

Quality Prediction of Refined, Bleached and Deodorised Palm Oil using Multiple Least Squares Regression

Ling Qian Tan^a, Nor Alafiza Yunus^{a,*}, Wai Hoong Khu^a, Mohd Kamaruddin Abd Hamid^b

^a School of Chemical and Energy Engineering, Faculty of Engineering, Universiti Teknologi Malaysia, 81310 UTM Johor Bahru, Johor Darul Takzim, Malaysia.

^b JMH Integrated Services Sdn. Bhd., 34, Jalan PI 10/3, Taman Pulai Indah, Bandar Baru Kangkar Pulai, 81300 Johor Bahru, Johor Darul Takzim, Malaysia.
alafiza@utm.my

The predictive modelling for refined, bleached, and deodorised palm oil (RBDPO) quality was performed with an aim to shape a smart and economically sustainable palm oil refining industry. The aim of this study is to improve the RBDPO predictive modelling framework via the novel use of moving average and moving windows form of prediction. Weighted Moving Average (WMA) data smoothing technique was introduced to investigate the effect of the moving average algorithm by comparing the outcomes with the unsmoothed dataset. The Multiple Least-Squares Regression (MLSR) technique was used for model training and prediction of the RBDPO quality. The models were performed using dynamic windows in an expanding and rolling windows form as well as the conventional static window form of prediction. The prediction improvement was consistently observed in all the methods used and was compared in terms of the Mean Squared Error of Prediction (MSEP). It was revealed that the data smoothing result with WMA reduced 33.73 % of the error compared to the unsmoothed data. By comparing the prediction accuracy of the order of the model, the process fitted the first-order model and gave 16 times error less than the second-order model. The dynamic windows prediction form reflected better precision with an average of 37 % error reduced dynamic form that fulfilled the transient nature of the palm oil refining process, in which the data distribution shifts over time.

1. Introduction

The palm oil industry's development has grown exponentially due to increasing palm oil demand worldwide. MPOB (2017) reported that Malaysia is one of the world's leading palm oil-producing countries, with 19.92 Mt of palm oil production in 2017. The demand for refined, bleached, deodorised palm oil (RBDPO) has increased linearly to meet the needs of RBDPO in the market globally. RBDPO is widely used in manufacturing biofuels, agri-food and body care sectors (Ayompe et al., 2020). The RBDPO quality must always meet the requirements for safe consumption. In the current palm oil industry, the quality control of RBDPO is monitored via chemical analysis in the laboratory. The off-specification products are sent back to the refinery tank to be refined again, which is costly and time-consuming (Sulaiman et al., 2018). By implementing predictive modelling, early remediations can be performed to avoid off-specification products. This approach minimises waste materials from excessive reagent spent, excessive bleaching earth dosage and reduces environmental pollution caused by waste disposal to the environment (Lau et al., 2019).

One of the most common and widely applied predictive tools to predict RBDPO quality in past studies and various fields is regression and Artificial Neural Networks (ANN). Although ANN can accurately fit in the non-linear variables, the black-box nature of ANN does not make it a viable option since analysts might not detect the source of error and it is difficult to tune parameters. The relationship between the variables cannot be seen clearly using ANN. The industry requires a more interpretable method for model training. In response to this issue, the Multiple Least-Squares Regression (MLSR) model training was implemented for RBDPO quality prediction with the aid of MATLAB (MATLAB, 2019). Che Ithnin and Hashim (2019) also applied MLSR method for the prediction of biogas generation from palm oil mill effluent. The refinery plant data contains noise and will

show low accuracy in prediction if the data is not clean. It is worth noting that in standard predictions, fluctuations often conceal the trend.

In the data cleaning approach, the moving average produces a smaller noise effect and provides a clearer trend for RBDPO quality prediction. By observing this trend, fault can be detected early and reduces time consumption in the process. A smart Weighted Moving Average (WMA) smoothing algorithm was proposed to contribute to an effective data cleaning approach in the industry. Previous study by Rashid et al. (2018) carried out the process in a conventional static window form and assumed the process is static all the time, disregarding any possible data distribution shifts over time. The process in a real plant has a transient behaviour. The data structure changes periodically makes the static window form is less reliable and resulting in low prediction accuracy. A more dynamic form that is better suited to the transient process is required in the industry.

The aim of this study is to develop an enhanced predictive modelling framework for RBDPO quality prediction using MATLAB. WMA algorithm and the effect of higher-order predictive modelling were introduced and investigated towards RBDPO prediction accuracy. Various model training techniques were simulated in a conventional static window form and dynamic window form of prediction.

2. Methodology

2.1 Data collection and preparation

The Sabah palm oil industry is a palm oil refinery plant that processes crude palm oil (CPO) into RBDPO, a widely used ingredient in the food industry. More than 300 sets of data were collected from the Sabah palm oil industry. In this study, there were a total of 21 input variables, including temperature and pressure of the processes, tanks level, pressure during deodorisation, percentage of free fatty acid (FFA), moisture value (MOIST) and iodine value (IV) of CPO. The output variables include important quality parameters of RBDPO, namely FFA, MOIST, IV and colour value (CV). The observations were collected for each variable every hour. After the data were collected, they were analysed and pre-processed using MATLAB to remove the data that run out of range and produces misleading results. Standardisation of data was carried out using mean-centred and standardised data for each column (Wan et al., 2014). As a result, there was no y-intercept and the generalised mathematical input-output model is computed as shown in Eq(1) (Khu et al., 2020).

$$Y = X\beta \quad (1)$$

The above model is bivariate with a dependent variable, Y , an independent variable, X and β as the predictor coefficient. The standardised data were then normalised for every variable using the 1.5 Interquartile Range (IQR) rule to remove the outliers. After data sorting, the processing time was determined using cross-correlation for each independent variable, X to Y . The processing time was the mode of the cross-correlation plots with a cross-correlation coefficient closest to 1 or -1. Next, the output observations were forward-shifted so that the input observations were mapped to output observations.

2.2 Data pre-processing

The data were divided into two sets, one set was without smoothing and the other set was carried out with WMA. This distinction was to test whether smoothing will give a better prediction result. Dataset without smoothing was the actual dataset, whereas the smoothed dataset had been assigned a weight to the most current data. All X variable observations were processed using Eq(2) and all x_{WMA} values were recorded in an m -by- n input training matrix (Riyadi et al., 2019).

$$x_{WMA} = \sum a_n x_i / \sum a_n \quad (2)$$

whereby a_n is the weight of data and x_i is the actual data of variable X . In this study, the second-order modelling was added to determine the non-linearity of RBDPO quality prediction. The WMA data and the non-smoothed dataset were further developed into the first-order and second-order datasets. The first-order dataset is the original smoothed and non-smoothed data, while the second-order data was obtained by squaring each observation. The second-order model is shown in Eq(3) (Rahman Khan et al., 2012).

$$Y = \sum_{i=1}^n \beta_{ii} X_i^2 + \sum_{i=1, i \neq j}^n \beta_{ij} X_i X_j + \sum_{i=1}^n \beta_i X_i \quad (3)$$

Y is the output variable, X_i is the input variables (first-order terms), X_i^2 and $X_i X_j$ are the second-order terms of the input variables, while β_{ii} , β_{ij} and β_i are the unknown predictor coefficients. The input variable, X matrix size, was enlarged to an m -by- q matrix due to the second-order terms in the model. The output variable, Y matrix size, remained.

2.3 Model training and prediction

The dataset was divided into a training and testing set with a ratio of 3:1. It is to make sure the training dataset is large enough to carry out the prediction. The training set is used for the regression, while the testing set is used for the prediction. The training input and output subsets were denoted as X_{TR} and Y_{TR} while the testing input and output subsets were denoted as X_{TS} and Y_{TS} . Model training and prediction were performed in static, expanding and rolling window forms using MATLAB software. The prediction analysis was carried out after model training by multiplying the predictor coefficient, β , with X_{TS} input subset to obtain the predicted output RBDPO data, \hat{y} . The MLSR technique was applied to determine β by using Eq(4).

$$\beta = X^T X / X^T Y \quad (4)$$

X is an m -by- n input training subset matrix with m observations and n variables, X^T is an n -by- m transpose matrix of X , Y is an m -by- p output training subset matrix with m observations and p variables, Y^T is a p -by- n transpose matrix of Y and β is an n -by- p predictor coefficient matrix. Each combination: the first and second orders without smoothing, the first and second orders smoothed with WMA had undergone model training and prediction. The methodology was followed by the three types of window prediction. The detailed algorithm was explained thoroughly in Khu et al. (2020). The schematic diagram for each form was illustrated in Figure 1.

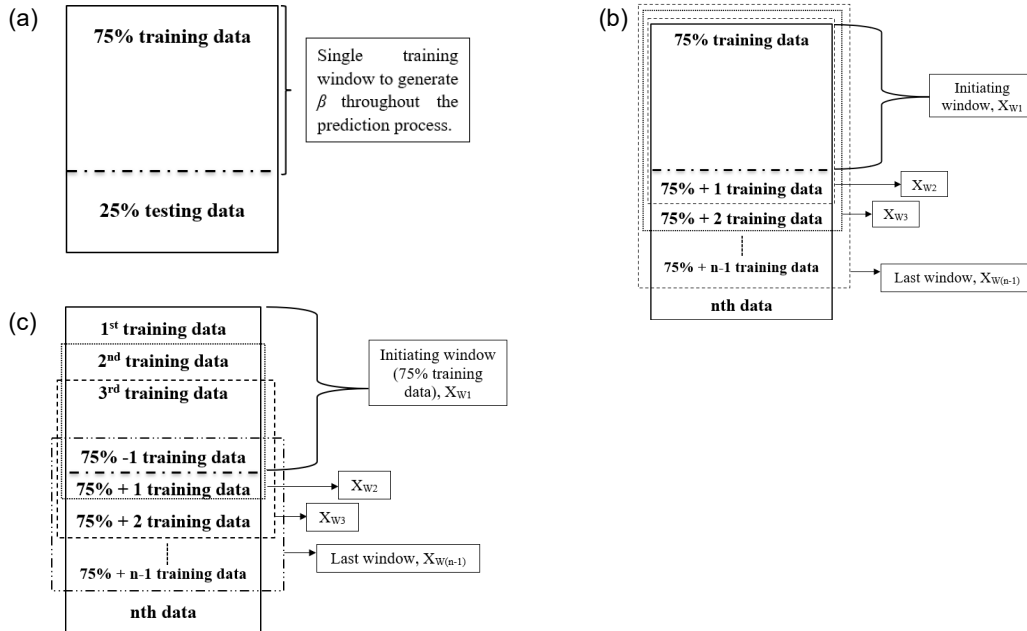


Figure 1: Schematic diagrams for (a) static; (b) expanding; (c) rolling windows of training and prediction

After model training and prediction, the Mean Squared Error of Prediction (MSEP) for predicted FFA, MOIST, IV and CV of RBDPO were calculated for every combination, followed by plotting the MSEP for each output variable of RBDPO using Eq(5).

$$\text{MSEP} = \frac{\sum_{i=1}^n (y_{TS,i} - \hat{y})^2}{N} \quad (5)$$

$y_{TS,i}$ is testing y -value at the i^{th} row, \hat{y} is the predicted value and N is the total number of rows. The MSEP analysis compares the prediction error of each combination.

3. Results and discussion

Data pre-processing, including normality and outlier tests, were conducted on the dataset using the 1.5 IQR rule via MATLAB. The results display that there were 112 observations left after outlier removal. From the pruned dataset, the 1st to 20th observations were analysed using a cross-correlation plot to determine the residence time. The optimum process lag was determined to be 1 and by multiplying 1 lag with one-hour time intervals in the process, the optimum processing time observed was one hour. The data was forward shifted to map output onto the input before performing model training and prediction.

3.1 Effect of moving average algorithm

The MSEP was calculated between the actual output data with a predicted value to investigate the prediction model efficiency. Figure 2 shows the bar chart of the MSEP for the predicted IV of six combinations: the first-order without smoothed dataset and the first-order smoothed with WMA dataset for static, rolling and expanding model. The MSEP closest to zero is regarded as the most accurate. Based on Figure 2, the MSEP values for the moving average data were relatively smaller than the data without moving average. The former set produced an error reduction of 33.73 % on average for IV prediction. Applying WMA before model training and prediction helped smooth the noisy data by filtering noise from dynamically fluctuated data (Hota et al., 2017). As a result, the dataset with WMA had a smaller MSEP.

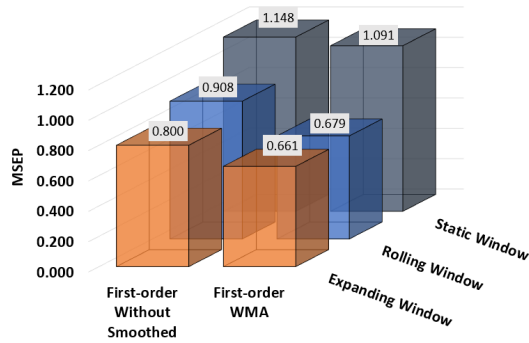


Figure 2: MSEP for predicted IV of RBDPO

3.2 Effect of a higher-order predictive modelling

Figure 3 illustrates the MSEP of FFA in RBDPO towards the first-order and second-order prediction results for static, window and rolling window prediction. The second-order MSEP values were significantly larger than the first-order MSEP values, resulting in an average of 16 times error increment compared to the first-order error. The RBDPO quality prediction did not fit well with the second-order, resulting in notable MSEP values. This result could be due to the interaction variable effect. The independent variables are multiplied with another independent variable and have a more complex model as the number of columns for input increases after multiplying with all the independent variables. The response was well modelled by a linear function, indicating that the RBDPO predictive modelling reflects the first-order behaviour more than a non-linear behaviour.

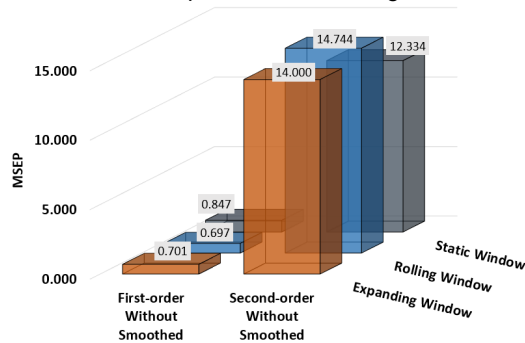


Figure 3: MSEP for predicted FFA of RBDPO

3.3 Effect of static and moving window form of model training and prediction

Table 1 shows the percentage improvement of rolling and expanding windows prediction for RBDPO FFA, MOIST, IV and CV over static window prediction for the first-order without the moving average dataset.

Table 1: Improvement (%) of rolling and expanding windows prediction over static window prediction for the first-order without the moving average dataset

	FFA	MOIST	IV	CV
Improvement of expanding window prediction (%)	20.80	0.00	43.58	58.14
Improvement of rolling window prediction (%)	21.51	0.00	26.45	39.70
Average improvement (%)	21.16	0.00	35.02	48.92

Based on the results in Table 1, the moving windows prediction had significantly increased the efficiency for model training and prediction. The percentage improvement for the MOIST is 0 % because the MOIST values remain the same throughout the column after normalisation and gave same predicted values for the three types of windows prediction. Table 1 implies that moving windows had increased the prediction accuracy for all RBDPO variables. Khu et al. (2020) achieved an average improvement of RBDPO dynamic windows prediction at 20.60 % for FFA, 55.92 % for MOIST, 32.55 % for IV and 35.66 % for CV. The efficiency of the dynamic window was higher because the process is transient, whereby the process variables changes with time. The moving windows training sets rolled and expanded for each iteration, the most recent data was included in the prediction and produced a more accurate approximation while the conventional static window utilised the same window throughout the process.

3.4 Improved predictive modelling framework for refined, bleached, deodorised palm oil

Figures 4a and 4b represent the control chart of RBDPO FFA and IV with seven trends: the actual data, the first-order data and the first-order with WMA data over static, rolling and expanding window form of prediction. Based on Figure 4a, the predicted data had a similar trend for the six combinations. It was observed that the dataset with a moving average produced more precise results at the 4th and 15th sample numbers. The precision of the outcomes was affected by the fluctuations of process parameters and disturbances of the process. The results were acceptable as all the predicted data were within the control limits.

The trend shown in Figure 4b is more precise and relatively more comparable. The first-order trend lines with WMA were closer to the actual data than the first-order dataset without a moving average. The 8th, 13th, 14th, 25th and 28th observations of the original first-order dataset showed a significant downward trend as opposed to the actual trend. The 28th observation of the static window first-order without a moving average set falls out of the control limit. This causes a false alarm in the real industry, elucidating the suitability of the WMA technique. By comparing static and moving window forms of the first-order with WMA data, the static window trend line indicated an upward trend with visually different trends from the actual data. It demonstrated that the trend of dynamic windows was closer to the actual data trend.

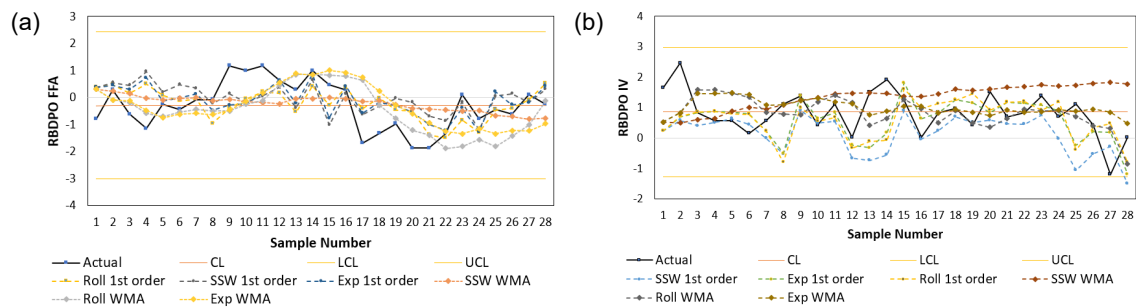


Figure 4: Control chart for (a) the FFA of RBDPO; (b) the IV of RBDPO

Based on the results, the improved predictive modelling framework for RBDPO quality is proposed, as shown in Figure 5.

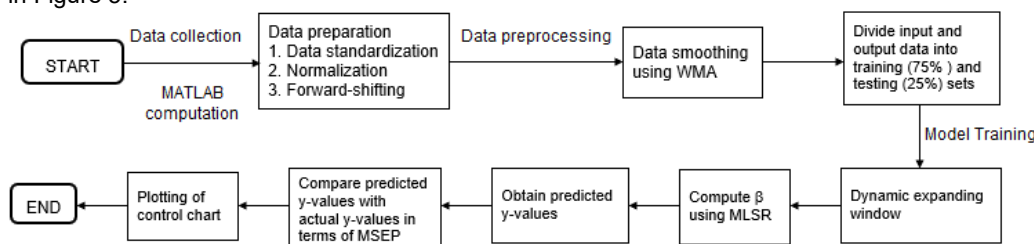


Figure 5: Results of the improved predictive modelling framework for RBDPO quality

The pre-processed data was divided into input and output matrices to be further processed for data smoothing. The MATLAB simulation results proved that the moving average with WMA produced a smaller MSEP and a clearer data trend than the dataset without smoothing. The second-order modelling increased significantly in its MSEP, indicating that the CPO-RBDPO quality prediction did not reflect a non-linear behaviour. The first-order modelling should be chosen since the model fits a linear response. By comparing static, expanding and rolling

windows, the dynamic window trend was closer to the actual data trend. The effect of two moving window forms could be compared in terms of percentage improvement. The expanding window form had an overall greater improvement (30.63 %) than the rolling window (21.92 %). Expanding window form was selected as a better method to improve the framework for RBDPO quality prediction.

4. Conclusion

An improved framework for the RBDPO prediction modelling has been developed successfully. The moving average smoothing algorithm was very useful in determining the RBDPO prediction accuracy, which minimised an average error of 33.73 %. Higher-order non-linear algorithm did not fit well with all the RBDPO variables as the response was well modelled by a linear model, which resulted in 16 times error reduction compared to the higher-order model. Most significantly, the rolling and expanding windows had notably raised the prediction accuracy compared to the conventional static window form of prediction. The rolling window form had also reduced 33.15 % of error while the expanding window improved 40.84 % of prediction accuracy compared to the static window prediction results. Further validation is suggested in future studies to ensure the reliability and practicality of prediction in the palm oil refining industries.

Acknowledgements

The authors would like to thank the Ministry of Higher Education Malaysia and Universiti Teknologi Malaysia (UTM) for providing the research fund for this study (Vote No. R.J130000.7351.4B504).

References

- Ayompe L.M., Schaafsma M., Egoh B.N., 2020, Towards sustainable palm oil production: The positive and negative impacts on ecosystem services and human wellbeing, *Journal of Cleaner Production*, 278, 123914.
- Che Ithnin N.H., Hashim H., 2019, Predictive modelling for biogas generation from palm oil mill effluent (POME), *Chemical Engineering Transactions*, 72, 313-318.
- Hota H.S., Handa R., Shrivastava A.K., 2017, Time series data prediction using sliding window based RBF neural network, *International Journal of Computational Intelligence Research*, 3(5), 1145-1156.
- Khu W.H., Rashid N.A., Mohd Noor M.A., Mohd Rosely N.A., Shamsuddin A., Ibrahim K.A., Abd. Hamid M.K., 2018, Moving windows prediction of refined, bleached and deodorised palm oil quality using multiple least squares and partial least squares regressions, *Chemical Engineering Transactions*, 78, 439-444.
- Lau S.Y., Phuan S.L., Danquah M.K., Acquah C., 2019, Sustainable palm oil refining using pelletised and surface-modified oil palm boiler ash (OPBA) biosorbent, *Journal of Cleaner Production*, 230, 527-535.
- MATLAB, 2019, 9.6.0.1072779 (2019a), The MathWorks, Inc., Natick, Massachusetts, United States.
- MPOB, 2017, Production of Crude Palm Oil, Malaysian Palm Oil Board <bepi.mpob.gov.my/index.php/en/production/production-2017/production-of-oil-palm-products-2017.html> accessed 21.02.2021.
- Rahman Khan M.A., Rahman M.M., Kadirgama K., Ismail A.R., 2012, Mathematical model for wear rate of negative graphite electrode in electrical discharge machining on Ti-5Al-2.5Sn, *Jurnal Teknologi*, 59(2), 55-59.
- Rashid N.A., Shamsuddin A., Mohd Rosely N.A., Mohd Noor M.A., Mazilamani L.S., Abd Hamid M.K., Asri Ibrahim K., 2018, Refined bleached deodorised palm oil quality prediction using kernel principal component regression, *The 7th International Graduate Conference on Engineering, Science & Humanities*, 264-266.
- Riyadi N., Mulki M.F., Susanto R., 2019, Analysis of customers purchase patterns of e-commerce transactions using apriori algorithm and sales forecasting analysis with weighted moving average (WMA) method, *Scientific Research Journal*, 7(7), 45-58.
- Sulaiman N.S., Mohd-Yusof K., Mohd-Saion A., 2018, Quality prediction modeling of palm oil refining plant in Malaysia using artificial neural network models, *International Journal of Engineering & Technology*, 7(3.26), 19-22.
- Wan X., Wang W., Liu J., Tong T., 2014, Estimating the sample mean and standard deviation from the sample size, median, range and/or interquartile range, *BMC Medical Research Methodology*, 14(1), 135.