CHEMICAL ENGINEERING TRANSACTIONS
VOL. 78, 2020
A publication of AIDIC
The Italian Association of Chemical Engineering
Online at www.cetjournal.it
Guest Editors: Jeng Shiun Lim, Nor Alafiza Yunus, Jiří Jaromír Klemeš

ISBN 978-88-95608-76-1; **ISSN** 2283-9216
DOI: 10.3303/CET2078074

Moving Windows Prediction of Refined, Bleached and Deodorized Palm Oil Quality Using Multiple Least Squares and Partial Least Squares Regressions

Wai Hoong Khu[a], Nor Adhihah Rashid[a], Mohd. Aiman Mohd. Noor[a], Nur Atikah Mohd Rosely[a], Azmer Shamsuddin[c], Kamarul Asri Ibrahim[b], Mohd. Kamaruddin Abd. Hamid[b,*]

[a]Process Systems Engineering Centre (PROSPECT), Research Institute of Sustainable Environment (RISE), Universiti Teknologi Malaysia, 81310 UTM Johor Bahru, Johor, Malaysia
[b]School of Chemical and Energy Engineering, Faculty of Engineering, Universiti Teknologi Malaysia, 81310 UTM Johor Bahru, Johor, Malaysia
[c]Lahad Datu Edible Oils Sdn. Bhd., KM 2, Jalan Minyak off Jalan POIC, Locked Bag 16, 91109 Lahad Datu, Sabah, Malaysia
kamaruddin@cheme.utm.my

With the goal of shaping a smart, economically-sustainable palm oil refining industry, this paper aims to improve the prediction modelling of Refined, Bleached and Deodorized Palm Oil (RBDPO) quality for early quality fault detection and remediation via the novel use of a moving windows form of prediction in addition to the conventional static window form. In this study, both Multiple Least Squares Regression (MLSR) and Partial Least Squares Regression (PLSR) model training techniques and prediction computations are carried out in the form of two moving data windows types, namely the expanding and rolling fixed-size data windows, together with the conventional static window form of prediction as control. Prediction improvement is observed consistently across both regression techniques when carried out in moving windows form, where both types of moving windows predictions, i.e. expanding and rolling windows, have fared significantly better than the static window form, with an average prediction error reduction of 20.6 % for Free Fatty Acid prediction, 55.9 % for Moisture Content prediction, 32.6 % for Iodine Value prediction and 34.2 % for Colour Value prediction in RBDPO. Among the moving windows themselves, the superiority of expanding windows over rolling windows and vice versa is insignificant. On the whole, this study paves the way for a revamped RBDPO quality prediction form which better reflects the dynamic, transient nature of the palm oil refining process, thus further consolidating its reliability for widespread practical use.

1. Introduction

While palm oil has been a commodity favored in oleochemical research, concerns have been raised regarding on the imbalance in research strategies since the efforts are geared towards technical topics in favor of sustainability issues in palm oil production (Hansen et al., 2015). This study is thusly indirectly oriented towards the economical and quality sustainability aspect of palm oil refining, in which the predictive modelling of RBDPO quality as explored here is expected to provide quality forecasts which allow operators to monitor product quality in advance and to perform early remediations of crude oil feed. If substandard RBDPO quality could be predicted and prevented, it will reduce the need for cost-and-energy intensive recycling processes, which according to Makky and Soni (2014), can amount up to an hourly opportunity loss of MYR 159,000 due to excessive energy dedicated to pumps and process equipment during re-processing. With improved economic performance, the sustainable development of palm oil refinery plants can be further encouraged, thus positively reinforcing the path towards a smarter, sustainable palm oil refining industry (Ngan et al., 2018). To a certain extent, the implementation of RBDPO quality prediction also caters to the environmental sustainability aspect, where scraps and waste can be minimized as the production of off-spec products can be

Paper Received: 14/05/2019; Revised: 04/08/2019; Accepted: 28/08/2019
Please cite this article as: Khu W.H., Rashid N.A., Mohd. Noor M.A., Mohd Rosely N.A., Shamsuddin A., Ibrahim K.A., Abd. Hamid M.K., 2020, Moving Windows Prediction of Refined, Bleached and Deodorized Palm Oil Quality Using Multiple Least Squares and Partial Least Squares Regressions, Chemical Engineering Transactions, 78, 439-444 DOI:10.3303/CET2078074

predicted beforehand. Excessive spent bleaching earth and other refining reagents, which could severely contaminate landfills and are in dire need of treatment before being disposal (Lau et al., 2019), can be reduced in the first place with less redundant processing in refineries.

Palm oil predictive modelling was first initiated by Yusof et al. (2003) via the use of Artificial Neural Networks (ANN). Due to the black-box nature of ANN and its difficult parameter tuning process (Ma and Gomez, 2015), Noor et al. (2017) and Rashid et al. (2017) utilized multivariate regression algorithms, which are relatively more interpretable, for model training instead. A distantly-related study on Palm Oil Mill Effluent by Che Ithnin and Hashim (2019) also implemented the use of regression, specifically the Multiple Least Squares Regression (MLSR) technique for predictive modeling, whereas the aforementioned works of Noor et al. (2017) particularly utilized a technique in the form of the Partial Least Squares Regression (PLSR) for RBDPO quality prediction. While all predictive modelling studies in the palm oil industry up until this point have come up with various techniques on model training, they are carried out in a conventional single static window form, where a generalized model is generated from a training dataset and is applied on an entirely different testing dataset window in one go to produce a series of forecasted values. A major limitation of static window usage according to Black et al. (2014) is that it assumes the process is constant for both training and testing dataset windows and ignores any possible data distribution shifts over time. Moving windows prediction on the other hand, which is divided into either expanding and rolling windows prediction, incorporate the effects of changing data structure over time where prediction is carried out successively while the training set is repeatedly updated. While moving windows are used in stock returns and financial forecasting with visible improvement (39.5 % – 45.9 % error reduction as documented by case study of Black et al. (2014)) over static prediction, it is rarely found in the predictive modelling of manufacturing processes, let alone the palm oil industry, due to its elusiveness outside of the economics field and its relative tediousness in execution.

In whole, the key objective of this paper is to extend the RBDPO quality prediction framework established by Noor et al. (2017) and other authors by integrating the novel use of moving windows strategy in palm oil prediction with available MLSR and PLSR training techniques, and comparing its prediction accuracy with the conventional static window prediction form in terms of Mean Squared Errors of Prediction (MSEP). MSEP for both rolling and expanding windows will be further compared among themselves to determine the superiority of one form over the other, if present.

2. Methodology

2.1 Data collection and preparation

Large amounts of data were collected from the Lahad Datu Edible Oils refinery plant. The input variables for this prediction study, which are increased substantially from previous studies, consist of the Free Fatty Acid percentage (% FFA) of crude palm oil (CPO), the moisture content percentage (% MC) of CPO, the Iodine Value (IV) of CPO, the Deterioration of Bleachability Index (DOBI) of CPO, the phosphoric acid flowrate during degumming, the bleaching earth dosage during bleaching, and the vacuum pressure during deodorizing. The output variables were selected based on the key commercial quality parameters of RBDPO, which consist of the % FFA of RBDPO, the % MC of RBDPO, the IV of RBDPO, and the Colour Value (CV) of RBDPO. 500 observations were collected for each variable at a 30-minute interval. Prior to model training, the collected data were pre-processed beforehand by using statistical signal processing tools as proposed by Rosely et al. (2017) to fulfil the Central Limit Theorem of a large dataset. Firstly, data standardization was carried out to ensure a uniform data distribution for every variable, which was followed by box-plotting to select the subset with the highest normality. For the selected subset, the corresponding autocorrelation plots for CPO and RBDPO variables were plotted to identify the optimum sampling time where data points exist at high randomness for fixed intervals. Finally, cross-correlation plots for each CPO-RBDPO variable pair were generated to determine the processing time based on the optimum sampling time. All output observations were forward-shifted for a unit of processing time so that the current input observations were mapped onto future output observations to allow the prediction of output RBDPO qualities in advance.

2.2 Model training using Multiple Least Squares and Partial Least Squares Regression techniques

The pruned dataset was divided into a training and a testing dataset, in which the former was used to generate a generalized input-output mathematical model $Y=X\beta$, where Y is a m-by-p output training subset matrix with p variables and m observations, X is a m-by-n input training subset matrix with n variables and m observations, and β is a n-by-p predictor coefficient matrix. To generate the model i.e. model training, β was computed differently using MLSR or PLSR techniques. For the MLSR technique, Eq(1) was used.

$$\beta = (X^T X)^{-1}(X^T Y) \quad (1)$$

For the PLSR technique, the Nonlinear Iterative Partial Least Squares (NIPALS) algorithm was used for β computation where X is decomposed into score matrix T and loading matrix P, and Y is decomposed into score matrix U and loading matrix Q in an iterative fashion while maximizing X-Y covariance, W, at the same time. The detailed algorithm is explained thoroughly in the works of Noor et al. (2017).

2.3 Model training and prediction via static, expanding and rolling windows forms

The prediction process was carried out after model training by multiplying the predictor coefficient β to the input subset X_{TS} of the testing dataset to generate the predicted output RBDPO data $\hat{Y}$. Model training and prediction were carried out in three main forms, where (b) and (c) are novel implementations.

a. Static window training and prediction
   During model training, regression was conducted on the entire training input subset window $\boldsymbol{X}$ and training output subset window Y to generate a single β matrix. After that, β was multiplied to a separate testing input subset window X_{TS} to generate a series of predicted values $\hat{Y}$ in one go, assuming the mutually exclusive training and testing datasets were having the same distribution.
b. Expanding windows training and prediction
   Prior to training and prediction, the training and testing input subsets (X and X_{TS}) were concatenated into a single joint input set X_j, whereas the training and testing output subsets (Y and Y_{TS}) were concatenated into a joint output set Y_j. The first m observations in X_j and Y_j were selected as the initiating windows X_{w1} and Y_{w1}, where the window size m is the original number of observations (number of rows) in the training input subset X. The first β, i.e. β_1, was calculated for the initiating windows X_{w1} and Y_{w1}, via either regression methods. β_1 was subsequently multiplied with the $(m+1)^{th}$ observation in X_j, i.e. x_{m+1}, to obtain the first forecasted output observation, $\hat{y}_1$. A new pair of input and output data windows X_{w2} and Y_{w2} were then assigned by adding the next X_j and Y_j observation into the original window. A new β_2 was computed by performing regression on the new windows X_{w2} and Y_{w2}, which was in turn multiplied with the $(m+2)^{th}$ observation in X_j, x_{m+2}, to give the next predicted observation, $\hat{y}_2$. The window assignment, training and prediction processes were repeated as the data windows got one size larger for each update. The predicted observations from every window $\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n$ were collected into the predicted output set $\hat{Y}$.
c. Rolling windows training and prediction
   The training and prediction process of the rolling windows form was the same as that of the expanding windows up until the assignment of the new window pairs X_{w2} and Y_{w2} and subsequent new window pairs. Now instead of just adding the next X_j and Y_j observation to the original input and output windows, an added step was required where the first observation in both the original input and output windows were simultaneously removed. Thus, the window size remained constant for each update as the data input and output data windows X_w and Y_w "roll on" instead of "expanding" to the next.

For both moving windows cases, there were (M-m) consecutive windows altogether with (M-m) predicted output observations $\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_{M-m}$, where M is the total row number of the joint input set X_j, and m is the initiating window size. Overall, to test the prediction accuracy, the MSEP metric was used (in agreement to previous predictive modelling studies), as shown in Eq(2).

$$MSEP = (\sum_{i=1}^{N} y_{TS,i} - \hat{y}_i)/N \quad (2)$$

Where N is the number of observations in the output testing dataset Y_{TS}, $y_{TS,i}$ is the i^{th} observation in Y_{TS}, and $\hat{y}_i$ is the i^{th} observation in $\hat{Y}$. Control chart plotting was optionally used as a mean to monitor the predicted palm oil quality as time progressed.

3. Results and discussion

The data pre-processing stage based on the methodology proposed by Rosely et al. (2017) yielded a pruned dataset with an optimal data range with high normality boxplots of 25 data points for training and subsequent 25 data points for testing. Among the autocorrelational plots of the selected dataset, the plots for the CPO % FFA showed 8 lags to fall beneath the confidence bands, which signified an optimum sampling time of 8 lags*original sampling time (0.5h) = 4 hours for high data randomness and minimum heredity effects. This was followed by cross-correlational plotting, where the plots of CPO-RBDPO % FFA variable pair once again showed 4 lags to cross the zero boundary, and the statistical processing time was thusly determined as 4 lags*optimum sampling time (4h) = 16 hours. This allowed the output quality of RBDPO in terms of % FFA, % MC, IV and CV to be predicted at 16 hours ahead of time, before CPO with a current % FFA, % MC, IV and DOBI was processed.

Subsequently, model training and prediction were carried out with a combination of either one of two model training techniques (MLSR and PLSR) and either one of three training and prediction forms (static window,

expanding windows and rolling windows). Altogether there were six training and prediction combinations for each of the four output RBDPO variables. The MSEP of each combination for each output variable (RBDPO % FFA, % MC, IV and CV) are depicted visually on bar charts in Figures 1a, 1b, 1c, and 1d.

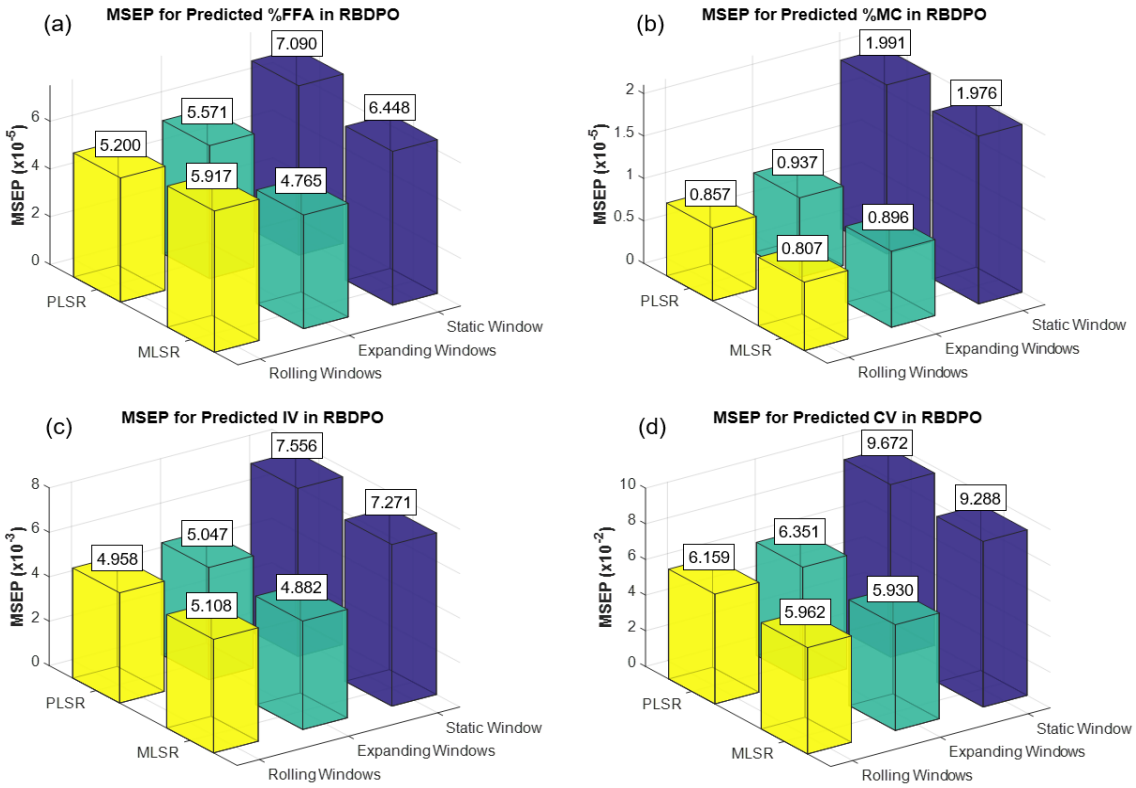

Figure 1: MSEP values for predicted (a) % FFA; (b) % MC; (c) IV; (d) CV in RBDPO

As a minor observation, the usage of either MLSR and PLSR gives visually similar results. The latter technique gives a slight error increase of 4.92 %, 3.83 %, 1.46 % and 4.85 % on average for % FFA, % MC, IV and CV predictions, which could be attributed to a slight over-removal of data in discarded components despite multicollinearity treatment during PLSR training, as hypothesized by Carley et al. (2004). Nevertheless, in this study, the effects of using either regression techniques are not nearly as pronounced and evident as the effects of selecting different types of windows for model training and prediction on the prediction accuracy. The novel usage of both dynamically expanding and rolling windows has led to a substantial decrease in MSEP compared to the conventional usage of a single static prediction window for all RBDPO output variables, as shown in Figure 1. The percentage of improvement of both moving windows prediction forms, measured in terms of percentage reduction in MSEP when compared to that of static window prediction, is recorded in Table 1 for each output variable.

Table 1: Percentage improvement of moving windows prediction over static window prediction

	% FFA	% MC	IV	CV
% Improvement of expanding windows prediction	23.76	53.80	33.03	35.24
% Improvement of rolling windows prediction	17.44	58.05	32.07	36.07
Average % improvement of moving windows prediction	20.60	55.92	32.55	35.66

As shown in Table 1, the improvement of selecting either form of moving windows prediction over static window is significant, with an average improvement of 20.60 % for % FFA prediction, 55.92 % for % MC prediction, 32.55 % for IV prediction, and 35.66 % for CV prediction. When moving windows are implemented, prediction error reduction is observed consistently across different types of regression techniques used, thus implying that the usage of moving windows objectively leads to a better prediction regardless of the model training techniques, though it should be noted that the study is only limited to two notable techniques, i.e. MLSR and PLSR. In comparison to the previous PLSR prediction study by Noor et al. (2017), which gave an

MSEP of 1.24×10^{-4} for % FFA prediction, 1.81×10^{-5} for % MC prediction, and 8.53×10^{-3} for IV prediction (CV was not considered in previous studies), this study has also shown a prominent improvement of 56.5 %, 50.4 % and 41.3 % on average for % FFA, % MC and IV. Similar to the outcomes in this study itself, the improvement of this study over the previous works by Noor et al. (2017) could be attributed to the utilization of moving windows prediction in favor of the conventional static window form, in addition to the usage of an increased amount of input variables for model training in this framework. Overall, the substantial improvement resulted from the implementation of moving windows is aligned with what has been observed in the economics field as well, with an error reduction of 39.5 % – 45.9 % in rolling window forecasts of stock returns compared to standard static window forecasts as recorded in the case study of Black et al. (2019). In this situation, when comparing among both moving windows, i.e. expanding and rolling windows themselves, the MSEP differences between both conditions are relatively insignificant, with expanding windows being superior for % FFA and IV prediction, and the vice versa for % MC and CV prediction, suggesting that the expanding windows form of prediction seemed to be able to capture the data drift as well as the rolling windows form did due to a small sample.

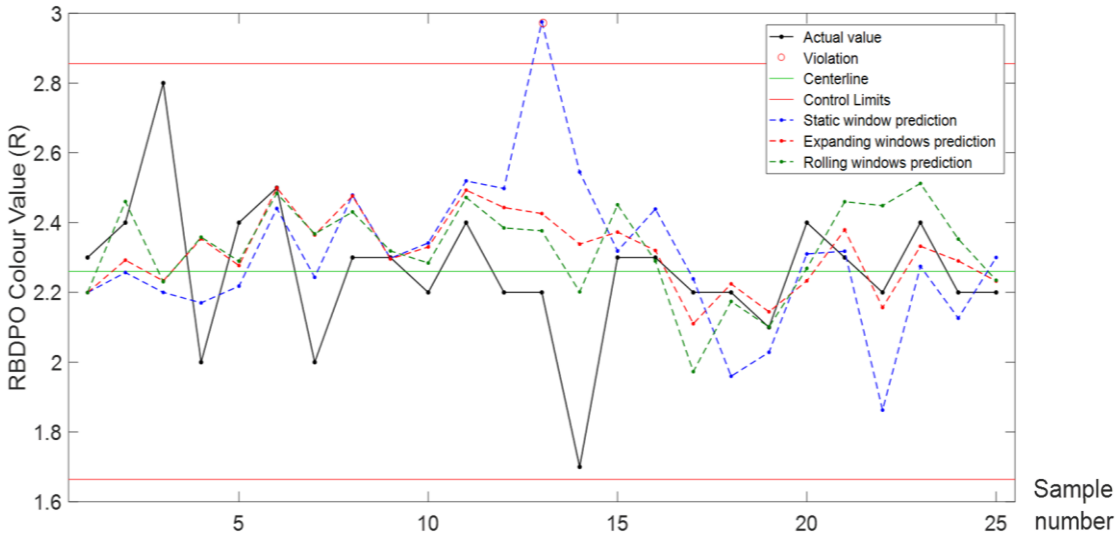

Figure 2: RBDPO Colour Value control chart with actual data, static, expanding and rolling window predictions

When the actual and predicted data for all window forms are plotted on a quality control chart, as shown in an example of a Shewhart chart of the CV parameter of RBDPO in Figure 2, it is evident that the predicted values from both expanding and rolling windows showed a significantly higher visual adherence than that of static window towards the actual testing data values on the chart. The static window form has also produced a false alarm, i.e. an out-of-control prediction, on the 13th observation whereas both similar predicted trends from moving windows are correctly lying within the control limits. When compared among themselves, the superiority of either rolling or expanding windows is insignificant, where each shows a better prediction accuracy over the other at both halves of the process separately.

Admittedly, there are shortcomings while implementing the moving windows strategy in this study. The downside of using dynamic moving windows compared to static window, is the relative tediousness of their execution (there were 50-25 = 25 windows in this case), but the huge increase in prediction reliability of 36.2 % on average warrants its slightly painstaking implementation. Another notable limitation in the study is the small sample size and prediction period, which should be extended to conclusively determine the disparity between expanding windows and rolling windows prediction in the long run. Meanwhile to improve model training, more process variables from the degumming, bleaching and deodorizing stages should be considered, though variable selection strategies should be introduced to remove excess irrelevant features from the mix. Nonetheless, the implication of this study in spite of its limitations is apparent as it paves the path towards a significantly improved palm oil quality prediction framework by integrating the usage of moving windows outside of their usual applications in financial forecasting. In a practical sense, the moving windows strategies in palm oil prediction enable the forecasts to more accurately match the data drifts in a transient refining process (as observed similarly in stock returns forecasting), thus resulting in an increased reliability in prediction which better reflects the production quality in the long run. This in turn allows the operators to perform correct remedial decisions more confidently in the event of subpar predicted qualities, while

consequently being less reliant on energy-and-resource-intensive recycling and redundant processing efforts. With sufficient data collection and input-output variable identification, the moving windows prediction framework as enumerated in this paper can even be reiterated and applied in other time series predictive modelling studies not just in palm oil refining, but in other manufacturing industries to reap its benefits as well.

4. Conclusions and recommendations

The predictive modelling of RBDPO refining process has been successfully performed by using simple MLSR and PLSR with an integrated use of moving windows strategies for model training and prediction. The novel usage of rolling and expanding windows significantly improves prediction accuracy over conventional static window predictions with a substantial average error reduction of 36.2 %, and the moving windows prediction improvement is observed objectively in spite of different regression techniques used in this case. As such, the moving windows strategies are greatly recommended for future predictive modelling studies despite their tediousness in execution. For such future studies, a larger sample size and longer prediction periods are recommended to further compare the superiority of rolling windows over expanding windows or vice versa in the long run, whereas the number of modelling variables could be increased with subsequent introduction of appropriate variable selection strategies to improve model training. A recommended addition to the predictive modelling framework would be an extra fault diagnosis step, where problematic input variables could be isolated in the event of predicted out-of-controls and appropriately targeted and remedied. In a greater sense, through this line of research, the reliability and practicality of prediction in palm oil refining industries could be greatly improved along the way, thus laying the groundwork that warrants a confident implementation of predictive modelling in palm oil refineries for economical and quality sustainability on a wider scale.

Acknowledgments

The financial support from Research University Grant (RUG) Tier 1 (Q.J130000.2546.12H67) and the School of Chemical and Energy Engineering, Universiti Teknologi Malaysia (UTM) is greatly acknowledged.

References

Black A.J., Klinkowska O., McMillan D.G., McMillan F.J., 2014, Forecasting stock returns: do commodity prices help, Journal of Forecasting, 33(8), 627-639.

Carley K.M., Kamneva N.Y., Reminga J., 2004, Response surface methodology: CASOS technical report. Carnegie Mellon University, Pennsylvania, United States of America.

Che Ithnin N.H., Hashim H., 2019, Predictive modelling for biogas generation from palm oil mill effluent (POME), Chemical Engineering Transactions, 72, 313-318.

Hansen S.B., Padfield R., Syayuti K., Evers S., Zakariah Z., Mastura S., 2015, Trends in global palm oil sustainability research, Journal of Cleaner Production, 100, 140-149.

Ma Y.Z., Gomez E., 2015, Uses and abuses in applying neural networks for predictions in hydrocarbon resource evaluation, Journal of Petroleum Science and Engineering, 133, 66-75.

Makky M., Soni P., 2014, In-situ quality assessment of intact oil palm fresh fruit bunches using rapid portable non-contact and non-destructive approach, Journal of Food Engineering, 120, 248-259.

Ngan S.L., Promentilla M.A.B., Yatim P., Lam H.L., Er A.C., 2018, Developing sustainability index for Malaysian palm oil industry with fuzzy analytic network process, Chemical Engineering Transactions, 70, 229-234.

Noor M.A.M., Rosely N.A.M., Rashid N.A., Moh Y.H., Shamsuddin A., Hamid M.K.A., Ibrahim K.A., 2017, Quality prediction of refined bleached deodorized palm oil (RBDPO) using partial least squares regression technique, Energy Procedia, 142, 3002-3007.

Lau S.Y., Phuan S.L., Danquah M.K., Acquah C., 2019, Sustainable palm oil refining using pelletized and surface-modified oil palm boiler ash (OPBA) biosorbent, Journal of Cleaner Production, 230, 527-535.

Rashid N.A., Rosely N.A.M., Noor M.A.M., Shamsuddin A., Hamid M.K.A., Ibrahim K.A., 2017, Forecasting of refined oil quality using principal component regression, Energy Procedia, 142, 2977-2982.

Rosely N.A.M., Rashid N.A., Noor M.A.M., Hawi N.D.A, Sepuan S.Q., Shamsuddin A., Ibrahim K.A., Hamid M.K.A., 2017, Product sampling time and process residence time prediction of palm oil refining process, Chemical Engineering Transactions, 56, 1411-1416.

Yusof K.M., Idris A., Lim J.S., Wong H.M., Morad N.A., 2003, Artificial neural network modelling of steady state chemical engineering system. In Malaysia-Japan Seminar on Artificial Intelligence Applications in Industry, June 24-25, Kuala Lumpur, Malaysia.