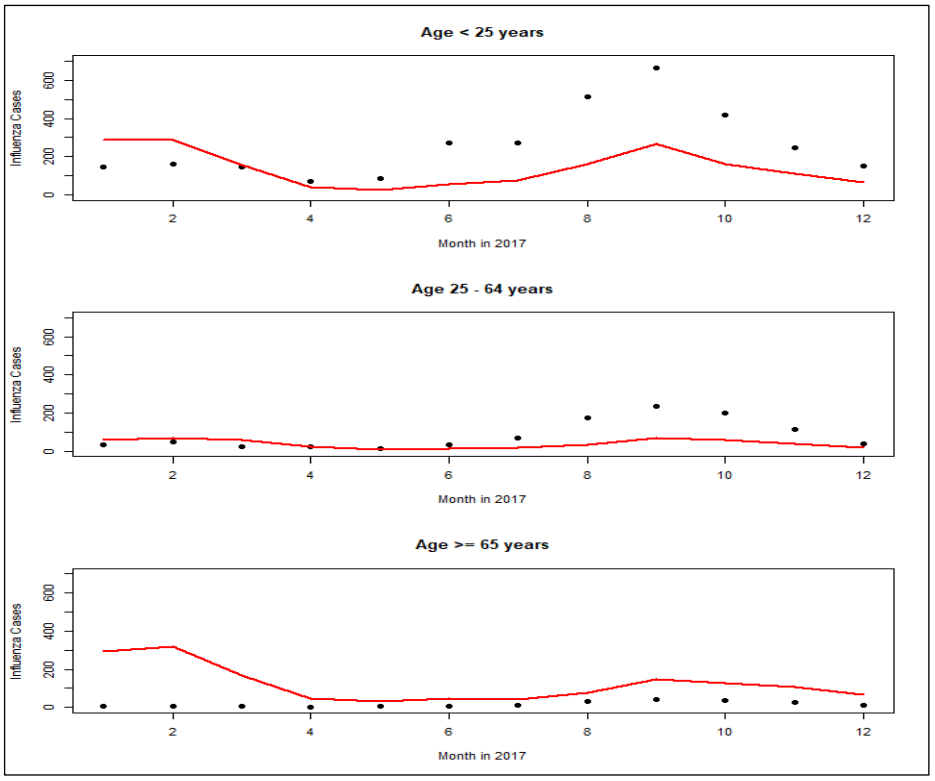


Figure 4. Monthly forecasted Influenza cases (line) for each data series compared with reported cases (dot) from January to December 2017: age < 25 years (top), age 25-64 years (middle) and age ≥ 65 years (bottom) in Phitsanulok province, Northern Thailand

The trend of Influenza cases for the three age groups; less than 25 years, 25-64 years and 65 years and older; had increased from 2009 to 2011, thereafter the trend was slightly decreased and increased in 2016. This was in accordance with the cyclic pattern which likely showed major peak in two consecutive years for every five years (2009-2010 and 2016-2017). However, due to only eight years of data series, it may not be clear to see the cyclic pattern of the disease. Various cyclic patterns of Influenza have been observed, such as every 1, 2-3, 5-6, 8, 10.6-11.3, 13 and 18-19 years. The cyclic patterns were associated with the cycle of natural environment related to Physico-chemical processes and irradiations of cosmic or terrestrial origin, called as space weather [9-12]. Some previous researches reported that Influenza epidemic changes had been 1.3 times more frequent in the years of the sudden increase in solar activity. Moreover, the cyclic variations of Influenza every 2-4 years can be associated with the frequency of such sudden changes [9, 10].

Seasonality, the major peak occurred in September and minor peak most likely occurred in February. The major peak in September may cause by rainfall, which increases humidity and precipitation. This corresponds with the highest cumulative rainfall in Phitsanulok province which was reported from August to September (247 mm) [13]. Those factors are related to increased Influenza activities [14]. In fact, year-round Influenza activity is more common in tropical countries [15, 16] including Thailand with the peak during rainy season between July and September [17, 18], and a seasonal pattern of Influenza in non-epidemic year caused by circulation of Influenza A [19]. The minor peak in February, during winter, may be due to the low

temperature which increases virus stability. This is in line with the minimum temperature in Phitsanulok province, which was 18.6°C in December to 25.4 °C in April [13]. The low temperatures and humidity enhance the temperate countries typically experience Influenza during winter season [20-23].

Temporal patterns of Influenza among the three age groups were not different. This corresponded with the results of synchronicity which found that Influenza occurrence of the three age groups co-varied or consistently fluctuated over time, both in year-round and each season. This reflects the disease occurrence are more likely affected by common factors rather than specific host factors in different age groups. The common factors can be an exogenous factor such as meteorological factors (temperature, rainfall, humidity) [20-23] and solar activity [10] or it can be some host factors found in all age groups, such as contact pattern [24, 25] and personal hygiene; hand washing, wearing a protective mask when illness [26]. Those factors concomitantly enhance the etiologic agent and spread of Influenza in all age groups. However, some previous studies reported the differences among age groups due to pre-existing specific immunity, accessibility to health services and physiological factors [27, 28].

The reported cases in age less than 25 and age 24-64 years were higher than the forecasted cases. This is in line with the nationwide epidemic of Influenza in 2017 which showed that there were 61 event-based surveillance of Influenza outbreaks in 2017, compared to 34 and 24 events in 2016 and 2018, respectively [29]. On the contrary, the reported cases in age ≥ 65 years were lower than the forecasted cases. It may be due to the reason that in 2017 Phitsanulok province has supported free vaccination for the four high risk groups which included elderly older than 65 years. It was found that the coverage of the vaccination in the elderly was quite high, which may be due to the campaign and accessibility to government services [30, 31]. Therefore, the reported cases were lower than the forecasted cases.

4. Conclusions

Trend of Influenza cases for the three age groups in Phitsanulok province slightly decreased from 2011 to 2015 and dramatically increased in 2016. The two peaks were observed; major peak in September and minor peak in February. The cyclic pattern likely observed major peak in two consecutive years for every five years. All pairs of data series co-varied over time. The forecasted cases of the three age groups were less than cases in 2016. The major peak in February and minor peak in September were observed in all data series. Age less than 25 years had the highest forecasted cases, followed by age 65 years and older and age 25-64 years. The forecasted cases in 2017 were lower than the reported cases in all data series, except for age ≥ 65 years. Prevention and control of Influenza should be strengthened before September (rainy season) and February (winter), focusing on the common factors enhancing the disease occurrence in all age, such as hygiene, hand washing, wearing a protective mask when illness. In order to get more accurate forecast, meteorological factors (temperature, rainfall, humidity), environment, space weather, and solar activity should be included in the forecasting model. The variables in surveillance system should be utilized to guide more specific for Influenza prevention and control.

5. Acknowledgements

We would like to thank the Bureau of Epidemiology, Department of Disease Control, Ministry of Public Health, Thailand for the database of Influenza cases.

References

[1] Punpanich, W. and Chotpitayasunondh, T., 2012. A review on the clinical spectrum and natural history of human Influenza. *International Journal of Infectious Diseases*, 16(10), e714-e723.

[2] World Health Organization, 2018. *Influenza (Seasonal)*. [online] Available at: https://www.who.int/news-room/fact-sheets/detail/Influenza-(seasonal).

[3] Bureau of Epidemiology, Ministry of Public Health, 2013. *Annual Epidemiological Surveillance Report 2013*.

[4] Bureau of Epidemiology, Ministry of Public Health, 2014. *Annual Epidemiological Surveillance Report 2014*.

[5] Paget, J., Spreeuwenberg, P., Charu, V., Taylor, R.J., Iuliano, A.D., Bresee, J., Simonsen, L., Viboud, C., Global Seasonal Influenza-associated Mortality Collaborator Network and GLaMOR Collaborating Teams, 2019. Global mortality associated with seasonal Influenza epidemics: New burden estimates and predictors from the GLaMOR Project. *Journal of Global Health*, 9(2), 020421. https://doi.org/10.7189/jogh.09.020421.

[6] Chadsuthi, S., Iamsirithaworn, S., Triampo, W. and Modchang, C., 2015. Modeling seasonal Influenza transmission and its association with climate factors in Thailand using time-series and ARIMAX analyses. *Computational and Mathematical Methods in Medicine*, 436495. doi:10.1155/2015/436495.

[7] Song, X., Xiao, J., Deng, J., Kang, Q., Zhang, Y. and Xu, J., 2016. Time series analysis of Influenza incidence in Chinese provinces from 2004 to 2011. *Medicine*, 95(26), e3929. doi:10.1097/MD.0000000000003929.

[8] Yang, L., Chan, KH., Suen, L.K., Chan, K.P., Wang, X., Cao, P., He, D., Peiris, J.S.M. and Wong, C.M., 2015. Age-specific epidemic waves of Influenza and respiratory syncytial virus in a subtropical city. *Science Report*, 5, 10390, https://doi: 10.1038/srep10390.

[9] Hoyle, F. and Wickramasinghe, N.C., 1990. Sunspots and Influenza. *Nature*, 343; 304, https://doi.org/10.1038/343304a0.

[10] Tapping, K.F., Mathias, R.G. and Surkan, D.L., 2001. Influenza pandemics and solar activity. *Canadian Journal of Infectious Diseases*, 12, 61-62.

[11] Webster, R.G., Bean, W.J., Gorman, O.T., Chambers, T.M. and Kawaoka, Y.,1992. Evolution and ecology of Influenza A viruses. *Microbiological Reviews*, 56(1), 152-179.

[12] Dimitrov B.D., and Babayev, E.S., 2015. Cyclic variations in the dynamics of flu incidence in Azerbaijan, 1976-2000. *Epidemiology & Infection*, 143(1),13-22.

[13] Moura, F.E., 2010. Influenza in the tropics. *Current Opinion in Infectious Diseases*. 23(5), 415-420.

[14] Xu, C., Thompson, M.G. and Cowling, B.J., 2017. Influenza vaccination in tropical and subtropical areas. *Lancet Respiratory Medicine*, 5(12), 920-922.

[15] Tamerius, J.D., Shaman, J., Alonso, W.J., Bloom-Feshbach, K., Uejio, C.K., Comrie, A. and Viboud, C., 2013. Environmental predictors of seasonal Influenza epidemics across temperate and tropical climates. *PLoS Pathogens*, 9(3), e1003194, https://doi.org/10.1371/journal.ppat. 1003194.

[16] Suthachana, S., Kaewnorkkao, W., Suangtho, P., Sayumpurujinun, S. and Kongyu, S., 2012 Forecasting the situation of Influenza in Thailand, 2012. *Bureau of Epidemiology, Department Disease Control, Minstry of Public Health, Workshop Conference in Analyzed and Forecasting*, Bangkok, Thailand, 12-14 September 2012.

[17] Prachayangprecha, S., Makkoch, J., Suwannakarn, K., Vichaiwattana, P., Korkong, S., Theamboonlers, A. and Poovorawan, Y., 2013. Epidemiology of seasonal Influenza in Bangkok between 2009 and 2012. *The Journal of Infection in Developing Countries*, 7(10), 734-740

[18] Prachayangprecha, S., Vichaiwattana, P., Korkong, S., Felber, J. A. and Poovorawan, Y., 2015. Influenza activity in Thailand and occurrence in different climates. *SpringerPlus*, 4, 356, https://doi: 10.1186/s40064-015-1149-6.

[19] Tamerius, J., Nelson, M.I., Zhou, S.Z., Viboud, C., Miller, M.A. and Alonso, W.J., 2011. Global Influenza seasonality: reconciling patterns across temperate and tropical regions. *Environmental Health Perspectives*, 119 (4), 439-445.

[20] Chowell, G., Viboud, C., Munayco, C.V., Gómez, J., Simonsen, L., Miller, M.A., Tamerius, J., Fiestas, V., Halsey, E.S. and Laguna-Torres, V.A., 2011. Spatial and temporal characteristics of the 2009 A/H1N1 Influenza pandemic in Peru. *PloS One*, 6(6), e21287, https://doi:10.1371/journal.pone.0021287.

[21] Lowen, A.C. and Steel, J., 2014. Roles of humidity and temperature in shaping Influenza seasonality. *Journal of Virology*, 88(14), 7692-7695.

[22] Liu, Z., Zhang, J., Zhang, Y., Lao, J., Liu, Y., Wang, H. and Jiang, B., 2019. Effects and interaction of meteorological factors on Influenza: Based on the surveillance data in Shaoyang, China. *Environmental Research*, 172, 326-332.

[23] Eccles, R., 2002. An explanation for the seasonality of acute upper respiratory tract viral infections. *Acta Oto-Laryngologica*, 122(2), 183-191.

[24] Mossong, J., Hens, N., Jit, M., Beutels, P., Auranen, K., Mikolajczyk, R., Massari, M., Salmaso, S., Tomba, G.S., Wallinga, J., Heijne, J., Sadkowska-Todys, M., Rosinska, M. and Edmunds, W.J., 2008. Social contacts and mixing patterns relevant to the spread of infectious diseases. *PLoS Medicine*, 25, e74, https://doi: 10.1371/journal.pmed.0050074.

[25] Schmidt-Ott, R., Schwehm, M. and Eichner, M., 2016. Influence of social contact patterns and demographic factors on Influenza simulation results. *BMC Infectious Diseases*, 16(1), 646. doi:10.1186/s12879-016-1981-5.

[26] Saunders-Hastings, P., Crispo, J.A.G., Sikora, L. and Krewski, D., 2017. Effectiveness of personal protective measures in reducing pandemic Influenza transmission: A systematic review and meta-analysis. *Epidemics*, 20, 1-20.

[27] Karageorgopoulos, D.E., Vouloumanou, E.K., Korbila, I.P., Kapaskelis, A. and Falagas, M.E., 2011. Age distribution of cases of 2009 (H1N1) pandemic Influenza in comparison with seasonal Influenza. *PloS One*, 6(7), e21690. https://doi:10.1371/journal.pone.0021690.

[28] Mertz, D., Kim, T.H., Johnstone, J., Lam, P.P., Science, M., Kuster, S.P., Fadel, S.A., Tran, D., Fernandez, E., Bhatnagar, N. and Loeb, M., 2013. Populations at risk for severe or complicated Influenza illness: systematic review and meta-analysis. *British Medical Journal*, 347, f5061. https://doi:10.1136/bmj.f5061.

[29] Suthachana, S. and Saritapirak, N., 2018. Event-based surveillance of Influenza outbreak in Thailand during 2016-2018. *Weekly Epidemiological Surveillance Report*, 49, 37. [In Thai]

[30] Payaprom, Y. and Suwannakeeree, W., 2018. Promoting Influenza vaccination among the risk population. *Journal of Nursing and Health Sciences*, 12, 13-23. [In Thai]

[31] Sato, A.P.S., Antunes, J.L.F., Moura, A.R.F., Andrade, F.B., Duarte, Y.A.O. and Lebrao, L., 2015. Factors associated to vaccination against Influenza among elderly in a large Brazilian metropolis. *PLOS ONE*, 10(4): e0123840, https://doi.org/10.1371/journal.pone.0123840

Current Applied Science and Technology Vol. 20 No. 2 (May-August 2020)

Effective Treatment of Oil Spills by Adsorbent Formed from Chitin and Polyurethane Foam

Tran Y. Doan Trang* and Zenitova Liubov Andreevna

Kazan National Research Technological University,
Kazan, Tatarstan, Russia

Received: 16 January 2020, Revised: 11 March 2020, Accepted: 3 April 2020

Abstract

Currently, one of the most serious environmental problems is water contamination by oil spills. The use of sorbents is considered one of the most promising approaches to treat this problem. In this study, polyurethane foam (PUF) was used as basal body material and chitin was used as a filler to decrease the cost of the sorbent (PPU10M). The technological parameters of the synthesis process and a heat resistance of the new sorbent were investigated. Three oil-water phases: distilled water, artificial seawater, and artificial river water were tested on the oil adsorption capacity of PPU10M in the ranges of contact time from 5 to 120 min. The results showed that PPU10M had high heat resistance, up to 300^{o}C and it only lost 5% of the mass. There was not much of the synthesis process between the combined sorbent and the primitive PUF. At the same time, the PPU10M reached the high oil adsorption capacity in all three oil-water phases; oil adsorption capacity reached 13.78 $g \cdot g^{-1}$ distilled water; 14.96 $g \cdot g^{-1}$ artificial seawater and 14.46 $g \cdot g^{-1}$ artificial river water. The removal percentage obtained was about 55-59% The study on adsorption kinetic of crude oil for adsorbent PPU10M showed that the pseudo-second-order model was best suitable for the crude oil adsorption process with the correlation coefficient $R^2 = 0.998$. The combined sorbent can be used up to 15 cycles while the oil capacity did not change significantly. The amount of regenerated oil reached up to 98%. Therefore, it showed that PPU10M was a potential oil adsorbent with many advantages for the removal process of oil pollution from water.

Keywords: adsorbent, adsorption capacity, chitin, oil spills, polyurethane foam
DOI 10.14456/cast.2020.19

*Corresponding author: Tel.: +7 96 25 49 47 56
E-mail: tydtrang@gmail.com

1. Introduction

Nowadays, the use and transportation of oil and oil products have increased rapidly so that oil spills and leakage accidents occur regularly all over the world. It has been estimated that the sea environment was polluted by more than 100 million gallons each year [1]. Such accidents can cause severe pollution of waters and our living environment, and they are also one of the causes of significant loss of valuable resources [2]. So, it is urgently needed to treat the oil spills in the water environment [3]. Different methods have been developed for removing oil from water and can be classified as mechanical, chemical, physicochemical, and biological [4]. Among these, sorbents are considered as one of the most promising approaches. Because they facilitate the sorption process given their large sorption capacity, high uptake rate, and potential recyclability [3, 5-7]. According to Adebajo *et al*. [8], the adsorbents for oil spill removal include three main groups: inorganic mineral sorbents, natural sorbents and synthetic sorbents. Generally, inorganic mineral products were not widely used for the treatment of oil pollution from the water environment because of their low buoyancy and absorbability in comparison to the other two types [9]. Natural materials showed low uptake, poor recyclability and buoyancy [1]. The adsorption capacity of synthetic polymer type was significantly higher than the other groups, but they had a high price and degrade very slowly. Moreover, the effects of this type of sorbent on environment and ecology were still unclear. Therefore, it was important to develop environmental-friendly oil absorbents with high absorption and selectivity, reasonable price as well as high recycling ability [2].

Recently, PUF (polyurethane foam) has attracted the attention of many scientists; its application was very wide in various fields such as: in the wastewater treatment and environmental protection, in pharmaceuticals, medicine, agriculture, or in biotechnology, and the construction industry. However, when they are saturated such sorbent has lost the ability to float making it prone to sink and therefore leading to secondary pollution for water environment [10]; and specifically, its cost also was very expensive. Increasing its buoyancy in the saturated state and decreasing the cost by using fillers from natural materials with low density was considered as a practical and possible way [10]. In there, chitin was one of the inexpensive and available natural materials. It was contained mainly in the crustacean's shells especially of crab, shrimp, and lobster, in insects, mollusks and fungal cell walls [11]. The natural sorption materials had more advantages such as low price, friendly with environment, and biodegradability but they had lower adsorption capacity and hydrophobicity in comparison to the synthetic ones [10]. To overcome these disadvantages, in this paper, synthetic materials were combined with natural sorbent materials.

In this work, a combined sorbent basic on the PUF and chitin was developed. The as-obtained sorbent was shown to be the significantly high oil adsorption capacity. Besides, this sorbent also had good buoyancy, can be reused and an increase of the amount recover oil after the sorption process. Notably, these sorbents had heat resistance over a wide range of temperatures. In the limit of this study, the adsorption in three oil-water phases: distilled water, artificial seawater, and artificial river water was investigated.

2. Materials and Methods

2.1 Materials

Chitin purchased from Co., Ltd Chitosan Vietnam (Ho Chi Minh City, Vietnam). Chitin flakes presented particles with size 1-3 mm, obtained from shrimp shells, were a white-yellowish color, moisture – 2.54%. The material was packed in sealed bags, stored in low-light and dry conditions

at the laboratory of Department Synthetic Rubber Technology, Polymer Institute, KNITU university (Kazan, Tatarstan, Russia).

Component A (elastic) and component B (elastic) (Dow-Izolan Co., Ltd. Vladimir, Russia) were used to make PUF. To simulate the oil spills at the water surface, a crude oil sample was used which has been produced from the Romashkinskoe region (Tatarstan, Russia), with density at 20°C, 0.894 $g \cdot cm^{-3}$, and kinematic viscosity at 20°C, 55.7 MPa.s.

2.2 Synthesis sorbent based on chitin and PUF

To synthesize the sorbent based on chitin and PUF (PPU10M), component A was mixed with chitin by agitator before the two components were mixed at room temperature. The weight ratio of component A and component B was 100:60. The rising ratio of the foaming for the adsorbent with a dose of 10% filler was observed best in the comparison in the other ratios of the filler. Moreover, the adsorbent with a dose of 10% chitin had the highest oil sorption capacity when compared with the other ratios of chitin [12]. Therefore, in this work chitin with size 1-3 mm was added to 10% total mass of PUF components. After the polymerization reaction and expansion, there was a 24 h cure time. Then the sorbents were cut with dimensions of 10x10x10 mm to determine the sorption capacity. The determination of the apparent density of sorbent was based on GOST 409-77 [13]. To study the effect on the thermal resistance of PPU10M, thermogravimetric analysis (TGA) was performed by a thermogravimetric analyzer (STA6000, PerkinElmer, USA) with heating rate of $3°C \cdot min^{-1}$, and the temperature range was set up between 30 - 300°C.

2.3 Sorption capacity experiment

2.3.1 Determination of the adsorption capacity in the system containing only oil or water

Two grams of the prepared adsorbents were placed into a beaker that contained 50 ml of the adsorbate. The adsorption process of the sorbent PPU10M was performed in 120 min. Next, the sorbent sample was removed from the beaker, drained for 2 min and weighed. The adsorption capacity of the PPU10M sample was determined according to the following equation (1):

$$\text{Oil/water adsorption capacity} = \frac{m_1 - m_0}{m_0} \quad (1)$$

where: m_1 - the mass of sorbent after adsorption, g; m_0 - the mass of initial sorbent, g.

2.3.2 Determination of the adsorption capacity in a simulated oil spill in the water

To simulate oil spills removal from sea and river, three different water phases such as artificial seawater, artificial river water, and distilled water were utilized in this experimental work. Artificial seawater was prepared according to method of ASTM D1141-98 [14], artificial river water was prepared according to Ruben *et al*. [15].

An oil-water mixture (50 g oil/250 ml water, the height of oil layer, d = 13 -14 mm) was placed in a 500 ml beaker and then 2 g sorbents (PPU10M) was placed in this mixture. After 5, 15, 30, 60, 90, 120 min of contact, the sorbent was withdrawn from the solution and the total mass of PPU10M after sorption was determined on the analytical balance after draining for 2 min.

Since the water was also incorporated into the sorbent material, further experiments were conducted to determine the water uptake of the adsorbent. To extract the adsorbed oil and water, 100 ml toluene was added into the remaining oil-water after the adsorption.

After that, the remaining water was extracted from the mixture by the separating funnel. The mass of the remaining water was weighed. Water uptake and oil uptake of the sorbent was determined using the equations (2), (3) and (4); oil removal percentage of adsorption PPU10M was determined by equation (5):

$$Q_w = \frac{M_{w0}\text{-}M_{w1}}{M_{c0}} \quad (2) \qquad Q_T = \frac{M_{c1}-M_{c0}}{M_{c0}} \quad (3)$$

$$Q_o = Q_T - Q_w \quad (4) \qquad R = \frac{M_{o0}-M_{o1}}{M_{o0}}.100\% \quad (5)$$

where: Q_w - water uptake of sorbent, $g \cdot g^{-1}$; Q_T - total uptake of sorbent, $g \cdot g^{-1}$; Q_o - oil uptake of sorbent, $g \cdot g^{-1}$; R - oil removal percentage of sorbent, %; M_{w1} - mass of water remaining after the sorption, g; M_{w0} - mass of the initial water, g; M_{c1} - total mass of the sorbent samples after sorption, g; M_{c0} - mass of the initial dry sorbent material, g; M_{o1}- mass of oil remaining after the sorption, g, M_{o0} - mass of the initial oil, g; respectively.

2.4 Kinetic adsorption for the sorbent PPU10M

2.4.1 Pseudo-first-order model

The pseudo-first-order model suggested that the uptake rate of the adsorbate with time was directly related to the available amount of the active surface sites [16]. The pseudo-first-order model was expressed as follow:

$$q_t = q_e - q_e e^{-k_1 t} \quad (6)$$

The linear of the pseudo-first-order model was described by equation (7):

$$\ln(q_e - q_t) = \ln q_e - k_1 t \quad (7)$$

where q_t and q_e - the adsorption capacities of the adsorbent at time t and at equilibrium, g/g; k_1 - the pseudo-first-order rate constant, min^{-1}.

2.4.2 Pseudo-second-order model

The pseudo-second-order kinetic model suggested that the chemical adsorption between sorbate and sorbent [17]. The form of the pseudo-second-order kinetic model was described by the following equation:

$$q_t = \frac{k_2 q_e^2 t}{1 + k_2 q_e t} \quad (8)$$

The linear expression of the pseudo-second-order model may be represented by equation:

$$\frac{t}{q_t} = \frac{1}{k_2 q_e^2} + \frac{t}{q_e} \quad (9)$$

where k_2 - the constant rate of the pseudo-second sorption, $g \cdot g^{-1} \cdot min^{-1}$; the initial adsorption rate - h ($g \cdot g^{-1} \cdot min^{-1}$) was expressed as following equation:

$$h = k_2 q_e^2 \quad (10)$$

2.4.3 The intraparticle diffusion model

The intraparticle diffusion model was used to check which diffusion process of solute occurred on the adsorbent: (1) diffusion of the solute through the interfacial film that surrounds the adsorbent (film diffusion); (2) diffusion of the solute to the pores and/or along of the porous walls (intraparticle diffusion) and (3) adsorption/desorption of solute in/out active sites of the adsorbent [18].
The linear of the model be expressed by the equation:

$$q_t = k_p t^{0.5} + C \quad (11)$$

Where k_p - the intraparticle diffusion rate constant, $g \cdot g^{-1} \cdot min^{-1}$; C - a constant related to the bounding layer thickness.
If the relationship between $t^{0.5}$ and qt was observed linear and this line passed through the origin, so it was possible to confirm that the adsorption process occurred due to the intraparticle diffusion [18].

2.4.4 Reusability of the combined sorbent PPU10M

Two grams of the prepared adsorbent were placed in a beaker containing 50 ml of oil. After 60 min, the saturated sorbents were removed, drained about 2 min, then wrung out and weighed. The sorption/desorption cycle was repeated 15 times to evaluate the reusability of the sorbent.

2.4.5 Statistical analysis of the data

The data experiments were calculated by using of Microsoft Excel. Each experiment was repeated three times. The linear regression analysis of the adsorption kinetic models was done through the least square method Microsoft Excel to directly estimate the model parameters from the linear equations.

3. Results and Discussion

3.1 Some properties of the original materials and combined sorbent

In this study, elastic sorbent based on PUF and chitin was investigated on the sorption capacity for removal oil spills from the sea and river. High filler content in the original PUF was impractical because this led to a change to some original properties of sorbent [19]. Therefore, the rate of filled chitin content of 10% (overweight total of PUF) was studied, with the purpose of a decrease in the cost of original sorbent, but maintained high oil sorption ability and buoyancy in water. Some properties of original PUF, chitin, and PPU10M were obtained as shown in Table 1.

The data of Table 1 presented that both start time and rise time of PPU10M were higher in comparison to the primitive PUF, but in general, there was not much change. This difference in the technological parameters was clearly, due to the adding of chitin led to hindering the contact between component A and component B, therefore, it took more time for the reaction of the synthetic adsorbent. The apparent density of PPU10M increased significantly due to the amount of chitin filled into the original PUF [19].

When chitin was used as filler with 10% content, its oil adsorption capacity was lower than the adsorbent PUF without filler; but was higher than the oil adsorption capacity of chitin.

For the removal of oil spills on the water surface, it was necessary to notice the water adsorption ability of the adsorbent. The lower the water adsorption efficiency, the greater the float ability of the adsorbent and the more effective process for the oil spill removal. The results showed that the combined adsorbent created basically on chitin and PUF significantly reduced the water adsorption capacity in comparison with PUF and chitin (water uptake of PPU10M, chitin, and PUF was 5.79 $g \cdot g^{-1}$; 6.34 $g \cdot g^{-1}$; 12.69 $g \cdot g^{-1}$, respectively). The selectivity of the adsorption was characterized by the coefficient a and was calculated by the formula:

$$a = \frac{\text{Water uptake}}{\text{Oil uptake}} \quad (12)$$

Table 1. Some properties of adsorbents original PUF, chitin and PPU10M

Parameters	**Original PUF**	**Chitin**	**PPU10M**
Start time, s	29	-	30-34
Rise time, s	156	-	204
Density, $kg \cdot m^{-3}$	59.10	13.12	69.91
Oil uptake, $g \cdot g^{-1}$ (*)	18.77	9.32	13.13
Water uptake, $g \cdot g^{-1}$ (**)	12.69	6.34	5.79
Coefficient a (equation 12)	0.68	0.68	0.45
Max weight loss at 300°C, %	5	13	5

(*), (**) - used oil adsorption test in the system contains only oil or only water.

In this case, the smaller the coefficient a, the greater the selection characteristic of oil and water uptake. The coefficient a in Table 1 presented clearly that the PPU10M sample creating on chitin and PUF contributed to increasing the selection characteristic of oil and water sorption in comparison to both original materials. Figure 1b showed that three different active functional groups in the urethane layer presented potential bonding sites for water, namely ether (*1), amine (*2) and carbonyl (*3). The hydroxyl and amine groups were both strongly polar molecules, that form strong hydrogen bonds with either hydrogen or oxygen molecules in water [20]. Ether, and carbonyl groups are all polar functional groups that are relatively weaker than hydroxyl and amine groups. But they still formed hydrogen bonds with water molecules [20]. Therefore, both PUF and chitin materials had "attraction" to water. The nature of the PUF-forming reaction was a combination of hydroxyl and isocyanate functional groups (Figure 1c). The hydroxyl groups in the chitin structure (Figure 1a) can have partly or completely participated in the formation of urethane bonds. Therefore, the structure of PPU10M was bulkier; leading to the more difficult to form bonds with water when compared to the original PUF and chitin.

Oiled-water pollution could occur in geographic areas with different various temperatures; therefore, the heat resistance of adsorbents was also a factor necessary to be considered. In this report, thermogravimetric analysis (TGA) was performed with the temperature range up to 300°C. The mass loss of adsorbents corresponded to their heat resistance. The results (Table 1) showed that the combined adsorbent PPU10M contributed to improving the heat resistance of chitin. Specifically, at 295°C the original chitin lost up to 13% of the mass, while the loss of adsorbent PPU10M was only 5% due to including the dehydration of the rings and decomposition of the acetylated and deacetylated units in chitin, [21, 22] that were less stable than the urethane linkages [23]. It showed that the filling of chitin to PUF did not affect the heat resistance of PUF but also improved thermal stability when comparing to the original chitin.

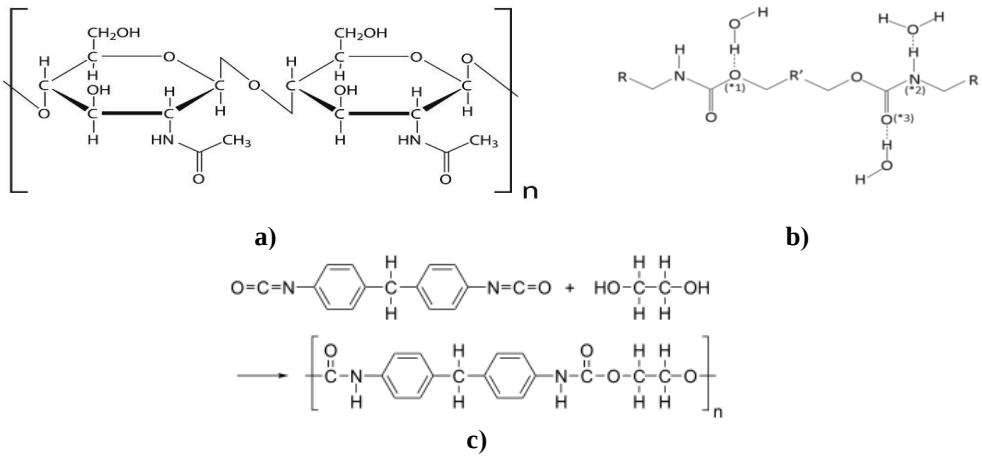


Figure 1. a) Formula of Chitin, b) Potential bonding location on the polymer of the urethane with water and c) The PUF-forming reaction

3.2 The adsorption capacity of the combined sorbent PPU10M

Oil spills often occurred in a water environment, so it was necessary to evaluate the adsorption ability of adsorbent in oil-water phases. A trial study of simulating oil spill with the thickness of oil layer, d = 5 mm, was conducted before studying the effect of adsorption time. The result was shown in Figure 2. As presented in section 3.1, because the combined adsorbent PPU10M had better selective adsorption to oil than water (coefficient a-Table 1). Therefore, when using PPU10M to remove oil spills, it first adsorbed priority the oil and then adsorbed the water. Thus, Figure 2 showed the significantly high potential of adsorbent sample PPU10M for treatment and remove oil spills. On the other hand, from Figure 2, it was observed that this adsorbent had high buoyancy in water, this solved the existing disadvantages of chitin and PUF - the lost the floating ability in the saturated state [24]. With the high buoyancy of PPU10M adsorbent, thus it was easy to remove by the oil-containing sorbents after the adsorption process by the simple and cheap mechanical methods such as fishing rackets or net or mesh screen.

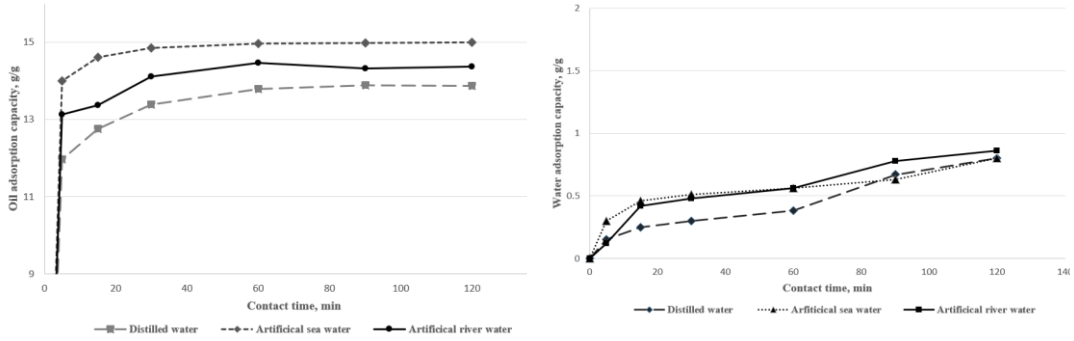
(a) Before adsorption (b) During adsorption process (c) After adsorption

Figure 2. The adsorption ability of PPU10M in oil-water phases with $d_{\text{oil layer}}$ = 5 mm

The contact time significantly affects adsorption capacity. Also, contact time can influence the economic efficiency of the treatment process of oil spills. Therefore, contact time is another performance governing factor in the adsorption process [25]. In this study, the adsorption process for the treatment of oil spills was evaluated from 5 to 120 min of the contact time for all three different oil-water phases: distilled water, artificial seawater and artificial river water (Figure 3).

Figure 3 showed the influence of contact time on the oil capacity of the PPU10M sample. For all types of oil spills, the speed of oil adsorption significantly increased in the first 30 min because of a large concentration difference in this stage. During the next 30 min, the concentration difference decreased so the oil sorption rate gradually lower. After 60 min, the oil adsorption reached into the balance and when increasing the contact time, the water adsorption capacity of adsorbent increased significantly. Therefore, the optimal contact time of the adsorption process was 60 min. The adsorption capacity of PPU10M adsorbent was explained due to the large number of pores in the structure of the adsorbent. Meanwhile, crude oil had a high viscosity and a hydrophobic, so it was quickly absorbed on the surface of the materials through intermolecular forces. Next, crude oil continued to transit into the capillary pores of PPU10M. At this time the pores in the adsorbent were occupied by crude oil, the diffusion of crude oil then became slower, so the amount of adsorbent tended to increase very slowly and reached the equilibrium state.

The results in Table 2 and Figure 3 presented that the oil in artificial seawater and artificial river water reached greater oil uptake than water uptake. The presence of salts in the composition of the seawater and river water can increase the electrical double layer between the adsorbate and adsorbent materials, thereby leading to an increased oil adsorption capacity [26, 27]. In general, the oil adsorption capacity among three types of water was not much different, ranged from 0.68-1.18 g/g. It showed that the chemical compositions of the water phases had not significantly affected the efficient removal of the oil spill by the basic combined adsorbent on PUF and chitin. On the other hand, the spill cleanup ability of an adsorbent was evaluated by oil removal percentage (equation 5). The results of this study showed that this combined adsorbent had a relatively high oil removal percentage, up to 55-59% for the height of the oil layer 13-14 mm. Therefore, for removal of oil spills in rivers as well as at sea, material PPU10M was a potential adsorbent with a high percentage of oil removal and also was possibly applied in different water environments.



(a) Oil adsorption capacity (b) Water adsorption capacity

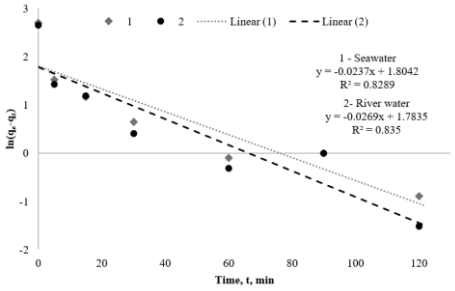
Figure 3. The relationship between the contact time and adsorption capacity of PPU10M

Table 2. Adsorption capacity and efficiency to remove oil spill of PPU10M in different water environment

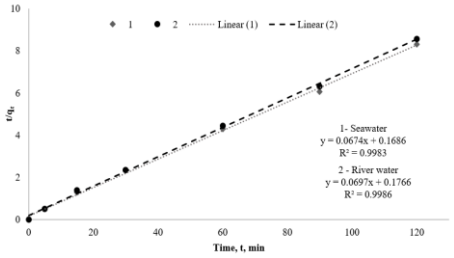
Type water	Oil uptake, $g \cdot g^{-1}$	Water uptake, $g \cdot g^{-1}$	Oil removal percentage, %
Distilled water	13.78	0.38	55.49
Artificial seawater	14.96	0.56	58.27
Artificial river water	14.46	0.56	58.61

3.3 Adsorption kinetic

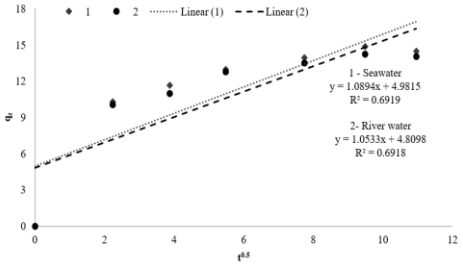
Valuable information on the mechanism of the adsorption process was provided when studying the adsorption kinetics model. Thus, it was important to study the kinetic of crude oil adsorption on the adsorbent PPU10M [28]. In fact, oil pollutions and spills only occur in the environment of the river and seawater. In this study, adsorption kinetic of crude oil for adsorbent PPU10M in two artificial river and seawater phases following the pseudo-first-order model, the pseudo-second-order model and the intraparticle diffusion model were used to fit the data of experiments. The linear equation, correlation coefficient R^2 and parameters of kinetics models were calculated and presented in Figure 4 and Table 3.



a. Pseudo-first-order model



b. Pseudo-second-order model



c. Intraparticle diffusion model

Figure 4. Kinetics of crude oil adsorption for adsorbent PPU10M in artificial seawater and river water phases

Table 3. Kinetic parameters of crude oil adsorption for sorbent PPU10M in artificial seawater and river water

Kinetic model	Seawater		River water	
	Parameters	**Values**	**Parameters**	**Values**
Pseudo-first-order model	$q_{e,1}$	6.075	$q_{e,1}$	5.951
	k_1	0.024	k_1	0.027
	R^2	0.8289	R^2	0.8350
Pseudo-second-order model	$q_{e,2}$	14.837	$q_{e,2}$	14.347
	k_2	0.027	k_2	0.028
	R^2	0.9983	R^2	0.9986
Intraparticle diffusion model	k_p	1.089	k_p	1.053
	C	4.982	C	4.810
	R^2	0.6919	R^2	0.6918

For the pseudo-first-order model and intraparticle diffusion model, the correlation coefficients R^2 were only 0.692 to 0.83, did not approach the $R^2 = 1$, and the value q_e of the model and of q_e of the experimental data were large difference ($q_{e,exp}$ = 14.46 $g \cdot g^{-1}$ and $q_{e,1}$ = 6.075). Therefore, we considered that the pseudo-first-order equation and intraparticle diffusion model were not suitable for the crude oil adsorption process of the sorbent PPU10M. The results revealed the best fit in the pseudo-second-order kinetic equation, the correlation coefficient R^2 reached 0.9983 for seawater and 0.9986 for river water, respectively. Besides, the values q_e of experimental data and of the pseudo-second-order model were approximately equal. For the comparison of the correlation coefficient R^2 and the values of qe, it was observed that the pseudo-second-order model gave a better fit for crude oil adsorption on the sorbent PPU10M with both the artificial seawater and river water environments. This means that the crude oil adsorption of the sorbent PPU10M had the appearance of the chemical interaction of adsorbates on the surface of the adsorbent [17]. Following the pseudo-second-order model, the initial adsorption rate h of crude oil on the sorbent PPU10M was calculated by equation (10) and reached from 5.76 (river water) to 5.94 (seawater), $g \cdot g^{-1} \cdot min^{-1}$.

3.4 Reusability

One of the factors affected the cost of adsorbents is its reusability. The reuse of adsorbents was one of the methods to significantly reduce the cost of treating water from oil pollution and accident oil spills. Also, after the adsorption process, the contaminated sorbents were one of the reasons to increase the amount of solid waste in the ecosystem. Thus, it was considered that the reusability of the sorbent was one of the measures to reduce waste entering the environment (Figure 5).

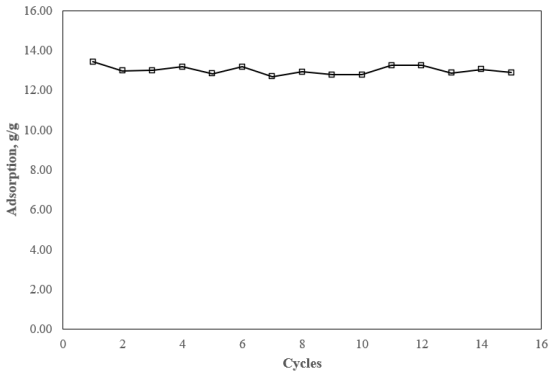


Figure 5. Reusability of the sorbent PPU10M

It was revealed that the PPU10M adsorbent with almost unchanged capacity can be used up to 15 cycles and an oil capacity reached ~ 13.00 $g \cdot g^{-1}$. It led to a significant reduction in the cost of the treatment of oil spills and pollution. At the same time, the amount of regenerated oil reached 98%. According to statistics, the world's oil reserves were becoming increasingly depleted, meanwhile, oil was an indispensable source of energy at present. The regeneration of absorbed oil after the adsorption process led to reducing the effects of oil depletion. At the same time, the regeneration of absorbed oil after the adsorption process also reduced the amount of oil entering into the environment. Thus, it led to reduce the negative effects of oil spills on humans and the environment. On the other hand, the proposed mechanical method for the recovery of absorbed oil was one of the simple and inexpensive methods. Therefore, it could be considered that oil sorbent PPU10M was a potential, economical and environmental friendly material.

Besides, the economic-benefit calculation of the combined adsorbent comparing with the original adsorbent was necessary. The addition of fillers to the component PUF can lead to a significant increase in the amount of the obtained adsorbent, thus its cost is reduced in comparíion to the original PUF. The calculation of the prices for materials showed that the use of filler-chitin into PUF components decreased about 870 $/ton materials when compared to PUF without fillers. This proved that the chitin-filling to PUF was meaningful in reducing the cost of the expensive PUF. Therefore, it showed that PPU10M was a potential oil adsorbent with many advantages for removal of oil pollution from water environments such as high oil adsorption capacity, selective adsorption to oil higher than water, high buoyancy at saturated state, well heat resistance, possible applications in many different water environments and especially more economical than PUF material.

4. Conclusions

In this study, the new sorption material PPU10M was created basically on PUF and chitin, which has been used widely for the treatment of oil spills due to its high adsorption capacity. There was nearly no difference in the technological parameters of the synthesis process between the sorbent material using chitin and the primitive PUF without filler. Especially, it had a great selection characteristic between oil and water adsorption and high oil adsorption which had an excellent prospect in practical application in comparison to the primitive PUF and chitin, where oil uptake reached 13.13 $g \cdot g^{-1}$ and water uptake was 5.79 $g \cdot g^{-1}$. Besides, PPU10M also had high heat

resistance, up to 300ºC and it only lost 5% of the mass. The use of the obtained sorbent for oil treatment was investigated in various conditions of water phases such as distilled water, artificial seawater and river water. The results showed that the oil absorption of this new kind of adsorbent was 13.78 g in distilled water, 14.96 g in artificial seawater and 14.46 g in artificial river water, respectively. Its removal percentage reached about 55-59%. At the same time, the combined adsorbent PPU10M reached the high oil adsorption capacity in all three oil-water phases: distilled water, artificial seawater and artificial river water, and the adsorption ability was not much different. The crude oil adsorption process of the sorbent PPU10M followed the application of the pseudo-second-order model for both artificial seawater and river water environments. This means that the crude oil adsorption process of PPU10M occurred the chemical interaction of adsorbates on the surface of the adsorbent. The reusability of sorbent was studied and the result showed that the combined sorbent of basic PUF and chitin could use up to 15 cycles without changing of oil capacity and reached about 13.00 $g \cdot g^{-1}$. The amount of regenerated oil reached up to 98%. It was shown that PPU10M was a new combined adsorbent that had the high removal of oil spills in rivers and also on the sea with many advantages such as high oil adsorption capacity, buoyancy and heat resistance ability. It can also be recycled and is environmentally friendly. Therefore, this was a viable adsorbent that can be widely used for oil spills removal.

References

[1] Guo, M.Y., Wang, H.W., Wang, Y.H., Si, Q.Y., Guo, Z.Z., Wei, X.F. and Song, H.H., 2014. Preparation of a new oil absorbing polyurethane foam. *Applied Mechanics and Materials*, 618, 131-135.

[2] Wu, Z.Y., Li, C., Liang, H.W., Zhang, Y.N., Wang, X., Chen, J.F. and Yu, S.H., 2014. Carbon nano fiber aerogels for emergent cleanup of oil spillage and chemical leakage under harsh conditions. *Scientific Reports*, 4 (4049), 1-6.

[3] Dong, X., Chen, J., Ma, Y., Wang, J., Chan-Park, M.B., Liu, X. and Chen, P., 2012. Superhydrophobic and superoleophilic hybrid foam of graphene and carbon nanotube for selective removal of oils or organic solvents from the surface of water. *Chemical Communications*, 48, 10660-10662.

[4] Santos, O.S.H., Coelho, M.D.S. and Yoshida, M.I., 2017. Synthesis and performance of different polyurethane foams as oil sorbents. *Journal of Applied Polymer Science.*, 2017, 1-8.

[5] Li, A., Sun, H.X., Tan, D.Z., Fan, W.J., Wen, S.H., Qing, X.J. and Deng, W.Q., 2011. Superhydrophobic conjugated microporous polymers for separation and adsorption. *Energy and Environmental Science*, 4, 2062-2065.

[6] Zhu, Q., Pan, Q. and Liu, F., 2011. Facile removal and collection of oils from water surfaces through superhydrophobic and superoleophilic sponges. *Journal of Physical Chemistry*, 115, 17464-17470.

[7] Bandura, L., Woszuk, A., Kołodyńska, D. and Franus, W., 2017 .Application of mineral sorbents for removal of petroleum substances :A review .*Minerals*, 7(37), 1-25 .

[8] Adebajo, M.O., Frost, R.L., Kloprogge, J.T., Carmody, O. and Kokot, S., 2003. Porous materials for oil spill cleanup: A review of synthesis and absorbing properties. *Journal of Porous Materials*, 10, 159-170.

[9] Wahi, R., Chuah, L.A., Choong, T.S.Y., Ngaini, Z. and Nourouzi, M.M., 2013. Oil removal from aqueous state by natural fibrous sorbent: An overview. *Separation and Purification Technology*, 113, 51-63.

[10] Hoang, P.H., Hoang, A.T., Chung, N.H., Dien, L.Q., Nguyen, X.P. and Pham, X.D., 2017. The efficient lignocellulose-based sorbent for oil spill treatment from polyurethane and

agricultural residue of Vietnam. *Energy Sources, Part A :Recovery, Utilization, and Environmental Effects*, 40 (3), 1-8 .

[11] Tri, Dang Le Minh 2012. *Study on the Adsorption of Reactive Dyestuffs in Textile Wastewater by Dyeing Chitosan Derived from Shrimp Shells.* Master Thesis. Hanoi National University.

[12] Trang, T.Y.D. and Zenitova, L.A., 2019. Study of the sorption ability of a sorbent for the elimination of oil spills based on polyurethane foam and chitin. *Bulletin of PNIPU. Chemical Technology and Biotechnology*, 2, 33-47.

[13] GOST 409-77, 2002. *Interstate Standard-Cellular Plastics and Rubber Sponge, Method for Determining the Appearance of Density*. Moscow: PC Publishing Standards.

[14] ASTM D1141-98, 2013. *Standard Practice for the Preparation of Substitute Ocean Water*. West Conshohocken: ASTM International.

[15] Ruben, A., Nobuyasu, Y., Katsuji, T. and Masao, N., 2003 .Change in the bacterial community of natural river biofilm during biodegradation of aniline-derived compounds determined by denaturing gradient gel electrophoresis. *Journal of Health Science*, 49(5), 379-385.

[16] Khalifa, R.E., Omer, A.M., Tamer, T.M ., Ali, A.A., Ammar, Y.A. and Mohy, E.M.S., 2019. Efficient eco-friendly crude oil adsorptive chitosan derivatives : kinetics, equilibrium and thermodynamic studies. *Desalination and Water Treatment*, 2019, 1-13.

[17] Hao, X., Liu, H., Zhang, G., Zou, H., Zhang, Y., Zhou, M. and Gu, Y., 2012. Magnetic field assisted adsorption of methyl blue onto organo-bentonite. *Applied Clay Science*, 55, 177-80.

[18] Vinhal, J.O., Nege, K.K., Lage, M.R., de M. Carneiro, J.W., Lima, C.F. and Cassella, R.J., 2017. Adsorption of the herbicides diquat and difenzoquat pn polyurethane foam : Kinetic, equilibrium anf computational studies. *Ecotoxicology and Environmental Safety*, 145, 597-604.

[19] Kuen, T.Q.A., Ivanova, M.A. and Zenitova, L.A., 2017. Polymer composition based on foam polyurethane and chitosan. *Bulletin of the Technological University*, 20(11), 32-35.

[20] Yukitoshi, T., Ethan, B., Seizo, S., Takashi, M. and Takashi, S., 2014. States of water adsorbed in water-borne urethane/epoxy coatings. *Polymer*, 55, 2505-2513.

[21] Tanodekaew, S., Prasitsilp, M., Swasdison, S., Thavornyutikarn, B., Pothsree, T. and Pateepasen, R., 2004. Preparation of acrylic grafted chitin for wound dressing application. *Biomaterials*, 25, 1453-1460.

[22] Khairkar, S.R. and Raut, A.R., 2014. Synthesis of chitosan-graft-polyaniline-based composites. *American Journal of Materials Science and Engineering*, 2 (4), 62-67.

[23] Matsui, M., Munaro, M. and Akcelrud, L., 2010. Chitin/polyurethane blends: a thermal and morphological study. *Polymer International*, 59, 1090-1098.

[24] Francisco, C. de F.B., Luiz, C.G.V., Tecia, V.C. and Ronaldo, do N., 2014. Removal of petroleum spill in water by chitin and chitosan. *The Electronic Journal of Chemistry*, 6 (1), 70-74.

[25] Iftekhar, S., Ramasamy, D.L., Srivastava, V., Asif, M.B. and Sillanpää, M., 2018. Understanding the factors affecting the adsorption of Lanthanum using different adsorbents: a critical review. *Chemosphere*, 204, 413-430.

[26] Bjelopavlic, M., Newcombe, G. and Hayes, R., 1999. Adsorption of NOM onto activated carbon: effect of surface charge, ionic strength, and pore volume distribution. *Journal and Interface Science*, 210, 271-280.

[27] Drikas, M., 1997. Adsorption of NOM onto activated carbon: electrostatic and non-electrostatic effects. *Carbon*, 35, 1239-1250.

[28] Sidik, S.M., Jalil, A.A., Triwahyono, S., Adam, S.H., MSatar, M.A.H. and Hameed, B.H., 2012. Modified oil palm leaves adsorbent with enhanced hydrophobicity for crude oil removal. *Chemical Engineering Journal*, 203, 9-18.

Finsler Metrics Induced by a Similarity Function

Nisachon Kumankat[1]*, Praiboon Pantaragphong[1] and Sorin V. Sabau[2]

[1] Department of Mathematics, Faculty of Science, King Mongkut's Institute of Technology Ladkrabang, Bangkok, Thailand
[2] School of Biological Science, Department of Biology, Tokai University, Sapporo, Japan

Received: 2 December 2019, Revised: 21 March 2020, Accepted: 2 April 2020

Abstract

In the present paper, the geometrical properties of a topological space endowed with a similarity was studied. Its relation with weighted quasi-metrics and Finsler metrics of Randers type was discussed. Finally, some applications to bioinformatics and computer science by relating similarities to dynamic programming algorithms are considered. In conclusion, the space containing the real-world data is non-symmetric and non-linear.

Keywords: Finsler metrics, similarity function, weighted quasi-metrics
DOI 10.14456/cast.2020.20

1. Introduction

Metric spaces, symmetric distances are used in different fields of pure mathematics like analysis, geometry and so on, as well as in applied mathematics, for instance in computer science, bioinformatics, data analysis, etc. These are the natural generalization of Euclidean and Riemannian spaces. On the other hand, non-symmetric distances, Minkowski norms, and Finsler metrics are also widely used in the analysis, differential geometry, data analysis, etc. We claim that symmetric distances, Euclidean and Riemannian metrics are just convenient approximations (usually obtained by averaging) of the real world. The real world, based on real data measurements is highly non-symmetric and non-linear. Of course, proving such a fact in its most generality is a very complex and difficult task, beyond the purpose of this paper.

However, we will argue that the similarity induced by the dynamic programming algorithm Needleman-Wunsch is actually equivalent in nature to non-symmetric distances (so-called quasi-distances) and Finsler metrics. Our main statement in this paper is that the following motions are equivalent in nature, i.e. symmetric similarity function, weighted quasi-distance and Finsler metrics of Randers type with reversible geodesics.

*Corresponding author: Tel.: +66 91 8517097
E-mail: Nisachon.Kumankat@gmail.com

The topic studied in the present paper is very important for analysis, geometry, computer science, data analysis, bioinformatics and so on because in some sense it shows that at least partially, the reality we are living in, is non-symmetric, non-linear, non-homogeneous and the study of data sets from the real world is actually equivalent to the use of weighted quasi-metrics on topological spaces, or of a Finsler metric of Randers type on smooth manifolds.

2. Similarities and Distances

We start by recalling the following definition [1, 2, 3].

Definition 2.1 Let X be a topological space. If $s : X \times X \to \mathbb{R}$ is a continuous mapping such that

i. $s(x,x) > 0$ for any $x \in X$,
ii. $s(x,x) \geq s(x,y)$ for any $x, y \in X$,
iii. if $s(x,y) = s(x,x)$ and $s(y,x) = s(y,y)$, then $x = y$ for any $x, y \in X$,
iv. $s(x,y) + s(y,z) \leq s(x,z) + s(y,y)$ for any , $y, z \in X$,

Then s is called a *similarity function* on X.
The relation with quasi-metrics is well-known [4, 5].

Definition 2.2 Let X be a non-empty set and d a real-valued function $d : X \times X \to [0, \infty)$ that satisfies:

i. $d(x,y) \geq 0$ and $d(x,y) = 0$ if and only if $x = y$ for any $x, y \in X$,
ii. $d(x,y) \leq d(x,z) + d(z,y)$ for any $x, y, z \in X$,
iii. if $d(x,y) = d(y,x) = 0$ then $x = y$ for any $x, y \in X$.

Then (X, d) is called a *quasi-metric space*.

Definition 2.3 A *weighted quasi-metric space* is a triple (X, d, w), where X is a non-emply set, $d : X \times X \to [0, \infty)$ and $w : X \to [0, \infty)$ that satisfies:

i. $d(x,y) \geq 0$ and $d(x,y) = 0$ if and only if $x = y$ for any $x, y \in X$,
ii. $d(x,y) \leq d(x,z) + d(z,y)$ for any $x, y, z \in X$,
iii. if $d(x,y) = d(y,x) = 0$ then $x = y$ for any $x, y \in X$,
iv. $d(x,y) + w(x) = d(y,x) + w(y)$ for any $x, y, z \in X$.

The function d is called a *quasi-metric*, and w is the *weight function*.

Proposition 2.4 *If $s : X \times X \to \mathbb{R}$ be a similarity function on X, then $d : X \times X \to \mathbb{R}$ defined by*

$$d(x,y) := s(y,y) - s(y,x), \qquad \text{for all } x, y \in X,$$

is a quasi-metric on X.

Proof Let $x, y, z \in X$. We verify the conditions in the definition of the quasi-metric.

i. Positiveness: $d(x,y) = s(y,y) - s(y,x)$,
Since $s(y,y) - s(y,x) \geq 0$, it is clear that $d(x,y) \geq 0$.
And $d(y,y) = s(y,y) - s(y,y) = 0$, then $x = y$.

ii. Triangle inequality:

$$\begin{aligned} d(x,y) &= s(y,y) - s(y,x) \\ &= s(y,y) - s(y,z) + s(y,z) - s(y,x) \\ &\leq s(y,y) - s(y,z) + s(z,z) - s(z,x) \\ &= d(z,y) + d(x,z) \\ &= d(x,z) + d(z,y). \end{aligned}$$

iii. Separation axiom:

$$\begin{aligned} d(x,y) = 0 \ \wedge \ d(y,x) = 0 &\Longrightarrow x = y \\ s(y,y) - s(y,x) = 0 \ \wedge \ s(x,x) - s(x,y) = 0 &\Longrightarrow x = y \\ s(y,y) = s(y,x) \ \wedge \ s(x,x) = s(x,y) &\Longrightarrow x = y. \end{aligned}$$

Therefore, (X, d) is a quasi-metric on X.

Proposition 2.5 *Let $s : X \times X \to \mathbb{R}$ is a similarity function on X. If s is a symmetric, i.e. $s(x,y) = s(y,x)$ for all $x, y \in X$, then (X, d) is a weighted quasi-metric space with weight function $w : X \to \mathbb{R}$, $w(x) = s(x,x)$.*

Proof By proposition 2.4, we have $d(x,y) = s(y,y) - s(y,x)$.
Let $x, y, z \in X$. We verify the conditions in the definition of a weighted quasi-metric.

i. Positiveness: $d(x,y) = s(y,y) - s(y,x)$,
since $s(y,y) - s(y,x) \geq 0$, it is clear that $d(x,y) \geq 0$.
And $d(y,y) = s(y,y) - s(y,y) = 0$, then $x = y$.

ii. Triangle inequality:

$$\begin{aligned} d(x,y) &= s(y,y) - s(y,x) \\ &= s(y,y) - s(y,z) + s(y,z) - s(y,x) \\ &\leq s(y,y) - s(y,z) + s(z,z) - s(z,x) \\ &= d(x,z) + d(z,y). \end{aligned}$$

iii. Separation axiom:

$$\begin{aligned} d(x,y) = 0 \ \wedge \ d(y,x) = 0 &\Longrightarrow x = y \\ s(y,y) - s(y,x) = 0 \ \wedge \ s(x,x) - s(x,y) = 0 &\Longrightarrow x = y \\ s(y,y) = s(y,x) \ \wedge \ s(x,x) = s(x,y) &\Longrightarrow x = y. \end{aligned}$$

iv. Let $w : X \to \mathbb{R}$, $w(x) = s(x,x)$, we have

$$\begin{aligned} d(x,y) + w(x) &= s(y,y) - s(y,x) + s(x,x) \\ &= s(x,x) - s(x,y) + s(y,y) \\ &= d(y,x) + w(y). \end{aligned}$$

Therefore, (X, d) is a weighted quasi-metric space.

We will consider in the following only symmetric similarity function.

Remark 2.6 :

1) Observe that the quasi-distance d and the weight function w are determined only by the similarity function s.

2) Let we assume the quasi-distance d is actually a distance function, i.e. $d(x,y) = d(y,x)$ for all $x, y \in X$, it follows
$$s(y,y) - s(y,x) = s(x,x) - s(x,y),$$
and if we take into account that s is symmetric, then we obtain
$$s(x,x) = s(y,y), \qquad \text{for all } x, y \in X.$$
In other words, the quasi-distance induced by a similarity function is a distance if and only if the similarity is the same on the diagonal.

3) The symmetrized distance induced by a quasi-distance d is
$$\rho(x,y) := \tfrac{1}{2}[d(x,y) + d(y,x)] = \tfrac{1}{2}[s(x,x) + s(y,y)] - s(x,y).$$

Conversely, a weighted quasi-metric space induces a symmetric similarity function. Indeed, we have the following proposition hold

Proposition 2.7 *Let (X,d) is a weighted quasi-metric space with the weight function $w : X \to \mathbb{R}$. Then the mapping $s : X \times X \to \mathbb{R}$,*
$$s(x,y) := w(x) - d(y,x), \qquad \text{for all } x, y \in X,$$
is a symmetric similarity function on X.

Proof Let $x, y, z \in X$. We verify the conditions in definition 2.1, we have:

i. Since $w : X \to \mathbb{R}$ is weight function, then $w(x) > 0$ and $d(x,x) = 0$.
It is clear that $s(x,x) = w(x) - d(x,x) = w(x) > 0$.

ii. We will show that $s(x,x) - s(x,y) \geq 0$, we have
$$\begin{aligned} s(x,x) - s(x,y) &= w(x) - d(x,x) - w(x) + d(y,x) \\ &= -d(x,x) + d(y,x) \\ &= d(y,x). \end{aligned}$$
Since $d(y,x) \geq 0$, thus $s(x,x) - s(x,y) \geq 0$, i.e. $s(x,x) \geq s(x,y)$.

iii. Suppose that $s(x,y) = s(x,x)$ and $s(y,x) = s(y,y)$, we have
$$\begin{aligned} s(x,y) = s(x,x) \ \wedge\ s(y,x) = s(y,y) &\implies x = y \\ w(x) - d(y,x) = w(x) - d(x,x) \ \wedge\ w(y) - d(x,y) = w(y) - d(y,y) &\implies x = y \\ d(x,x) - d(y,x) = w(x) - w(x) \ \wedge\ d(y,y) - d(x,y) = w(y) - w(y) &\implies x = y \\ d(x,x) - d(y,x) = 0 \ \wedge\ d(y,y) - d(x,y) = 0 &\implies x = y \\ d(x,x) = d(y,x) \ \wedge\ d(y,y) = d(x,y) &\implies x = y. \end{aligned}$$

iv. We will show that $s(x,y) + s(y,z) \leq s(x,z) + s(y,y)$, we have
$$\begin{aligned} s(x,y) + s(y,z) &= w(x) - d(y,x) + w(y) - d(z,y) \\ &= w(x) + w(y) - d(z,y) - d(y,x) \\ &\leq w(x) + w(y) - d(z,x) \\ &= w(x) - d(z,x) + w(y) - 0 \\ &= w(x) - d(z,x) + w(y) - d(y,y) \\ &= s(x,z) + s(y,y). \end{aligned}$$

Therefore, (X,s) is a symmetric similarity function on X.

Example 2.8 Let us consider the metric space (S,ρ) and the interval $I := (0,\infty)$. It is known that the product space $G := S \times I$ inherits a natural structure of generalized weighted quasi-metric structure (G,Q,W), where
$$Q : G \times G \to I, \quad Q(u,v) := \rho(x,y) + \eta - \xi,$$

$$W : G \to I, \qquad W(u) := 2\xi,$$

for any $u = (x, \xi),\ v = (y, \eta)$ on $G = S \times I$.

The similarity function $\varphi : G \times G \to \mathbb{R}$ induced by the weighted quasi-metric structure (G, Q, W) is given by

$$\varphi(u, v) := -\rho(x, y) + \xi + \eta, \qquad \text{for any } u = (x, \xi),\ v = (y, \eta) \in G.$$

Clearly this is a symmetric similarity function on G.

Indeed, let $u = (x, \xi),\ v = (y, \eta),\ l = (z, \zeta) \in G$, then

i. It is clear that

$$\varphi(u, u) = \varphi\big((x, \xi), (x, \xi)\big) = -\rho(x, x) + \xi + \xi = 2\xi > 0.$$

ii. We will show that $\varphi(u, u) - \varphi(u, v) \geq 0$, we have

$$\begin{aligned} \varphi(u, u) - \varphi(u, v) &= -\rho(x, x) + \xi + \xi + \rho(x, y) - \xi - \eta \\ &= -\rho(x, x) + \rho(x, y) + \xi - \eta \\ &= \rho(x, y) + \xi - \eta \\ &= \rho(y, x) + \xi - \eta \\ &= Q\big((y, \eta), (x, \xi)\big) = Q(v, u) \quad \geq 0. \end{aligned}$$

So, $\varphi(u, u) - \varphi(u, v) \geq 0$.

iii. Suppose that $\varphi(u, v) = \varphi(u, u)$ and $\varphi(v, u) = \varphi(v, v)$, that is

$$\begin{gathered} \varphi\big((x, \xi), (y, \eta)\big) = \varphi\big((x, \xi), (x, \xi)\big) \ \wedge \ \varphi\big((y, \eta), (x, \xi)\big) = \varphi\big((y, \eta), (y, \eta)\big) \\ -\rho(x, y) + \xi + \eta = -\rho(x, x) + \xi + \xi \ \wedge \ -\rho(y, x) + \eta + \xi = -\rho(y, y) + \eta + \eta \\ \rho(x, x) - \rho(x, y) + \eta - \xi = 0 \ \wedge \ \rho(y, y) - \rho(y, x) + \xi - \eta = 0, \end{gathered}$$

by subtracting these two equalities we get, $\rho(x, x) - 2\xi = \rho(y, y) - 2\eta$, whence $x = y$, so that $\xi = \eta$. Hence $u = v$.

iv. We will show that $\varphi(u, v) + \varphi(v, l) \leq \varphi(u, l) + \varphi(v, v)$, we have

$$\begin{aligned} \varphi(u, v) + \varphi(v, l) &= \varphi\big((x, \xi), (y, \eta)\big) + \varphi\big((y, \eta), (z, \zeta)\big) \\ &= -\rho(x, y) + \xi + \eta - \rho(y, z) + \eta + \zeta \\ &= -\rho(x, y) - \rho(y, z) + \xi + \zeta + \eta + \eta \\ &\leq -\rho(x, z) + \xi + \zeta + \eta + \eta \\ &= -\rho(x, z) + \xi + \zeta - \rho(y, y) + \eta + \eta \\ &= \varphi\big((x, \xi), (z, \zeta)\big) + \varphi\big((y, \eta), (y, \eta)\big) \\ &= \varphi(u, l) + \varphi(v, v). \end{aligned}$$

Therefore, φ is a symmetric similarity function on G.

Example 2.9 Let (S, ρ) be a metric space and $f : S \to (0, \infty)$ a Lipschitz function with respect to ρ. Then it is known that the graph of f, i.e. $G_f = \big\{\big(x, f(x)\big) : x \in S\big\}$ has a weighted quasi-metric space structure $\big(G_f, Q, W\big)$ given by

$$\begin{aligned} &Q : G_f \times G_f \to (0, \infty), \quad Q(u, v) := \rho(x, y) + f(y) - f(x), \\ &W : G_f \to (0, \infty), \qquad W(u) := 2f(x), \end{aligned}$$

for any $u = \big(x, f(x)\big),\ v = \big(y, f(y)\big)$ on G_f.

It follows that the function $\varphi_f : G_f \times G_f \to \mathbb{R}$ given by

$$\varphi_f(u, v) := -\rho(x, y) + f(x) + f(y),$$

for any $u = \big(x, f(x)\big),\ v = \big(y, f(y)\big) \in G_f$, is a symmetric similarity function on G_f.

We can conclude that a metric space (X, ρ) with a Lipschitz function $f : X \to \mathbb{R}$ induces a similarity function on X.

The similarity space (G_f, φ_f) constructed here is called the bundle over the metric space (S, ρ).

3. Embeddings and Relation to Finsler space

Let (X, σ) and (Y, τ) be two topoltgical spaces with similarities functions σ and τ, respectively. A continuous mapping $\varphi : X \to Y$ is called a *similarity embedding* if

$$\tau\big(\varphi(x), \varphi(y)\big) = \sigma(x, y),$$

for all $x, y \in X$.

Proposition 3.1 *Let (X, q, w) and (Y, p, u) be two weighted quasi-metric spaces with the associated similarities σ and τ, respectively. The continuous function $\varphi : X \to Y$ is a similarity embedding if and only if it is an embedding of weighted quasi-metric spaces.*

Proof We assume that $\varphi : (X, \sigma) \to (Y, \tau)$ is a similarity embedding, i.e.

$$\tau\big(\varphi(x), \varphi(y)\big) = \sigma(x, y), \qquad \forall x, y \in X.$$

The weighted quasi-metric (d, w) associated with the similarity function σ on X is given by

$$d(x, y) = \sigma(y, y) - \sigma(y, x), \qquad \forall x, y \in X,$$
$$w(x) = \sigma(x, x), \qquad \forall x \in X.$$

The weighted quasi-metric $(\hat{d}, \hat{w})$ associated with the similarity function τ on Y is given by

$$\hat{d}(x, y) = \tau(y, y) - \tau(y, x), \qquad \forall x, y \in Y,$$
$$\hat{w}(x) = \tau(x, x), \qquad \forall x \in Y.$$

We compute

$$\begin{aligned} \hat{d}\big(\varphi(x), \varphi(y)\big) &= \tau\big(\varphi(y), \varphi(y)\big) - \tau\big(\varphi(y), \varphi(x)\big) \\ &= \sigma(y, y) - \sigma(y, x) \\ &= d(x, y), \end{aligned}$$

for all $x, y \in X$. Likewise,

$$\hat{w}\big(\varphi(x)\big) = \tau\big(\varphi(x), \varphi(x)\big) = \sigma(x, x) = w(x),$$

and hence it results that $\varphi : (X, d, w) \to \big(Y, \hat{d}, \hat{w}\big)$ is an embedding of weighted quasi-metrics spaces.

Conversely, we assume that $\varphi : (X, d, w) \to \big(Y, \hat{d}, \hat{w}\big)$ is an embedding of weighted quasi-metrics spaces, i.e.

$$\hat{d}\big(\varphi(x), \varphi(y)\big) = d(x, y),$$
$$\hat{w}\big(\varphi(x)\big) = w(x),$$

for all $x, y \in X$. Using now relations

$$\sigma(x, y) = w(x) - d(y, x), \qquad \forall x, y \in X,$$
$$\tau(x, y) = \hat{w}(x) - \hat{d}(y, x), \qquad \forall x, y \in Y.$$

We have

$$\begin{aligned} \tau\big(\varphi(x), \varphi(y)\big) &= \hat{w}\big(\varphi(x)\big) - \hat{d}\big(\varphi(y), \varphi(x)\big) \\ &= w(x) - d(y, x) \\ &= \sigma(x, y), \end{aligned}$$

for all $x, y \in X$, therefore $\varphi : (X, \sigma) \to (Y, \tau)$ is a similarity embedding.

Theorem 3.2 *Every space with a symmetry function (X, s) is embeddable in a bundle over a suitable metric space (S, ρ).*

Proof The proof is quite straightforward by taking into account the construction in Example 2.9. This result can also be proved directly, using the fact that any weighted quasi-metric space is embeddable in a bundle over a suitable metric space [6, 7].

Theorem 3.3 *1. Let (S, ρ) be a metric space and $f : S \to [0, \infty)$ a Lipschitz function on this metric space. Then the graph of f admits a similarity function $\varphi : G_f \times G_f \to \mathbb{R}$ that depends on ρ and f only.*

2. Conversely, every similarity space (X, s) can be constructed in this way.

Proof (1) Statement 1 follows immediately from Example 2.9.

(2) Conversely, if we start with a similarity space (X, s), then we can consider :

(a) the associated weighted quasi-metric space (X, d, w), where d and w are given in Proposition 2.4 and 2.5.

(b) the symmetrized associated distance s has given in Remark 2.6.

By putting $f := \frac{1}{2}w$, using (X, s) and f, the construction from Statement 1 given that (G_f, φ) is a similarity space.

Moreover, (X, s) can be embedded in (G_f, φ) and the conclusion follows.

Lemma 3.4 *Let M be a compact smooth manifold. If (M, ρ) is an upper and lower curvature bounded metric space, then there exists a Riemannian metric g on M whose distance function coincides with p.*

Proof The proof is quite obvious. Every an upper and lower curvature bounded metric space (M, ρ) constructed on a n -dimensional compact smooth manifold M can be embedded isometrically in the Euclidean space $\mathbb{R}^{2n+1}$ with the canonical metric. On the other hand, the manifold M as a submanifold in $\mathbb{R}^{2n+1}$ inherits a canonical Riemannian metric [8, 9] from the embedding in the Euclidean space whose distance function obviously coincides with ρ.

Theorem 3.5 *1. If $(M, F = \alpha + \beta)$ is a simply connected Randers defined by a Riemannian metric $\alpha = \sqrt{a_{ij}(x)y^i y^j}$ and a closed 1-form β, then M is endowed with a naturally induced similarity function.*

2. Conversely, let s be a symmetric, similarity function defined on a compact differentiable manifold M whose associated distance p is upper and lower curvature bounded. If s is differentiable, then there exists a naturally constructed Randers metric on M that depends on s only.

Proof 1.It is clear since a Randers metric with β closed induces a weighted quasi metric on M.

2.Conversely, the similarity metric induces a weighted quasi metric (M, d, w). From Lemma 3.4 we can see that there exists a Riemannian metric on M whose distance function is exactly ρ, and since s was assumed smooth we can define $\beta := dw$, where w is the weight induced by s.

4. Relation with Bioinformatics and Computer Science

In order to assess the application of this theory, we start by recalling that Dynamic Programming is, at the same time, a mathematical optimization method as well as an algorithmic method in computer science. Dynamic Programming originates in the research of R. Bellman in the 1950s and it was applied eventually in many fields of science like engineering, economics and others. In the majority cases, this method works by simplifying a much more complicated problem by dividing it into much easier small problems using a recursive way. It is known that if a problem in computer science can be solved optimally by dividing it into smaller problems and then recursively determine the optimal solutions to these small problems, then the original problem has an optimal substructure. The algorithms involving Dynamic Programming are popular in the field of bioinformatics being extremely useful for some specific problems as DNA or amino acids sequences alignment, RNA structure prediction, protein structure research, and others [4, 5].

In the case of sequence comparison analysis in Bioinformatics, a similarity measure on Σ together with a gap penalties function can be used to define the global similarity between two sequences in Σ^*. The computation is handled using the Needleman-Wunsch dynamic programming algorithm which is quite similar to the W-S-B algorithm for computation of distances. It is possible to define global similarity using a dynamic programming matrix.

To be more precise, let Σ be a non-empty set. Then a *free monoid* Σ^*on Σ is the monoid whose elements are all *finite* sequences of zero or more elements, from Σ with the operation of *concatenation*. The set $\Sigma = \{A, B, C, \dots, Z\}$ is called *alphabet*, and its elements $A, B, C, \dots, Z$ are called **letters** of the alphabet, or *generators*. The elements $u \in \Sigma^*$ are called *words* or *strings*. The unique sequence of zero letters (the empty word) denoted by e is the *identity element* in Σ^*.

The *free semigroup* Σ^+on Σ is defined as $\Sigma^+ := \Sigma^* \backslash \{e\}$.

Remark 4.1 Biological motivation

The macromolecule that contains the essential information of living cells can be represented as a family of words over a finite alphabet. Consequently, DNA (or RNA) molecules can be seen as long words in the free semigroup with the generators $\Sigma = \{A, C, T, G\}$ (nucleotide alphabet). Proteins molecules can be regarded as words in the free semigroup whose generators are the 20 amino acids which compose the proteins in living cells (aminoacids alphabet) Σ_{AA}.

As an example, we mention here the insulin, whose intensive research, starting around 1950, has facilitated the development of the theory of molecular evolution of living organisms. Insulin is present in almost all living organisms on the Earth, hence by comparing the insulin sequences found in different species and computing their similarity, one can get a very detailed insight into the evolution of life on Earth. Sequence comparison, similarity estimation and so on, is one of the most fundamental research topics in bioinformatics [6].

We define global similarity using a dynamic programming matrix.

Definition 4.2 Let Σ be a set, $x, y \in \Sigma^*$, $s : \Sigma \times \Sigma \to \mathbb{R}$ and $g, h : \mathbb{N}^+ \to \mathbb{R}^+$. Let $x, y \in \Sigma^*$ and let $m = |x|$ and $n = |y|$. The *Needleman-Wunsch* dynamic programming matrix denoted $\mathrm{NW}(x, y, s, g, h)$, is an $(m+1) \times (n+1)$ matrix S with rows and columns indexed from 0 such that $S_{0,0} = 0$, $S_{i,0} = \max_{1 \le k \le i} \{S_{i-k,0} - h(k)\}$, $S_{0,j} = \max_{1 \le k \le j} \{S_{0,j-k} - g(k)\}$ and for all $i = 1,2, \dots, m$ and $j = 1,2, \dots, n$

$$S_{i,j} = \max\left\{S_{i-1,j-1} + s(x_i, y_j),\ \max_{1\le k\le i}\{S_{i-k,j} - h(k)\},\ \max_{1\le k\le j}\{S_{i,j-k} - g(k)\}\right\}.$$

We define the *global similarity* between the sequences x and y (given s, g and h), denoted $S(x, y)$, to be the value $S_{m,n}$.

The global similarity $S(x, y)$ between sequences x and y defined above satisfies all conditions in the definition of similarity [4].

Theorem 4.3 *Let* Π *be a finite set of biological sequences and let* $S(x, y)$ *be the global similarity function given by the Needleman-Wunsch dynamic programming algorithm. If the associated distance p is upper and lower curvature bounded, then there exists a metric of Randers type whose distance function coincides with the weighted quasi-metric induced by* $S(x, y)$.

Proof Based on our Theorem 3.5 we can explain as follows. Let us consider a finite set Π of biological sequences like, for instance, the insulin sequences in all species one can find in the NCBI database. Clearly, this is a finite set of sequences, a finite set of data that can be considered as a compact set in Σ^*. By using the dynamic programming, we can endow this compact set with a symmetric similarity function $S(x, y)$. Theorem 3.5 implies that there is always a Randers type metric whose associated distance function coincides with the weighted quasi distance function obtained from the similarity function.

5. Conclusions

In this paper, we introduce a definition of the similarity function, quasi-metric space and weighted quasi-metric space. We also study the geometrical properties of a topological space endowed with a similarity. The relation with embeddings and bundle and Finsler metrics of Randers type has been explained. Moreover, we present the relation of the mathematical concepts with computer science and bioinformatics. In conclusion, there is always a Randers type metric whose associated distance function coincides with the weighted quasi-metric induced by a similarity function.

References

[1] Sabau, S.V., Shibuya, K. and Shimada, H., 2014. Metric structures associated to Finsler metrics. *Publications Mathematicae Debrecen*, 84(1-2), 89-103.

[2] Shanker, G. and Rani, S., 2018. Weighted quasi-metrics associated with Finsler metrics. [online] Available at : https://www.researchgate.net/publication/322567912_Weighted_quasi-metrics_associated_with_Finsler_metrics

[3] Vitolo, P., 1999. The representation of weighted quasi-metric spaces. *Rendiconti dell'Istituto di Matematica dell'Università di Trieste*, 31, 95-100.

[4] Stojmirovic, A., 2008. *Quasi-metrics, Similarities and Searches: Aspects of Geometry of Protein Datasets*. Ph.D. Victoria University of Wellington.

[5] Stojmirovic, A., 2004. Quasi-metric space with measure. *Topology Proceedings*, 28 (2), 655-671.

[6] Pevsner, J., 2017. *Bioinformatics and Functional Genomics*. 3rd ed. New Delhi: Wiley India.

[7] Vitolo, P., 1995. A representation theorem for quasi-metric spaces. *Topology and Its Applications*, 65, 101-104.

[8] Bao, D., Chern, S.S. and Shen, Z., 2000. *An Introduction to Riemann-Finsler Geometry*. New York: Springer-Verlag.
[9] Chern, S.S. and Shen, Z., 2005. *Riemann-Finsler Geometry*. 6th ed. Singapore: World Scientific.

Instructions for Authors

Current Applied Science and Technology journal contains research reports, articles concerning development work, reviews of the literature and research activities. The objectives are to publish and promote research contributions and innovative work in fields associated with applied science and technology. An electronic journal is provided on the website (https://www.tci-thaijo.org/index.php/cast/index). The editors reserve the right to require revision of the submitted manuscript as a condition for final acceptance.

The institute and the editorial board claim no responsibility for the contents or views expressed by the authors of individual articles. Copying is allowed provided that acknowledgement is made. All articles submitted for publication will be assessed by a group of distinguished.

Ethics:

The journal is committed to maintaining the high level of integrity in the content published and has a Conflict of Interest policy in place. The journal uses plagiarism detection software to screen the submissions. If plagiarism is found, the COPE guidelines on plagiarism will be followed. For more details, please see https://www.tci-thaijo.org/index.php/cast/navigationMenu/view/Publication_Ethics.

Page Charge: Free

Submission of Manuscripts:

Manuscripts must be written in English and submitted online. Manuscripts are to be reviewed (double blinded) by at least 3 referees specializing in relevant fields. Revised manuscripts have to be sent online.

All manuscripts should be submitted to: https://www.tci-thaijo.org/index.php/cast/index

Contact:

Editor of Current Applied Science and Technology
King Mongkut's Institute of Technology Ladkrabang
1 Soi Chalongkrung 1, Ladkrabang District,
Bangkok 10520, Thailand
Tel: 662-329-8136
Fax: 662-329-8221
Email: cast@kmitl.ac.th

Manuscript Preparation Guide:

General: Manuscripts must be typewritten using *Microsoft Word for Windows*, single-spaced with margin set-up (in page set up menu) as follows (see also the document template):

Top Margin	1.5"	Bottom Margin	1.5"
Left Margin	1.5"	Right Margin	1.5"

Good quality printouts using A4 paper size are required. Format should be a single column. Times New Roman font type is required. Font sizes for various text functions are as follows:

Text functions	Size *	Typeface
Title	14 (CT)	Bold
Author and co-authors	11 (CT)	Normal
Address for correspondence	11 (CT)	Normal
Abstract and main text	10 (LJ)	Normal
Section heading and number including "Abstract", "Acknowledgement", "References"	12 (CT)	Bold
Subsection heading and number	11 (LJ)	Bold

* CT = Center Text, LJ = Left Justified.

The corresponding author should be noted (included a Fax number and E-mail address) and indicated with an asterisk. Full postal addresses must be given for all co-authors, keyed to names (if required) by superscripted Arabic numbers.

Paper Length: Should not normally exceed 10 pages including figures and tables
Spelling: American English
Abstract: Should not exceed 250 words
Keywords: Should not exceed 8 keywords
Text: Authors are requested to use the following order when typing:-

All research reports (Full Length or Short Reports): Title, Authors, Affiliations, Abstract, Keywords, Introduction (in reserch papers this must be confined to relevant matters, and must not be a general review of cognate literature), Materials and Methods, Results and Discussion, Conclusions, Acknowledgements, References.

Reviews and Discussion Papers will be considered in any format appropriate to the purposes of the authors, although adherence to the general guidelines described above is encouraged.

Line Art Figures: Figures can be drawn using several packages such as Win Draw®, Auto CAD®, Corel Draw®, VISIO® etc.

Photographs: Actual size photographs are acceptable. However, they can also be put into a text stream using a good-resolution scanner. All photographs must be clear when printed in monochrome.

Graphs: Several packages available today can produce attractive and professional graph presentation. Some also provide curve-fitting function, which can be useful. However, two-dimensional bar charts are preferred. All graphs must be clear in monochrome printing.
Equations and complex expressions: Math CAD®, Math Writer® and Equation Editor® (included in Microsoft Word®) are acceptable for presentation of this type of material.

Citations: Citations in the text should be denoted by numbers between square brackets (i.e. [1, 2], [1-3], [1, 3-8]…) *in the order of first appearance in the text.*

References: References should be numbered to correspond with the text citations. References must be arranged as follows:

Books

Authors, Initials., Year. *Title of Book.* Edition (only include this if not the first edition). Place: Publisher.

Example:

[1] Barker, R., Kirk, J. and Munday, R.J., 1988. *Narrative Analysis*. 3rd ed. Bloomington: Indiana University Press.

Chapters of edited books

Chapter authors' surname(s), initials., Year of book. Title of chapter. In**:** Book editor(s) initials. surnames, ed. or eds. *Title of Book*. Place of publication: Publisher. Chapter number or first and last page numbers.

Example:

[2] Samson, C., 1970. Problems of information studies in history. In: S. Stone, ed. *Humanities Information Research*. Sheffield: CRUS, pp. 44-68.

E-books

Author, Year, *Title of Book*. [e-book] Place of publication: Publisher. Available through: include e-book source/database, web address or URL.

Example:

[3] Carlsen, J. and Charters, S., eds. 2007. *Global Wine Tourism*.[e-book] Wallingford: CABI Pub. Available through: Anglia Ruskin University Library website <www.libweb.anglia.ac.uk>.

Journal articles

Author, Initials., Year. Title of article. *Full Title of Journal*, Volume number (Issue number), Page numbers.

Example:

[4] Ross, A.B., Junyapoon, S., Jones, J.M., Williams, A. and Bartle, K.D., 2005. A study of different soots using pyrolysis–GC–MS and comparison with solvent extractable material. *Journal of Analytical and Applied Pyrolysis*, 74(1-2), 494-501.

In case of online journal articles without page number:

Author, Initials., Year. Title of article. *Full Title of Journal*, Volume number (Issue number), DOI-based link.

Example:

[5] Mondal, N.K. and Basu, S., 2019. Potentiality of waste human hair towards removal of chromium (VI) from solution: kinetic and equilibrium studies. *Applied Water Science*, 9(49), https://doi.org/10.1007/s13201-019-0929-5.

Proceedings

Author, Initials., Year. Title of article. *Full Title of Proceedings*, Place of Conference, Date, page.

Example:

[6] Thanaboripat, D., Ruangrattanametee, V. and Srikitkademwat, K., 2010. Control of growth and aflatoxin production of aflatoxin producing fungi in corn by salts. *Proceeding of the 8th International Symposium on Biocontrol and Biotechnology*, Pattaya, Thailand, October 4-6, 2010, 283-289.

Patent
Inventor name, Initial(s)., Assignee., Year. *Title*. Place. Patent number (status, if an application).
Example:
[7] Leonard, Y., Super Sports Limited., 2008. *Tin Can Manufacture and Method of Sealing*. Canada. Pat. 12,789, 675.

Dissertation
Author, Year of publication. *Title of Dissertation*. Level. Official name of University.
Example:
[8] Richmond, J., 2005. *Customer Expectations in the World of Electronic Banking: a Case Study of the Bank of Britain.* Ph.D. Anglia Ruskin University.

Websites
Authorship or Source, Year. *Title of Document*. [online] Available at: include web site address/URL (Uniform Resource Locator).
Example:
[9] NHS Evidence, 2003. *National Library of Guidelines*. [online] Available at: http://www.library.nhs.uk/guidelines

Acknowledgements: These should be as brief as possible.

Proofs:
Proofs will be sent to the corresponding author and *must* be returned as soon as possible. Corrections should be restricted to typesetting errors.

Copyright:
The author(s) transfer(s) the copyright of the article to Current Applied Science and Technology effective if and when the article is accepted for publication.

Page Numbering:
All pages must be sequentially numbered, preferably by using the automatic page numbering function on your computer.

Copyright Material:
It is the authors' responsibility to obtain written permission from the copyright holder (usually the book or journal publisher) to use copyright material, and to send a copy of this consent with the manuscript. This consent is not normally denied but it is an international legal requirement that it be obtained.

Note:

Please note that authors are urged to check their proofs carefully before returning, since the inclusion of late corrections cannot be guaranteed.

Author(s) are responsible for ensuring that the submitted manuscript fully meets the requirements specified in the above Instructions. Manuscripts which fail to do so will be returned unedited to the Author(s) for correction in accordance with the above requirements, before they can be submitted to the processes of Referee evaluation.

1.5 ”

TEMPLATE

Enter title here (14 PT type size, Title Case, Bold, Centered)

Author Information entered here:
Name (in full)

Affiliation
City
Country
(11 pt type size, upper and lower case, centered under the title)

How to Use This Document Template

Insert the information in your document in place of the text here. For the body of your document, use Times New Roman 10 pt. Font, upper and lower case, double-spaced. Allow an extra half space above a line containing superscripts and/or below a line containing subscripts. The whole text should be left-justified. The headings should be 12 pt size, uppercase, bold and centered.

1.5” 1.5”

Abstract (12pt)

Maximum 250 words here. (10 pt)
.....
.....
.....
.....
.....

Keywords: (10 pt)

Maximum of 8 words

1. Introduction (12 pt)

Clearly explain the nature of the problem, previous work, purpose, and contribution of the paper (10 pt).
.....
.....
.....
.....
.....

*Corresponding author: Tel.: ……..... Fax: …….
E-mail: ……

2. Materials and Methods (12 pt)

.....
.....
.....
.....

3. Results and Discussion (12 pt)

.....
.....

4. Conclusions (12 pt)

Clearly indicate advantages, limitations and possible applications (10 pt).
.....
.....

5. Acknowledgements (12 pt)

A brief acknowledgement section may be included here (10 pt).
.....
.....

References (12 pt)

References must be numbered in the order cited in the manuscript and indicated in the text by a number in square brackets (e.g. [1, 2]) (10 pt).

Example of References must be arranged as follows:

[1] Barker, R. Kirk, J. and Munday, R.J., 1988. *Narrative Analysis*. 3rd ed. Bloomington: Indiana University Press.
[2] Samson, C., 1970. Problems of information studies in history. In: S. Stone, ed. *Humanities Information Research*. Sheffield: CRUS, pp. 44-68.
[3] Carlsen, J. and Charters, S., 2007. *Global Wine Tourism*. [e-book] Wallingford: CABI Pub. Available through: Anglia Ruskin University Library website <www.libweb.anglia.ac.uk>.
[4] Ross, A.B., Junyapoon, S., Jones, J.M., Williams, A. and Bartle, K.D., 2005. A study of different soots using pyrolysis-GC-MS and comparison with solvent extractable material. *Journal of Analytical and Applied Pyrolysis*, 74(1-2), 494-501.
[5] Thanaboripat, D., Ruangrattanametee, V. and Srikitkademwat, K., 2010. Control of growth and aflatoxin production of aflatoxin producing fungi in corn by salts. *Proceeding of the 8th International Symposium on Biocontrol and Biotechnology*, Pattaya, Thailand, October 4-6, 2010, 283-289.

[6] Leonard, Y., Super Sports Limited., 2008. *Tin Can Manufacture and Method of Sealing*. Canada. Pat. 12,789, 675.
[7] Richmond, J., 2005. *Customer Expectations in the World of Electronic Banking: a Case Study of the Bank of Britain.* Ph.D. Anglia Ruskin University.
[8] NHS Evidence, 2003. *National Library of Guidelines*. [online] Available at: http://www.library.nhs.uk/guidelines

Note:

Tables and Graphs: Minimum of 10 pt type size, all captions should be upper and lower case, centered. Each table and figure must be on a separate page (or pages if required), ***and must be embedded in the text.***

Illustrations and Photographs: Halftones, minimum of 10 pt type size, bold, captions should be in upper and lower case, centered. Images must be computer-designed with clearly visibility.

Contact

Editor of Current Applied Science and Technology
King Mongkut's Institute of Technology Ladkrabang
1 Soi Chalongkrung 1, Ladkrabang District
Bangkok 10520, Thailand
Tel: 662-329-8136 Fax: 662-329-8221
E-mail: cast@kmitl.ac.th
Website: https://www.tci-thaijo.org/index.php/cast/index

KING MONGKUT'S INSTITUTE OF TECHNOLOGY LADKRABANG
1 Soi Chalongkrung 1, Ladkrabang District Bangkok 10520, Thailand
Tel: 662-329-8136 Fax: 662-329-8221
E-mail: cast@kmitl.ac.th
Website: https://www.tci-thaijo.org/index.php/cast/index