Quality prediction and diagnosis of refined palm oil using partial correlation analysis

N A Rashid¹, K W Hoong¹, A Shamsuddin¹,², N A M Rosely¹, M A M Noor¹, K W Jin¹, M H Lee³ and M K A Hamid⁴
¹ Process Systems Engineering Centre (PROSPECT), Research Institute of Sustainable Engineering (RISE), Universiti Teknologi Malaysia, 81310 UTM Johor Bahru, Johor, MALAYSIA
² Lahad Datu Edible Oils Sdn Bhd, KM 2, Jalan Minyak off Jalan POIC, Locked Bag 16, 91109 Lahad Datu, Sabah, MALAYSIA
³ Department of Mathematical Science, Universiti Teknologi Malaysia, Johor, MALAYSIA
⁴ School of Chemical and Energy Engineering, Universiti Teknologi Malaysia, 81310 UTM Johor Bahru, Johor, MALAYSIA

E-mail: noradhiah@gmail.com (N A Rashid) / kamaruddinhamid@utm.my (M K A Hamid)

Abstract. Regression technique such as partial correlation analysis has been widely used as tool of prediction in business, finance and biomedical field. However, the application of predictive analysis in chemical process, specifically palm oil refinery process has rarely been done. Therefore, the objective of this paper is to present a quality prediction and diagnosis tool using partial correlation analysis, with the aim to predict the quality of refined palm oil and to diagnose the crude palm oil and process variables. Several statistical analysis are applied in data pre-process to obtain statistical sample size, optimum sampling and processing time of the process. The predictor coefficient is developed using partial correlation analysis while control chart is used to monitor the process behavior of both predicted and actual output value. The monitored out-of-control behavior is then diagnosed using SPE-contribution plot to identify the faulty input variables, thus pre-treatment can be executed before the refining process. The predicted model is successfully developed with MSE value less than 0.01 and three faulty variables are identified.

1. Introduction
Natural variation in Crude Palm Oil (CPO) and process changes in palm oil refinery process have greatly influence the quality of Refined Bleached Deodorized Palm Oil (RBDPO). The quality analysis can only be done after completion of batch run which may take up to 4-6 hours, thus causing a time-delay to the quality control system without control over the on-going production. Consequently, if the product quality does not achieve the standard quality specifications, it has to be recycled into the refining process. The recycle process of off-specification products quality does not only resulted in large time-delay and, on top of that, it will also increase the loss of profit, approximately RM 159,900.00 per hour due to the opportunity loss and production cost. To sustain the production of high quality final product, a process should be producing low variability product. Rather than to discover
the poor quality final product at the end of process, real-time process monitoring should be conducted to discover the unusual variability, hence process adjustment can be made beforehand.

In this work, a quality prediction and diagnosis tool is developed by using Partial Correlation Analysis (PCorrA) and Squared Prediction Error (SPE) contribution plot. The historical data is analyzed and transformed into a prediction model which then applied to predict future output quality parameter, given set of input parameter [1]. Diagnosis is capable to identify the faulty process variables which lead to the out-of-control behavior in a monitoring system. The process information on the process behavior and stability is extracted, so that the variability in the process can be determined and the source of out-of-control behavior can be identified [2].

Therefore, by developing the quality prediction and diagnosis tool, the RBDPO quality can be effectively predicted ahead of time based on the raw CPO data and the process will have adequate time for counteractions such as reconfiguration, maintenance or repair [3]. Besides, the clarification of the process factor that affects the RBDPO quality allows the manufacturer to make a better choice of process parameters at early design stage. In such a way, the production of high quality RBDPO can be sustained. This can reduce the dependence of recycle streams to rework off-specification RBDPO, thus saving time and cost for the whole RBDPO production for the long run.

2. Methodology

2.1 Pre-Processing Data

The data collected are standardized to convert each variable involved containing into set of data with uniform range of scale, standard deviation, and mean-centred [4]. The standardized data are distributed into different sample size with different boxplots and histograms drawn for each sample size. The skewness and kurtosis of each sample size is checked based on central limit theorem. Autocorrelation is then used to determine the optimum sampling time, followed by cross-correlation to determine the optimum processing time.

2.2 Development of Quality Prediction Tool

In order to improve the accuracy of the prediction model, variable selection is conducted using reliefF algorithm to select important features and removing insignificant variables from the dataset [5]. Apart from the four main CPO process variables (FFAin, MOISTin, IVin, DOBIin), more process variables are selected as the model input variables. Partial Correlation Analysis (PCorrA) is applied to derive the predictor coefficient, \( k \). PCorrA is a great tool in determining the correlation between two variables while holding other extraneous and intervening variables at constant values [6] as expressed in equation (1).

\[
k(X,Y :Z) = \frac{k(X,Y)- k(X, Z)k(Y, Z)}{\sqrt{[1-k^2(X, Z)][1-k^2(Y, Z)]}}
\]

(1)

Where \( X \) and \( Y \) are the matrix of input and output variables; \( Z \) is the matrix of intervening variables; \( k \) is the Pearson correlation between the variables. It is noted that the final \( k \) value is the predictor coefficient. Multivariate regression is applied to predict the output quality variable using predictor coefficient as expressed in equation (2).

\[
\hat{y}_i = k_1x_{i,1} + k_2x_{i,2} + k_3x_{i,3} + \ldots + k_px_{n,p}
\]

(2)

Control chart is used to monitor the process changes over time [7]. The process is said to be in control if the data lie within the upper and lower control, whereas the out-of-control process behavior occur when the data fall beyond the control limit. The point beyond the control limit depicts the quality deterioration due to some assignable causes. Both actual and predicted output values are plotted on the chart to visually display the deviation error and to pinpoint the possible out-of-control behaviour.
2.3 Development of Diagnosis Tool
The out-of-control observation monitored from the control chart is diagnosed using Square Prediction Error (SPE) contribution plot. Variables with the largest residuals or errors indicate that it produce the worst compliance to the prediction model, hence it is diagnosed as the main source of quality deterioration among the input variables [8].

\[ y_p = \hat{y}_{p1} + \hat{y}_{p1} + \hat{y}_{p1} + \ldots + \hat{y}_{pn} \]  

(3)

\[ \hat{y}_{pn} = k_n x_n \]  

(4)

\[ C_p = (y_p - \hat{y}_p)^2 \]  

(5)

Where \( y_p \) is the sum of predicted output value; \( k_n \) is the predictor coefficient calculated using PCorrA; \( x_n \) is the input variables; \( C_p \) is the contribution of the variable to the sum of predictor errors.

3. Results
Based on the boxplot and histogram plots in figure 1 and figure 2, sample size of 25 was determined as the best sample size, with balanced visual distribution for boxplot and consistent bell-shaped curve resemblance for histogram. For the autocorrelation plots from figure 3, the optimum sampling time was found to be 8 h. It was calculated by determining the largest lag value fell below the threshold value for all plots, which was then multiplied with the original time interval of the data (1 lag*8 h= 8 h optimum sampling time). Similar procedures were conducted for the cross-correlation plots, except the greatest lag value determined was multiplied with the optimum sampling time. The optimum processing time was found to be 24 h (3rd lag*8 h = 24 h optimum processing time) as plotted in figure 4.

Figure 1. Boxplot of 25 sample size.

Figure 2. Histogram of 25 sample size.
For variable selection, eight significant process variables were identified in which theoretically, these input variables greatly influenced the output RBDPO quality [9]. The identified input variables were FFA of CPO (X1), Moisture Content of CPO (X2), Iodine Value of CPO (X3), Deterioration of Bleachability Index (X4), Temperature of CPO feed oil (X5), Temperature of exchanger (X6), Level of dryer tank (X7), CPO flowrate (X8), Pressure of Niagara Filter 2 (X9), Temperature of Spiral 3 (X10), Pressure of Deodorizer 2 (X11) and Pressure of Deodorizer 3 (X12).

Mean Square Error (MSE) technique was used to evaluate the performance of the prediction models developed. The MSE for both training and testing data were compared as shown in figure 5. The MSE value of training data for FFAout, IVout and COLORout quality variables prediction were lower than the testing data, which indicates the model is overfit to the training data and poor generalization capability [10]. However, the testing data has lower MSE than the training MSE for MOISTout prediction and that indicates the prediction model was well generalized. Although the prediction model produced inconsistent MSE value, the prediction tool was considered reliable, given all the MSE values were approximately zero. Hence, it was proven that PCorrA technique was effective as quality prediction tool.

Figure 3. Autocorrelation plot.

Figure 4. Cross-correlation plot.

Figure 5. MSE of output quality variables prediction for training and testing data.
Control chart was plotted to monitor process variation over time and to identify the abnormal process behavior or unusual variation. The control chart was constructed for the actual and predicted value of RBDPO quality. As shown in figure 6, the actual output of MOISTout quality contains a few out-of-controls at observation number 3, 14 and 17. The sudden out-of-control spikes of MOISTout indicate that the process exhibit unnatural variability due to some assignable cause or known as special cause variation [11]. Special cause variation is a shift in output caused by specific factor such as environmental condition or process input parameters which can be identified and potentially be removed from the process [12].

![Figure 6. Control chart of moisture content quality prediction.](image-url)

The special cause variation of out-of-control spike monitored from the control chart was identified through diagnosis. Square Prediction Error (SPE) contributions were plotted for each out-of-control observation to identify the faulty input variables. As shown in figure 7, each out-of-control observation was affected by different input variables. The variable X3 (iodine value) was identified as the faulty variable for 3rd observation. In contrast, variable X12 (Pressure of Deodorizer 3) was identified as main source of out-of-control for 14th observation while the out-of-control for 17th observation was affected by the variable X7 (Level of dryer tank). Hence, it was shown that SPE-contribution plot was able to diagnose the source of fault in the output quality variable, given the location or observation number of out-of-control point.
Figure 7. Contribution plot of moisture content quality during; (a) observation number 3; (b) observation number 14 and (c) observation number 17.

4. Conclusion
Partial Correlation Analysis was proved as a competent tool for quality prediction. The special cause variation for the out-of-control behavior which were iodine value of CPO, pressure of deodorizer 3 and level of dryer tank have been identified using SPE contribution plot. Further corrective action and adjustment should be executed to the major causes of the out-of-control output quality variables so that optimal plant performance can be fine-tune. However, the quality prediction tool was developed with the assumption of the process to be steady state, thus dynamic model will be studied in future work.

Acknowledgment
The financial support from Universiti Teknologi Malaysia (RUGS Tier 1 Q.J130000.2546.19H72) is highly acknowledged.

References
[1] Zou M, Zhao L, Wang S, Chang Y and Wang F 2018 Quality analysis and prediction for start-up process of injection molding processes IFAC-Papersonline51 18 233-38
[2] De Ketelaere B, Rato T, Schmitt E and Hubert M 2016 Statistical process monitoring of time-dependent data Quality Engineering 28(1) 127-42
[3] Kargar S M, Salahshoor K and Yazdanpanah M J 2014 Multiple model-based fault detection and diagnosis for nonlinear model predictive fault-tolerant control Arabian Journal for Science and Engineering 39 7433-42
[4] Rashid N, Mohd Rosely N, Mohd. Noor M, Shamsuddin A, Abd. Hamid M and Asri Ibrahim K 2017 Forecasting of refined palm oil quality using principal component regression Energy Procedia 142 2977-82
[5] Jain D and Singh V 2018 Feature selection and classification systems for chronic disease prediction: a review Egyptian Informatics Journal 19 179-89
[6] Li X, Qiu T, Chen G, Zhong L-X and Wu X-R 2017 Market impact and structure dynamics of the chinese stock market based on partial correlation analysis Physica A: Statistical Mechanics and its Applications 471 106-13
[7] Hrehova S 2016 Predictive model to evaluation quality of the manufacturing process using matlab tools Procedia Engineering 149 149-54
[8] Bendjama H, Boucherit M S, Bouhouche S and Bast J 2011 Fault detection and diagnosis using principal component analysis application to low pressure lost foam casting process Int. J. Modelling, Identification and Control 14 102-11
[9] Ishibashi K, Iwakoji T, Otomasa S and Yada K 2016 Model selection for financial statement analysis: variable selection with data mining technique *Procedia Computer Science* **96** 1681-90.

[10] Kracik J and Strnadel B 2018 A statistical model for lifespan prediction of large steel structures *Engineering Structures* **176** 20-27.

[11] Bersimis S, Sgora A and Psarakis S 2016 The application of multivariate statistical process monitoring in non-industrial processes *Quality Technology & Quantitative Management* **15(4)** 526-49.

[12] Gejdoš P 2015 Continuous quality improvement by statistical process control *Procedia Economics and Finance* **34** 565-72.