Pertanika J. Sci. & Technol. 28 (4): 1179 - 1202 (2020)

SCIENCE & TECHNOLOGY

Journal homepage: http://www.pertanika.upm.edu.my/

Ridge Regression as Efficient Model Selection and Forecasting of Fish Drying Using V-Groove Hybrid Solar Drier

Hui Yin Lim[1], Pei Shan Fam[1]*, Anam Javaid[1,2] and Majid Khan Majahar Ali[1]

[1]*School of Mathematical Sciences, Universiti Sains Malaysia 11800 USM, Penang, Malaysia*
[2]*Department of Statistics, The Women University, Multan, Pakistan*

ABSTRACT

Application of the Internet of things (IoT) for data collection in solar drying can be very efficient in collecting big data of drying parameters. There are many variables involved so it is hard to find a model to predict the moisture content of the food product during drying. In model building, interaction terms should be incorporated because they also contribute to the model. Eight selection criteria (8SC) is a very useful method in model building. This study applied ordinary least squares (OLS) regression and ridge regression with 8SC in model building to predict the moisture content of drying fish. A total of eighty models were considered in this study. One best model was chosen each from OLS regression and ridge regression. M78.7.3 with a total of eleven independent variables was the best OLS model after conducting multicollinearity and coefficient test. Next, the best ridge model M56.0.0 was obtained after the coefficient test. The mean absolute percentage error (MAPE) was used to measure the accuracy of the prediction model. For OLS model M78.7.3, the MAPE value was 15.7342. The MAPE value for ridge model M56.0.0 was 17.4054. From the MAPE value, OLS model M78.7.3 provided a better estimation than the ridge model M56.0.0. However, OLS model M78.7.3 violated the normality assumptions of residuals. This is highly caused by the outlier problem. So, due to non- normality of the residuals and presence of outliers in the dataset, ridge regression is preferred for the best forecast model.

Keywords: Eight selection criteria, IoT, model selection, ordinary least squares, ridge regression

ARTICLE INFO

Article history:
Received: 27 March 2020
Accepted: 9 June 2020
Published: 21 October 2020

DOI: https://doi.org/10.47836/pjst.28.4.04

E-mail addresses:
limhuiyincorrine@student.usm.my (Hui Yin Lim)
fpeishan@usm.my (Pei Shan Fam)
anamjavaid7860@yahoo.com (Anam Javaid)
majidkhanmajaharali@usm.my (Majid Khan Majahar Ali)
*Corresponding author

ISSN: 0128-7680
e-ISSN: 2231-8526

INTRODUCTION

The global food demand grows rapidly due to the increase of world population (Bodirsky et al., 2015). Hence, the rising of the food demand brings to food insecurity issues. Food security is defined as “all people, at all times, have physical and economic access to sufficient, safe and nutritious food to meet their dietary needs and food preferences for a healthy and active life” (FAO, 1996). Therefore, to deal with food insecurity, substantial improvements in food processing are required to satisfy the increased food demand.

Drying is one of the food post-processing techniques, which plays a vital role in the preservation of agriculture crops and marine harvest (Silva et al., 2017 and Ali et al., 2017b). It reduces the moisture content of food to inhibit the growth of microorganisms. The advantages of drying include longer shelf life, smaller size for storage purpose and lighter weight for transportation (Ertekin & Yaldiz, 2004). Traditional drying involves the process of drying agriculture crops or marine harvest under the direct sun exposition (Tiwari, 2016).

However, dehydrated food products will be contaminated easily due to the exposure of direct sunlight in open space. Besides, non-uniform sun-drying under open space increases the chance of fungal attack and the growth of microorganisms (Tiwari, 2016). Open sun drying also cannot control the drying parameter due to weather uncertainties. Furthermore, this conventional drying method is very time-consuming. The conventional method of fish drying that is still being used is shown in Figure 1.

Therefore, the effort to improve sun drying has led to the usage of renewable energy, specifically solar drying. For instance, Ali et al. (2017a), Stiling et al. (2012), Hossain and Bala (2007), Alfiya et al. (2018) and many other researchers applied solar drying by

Figure 1. Traditional method of fish drying under direct sunlight

using solar drier in their study. Furthermore, the Internet of things (IoT) based solar drying system using v-Groove Hybrid Solar Drier (v-GHSD) by Ali et al. (2017a, 2017b) was more effective in monitoring the drying behavior.

Since the development of drier, especially v-GHSD provides more benefits in terms of quality and hygienic aspects, all the important factors involved with the solar drying system should be investigated.

Drying parameters play an important role in the drying process. Tiwari (2016) stated that temperature, air humidity, area of exposed surface and pressure had effects on the removal of the moisture content. Besides, Silva et al. (2017) found out air temperature was a very important factor that would affect the drying process. Furthermore, Krokida et al. (2003) found out drying temperature had more influence than the air velocity and air humidity during the drying process. Hence, all of these drying parameters may contribute to the fish drying process. However, there is a very limited research study on the effect of important drying parameters and its interaction terms for fish drying using solar drier towards the fish drying model.

Furthermore, Javaid et al. (2020) found that there were significant interactions among variables in the drying seaweed process. Hence, regression analysis is one of the existing methods to investigate the relationship between variables in a data set and a continuous response variable with the interaction terms.

Ordinary Least Squares (OLS) is one of the popular estimation methods for the linear regression model. OLS regression estimates the functional relationship by minimizing the sum of squares differences between the observed and predicted response variable. It produces unbiased estimates with the smallest standard errors and provides the best linear unbiased estimator (BLUE) if all the model assumptions are satisfied (Wen et al., 2013). However, real data always suffer from multicollinearity. The application of least squares method in parameter estimation in the presence of multicollinearity may cause the estimates becoming unstable (Mahajan et al., 1977).

Apart from multicollinearity, the outlier is also one of the problems in regression analysis. Rajarathinam and Vinoth (2014) stated that outliers were commonly present in agriculture production data due to uncontrolled factors. Outliers will inflate the error variance as well as the standard errors. OLS estimator is extremely sensitive to outliers in linear regression analysis. However, agriculture and marine production data always suffer from multicollinearity and outlier problems. Hence, a suitable method should be done to solve these problems in the fish drying data. The initial moisture content of fish is between eighty-two percent, and the moisture content needs to be reduced to thirty-five percent after drying in the solar drier to achieve Equilibrium Moisture Content (EMC).

To overcome the limitations of the OLS estimator, researchers implemented a few methods. Regularization is one of the most common approaches to solve multicollinearity.

Regularization methods can be applied to control the instability of OLS estimates. Ridge regression is one of the regularization methods that shrinks the coefficients towards zero by minimizing the mean square error of the estimates (Ullah et al., 2018).

Furthermore, Steece (1986) concluded that ridge estimation was able to curb outliers in regressor space by downweighting their influence. Besides, Chatterjee and Hadi (2015) also stated that ridge estimators were stable as they were not affected by slight variations in the estimation data. Hence, ridge regression provides estimates that are more robust as compared to least squares estimates for small perturbations in the data.

Many researchers such as Delaney and Chatterjee (1986), Golub et al. (1979) and Kennard (1971) studied on estimation of biasing parameter in the ridge regression. There are also many proposed methods in selecting the biasing parameter but it does not have a general agreement on the best way to choose an optimal value of the biasing parameter (Khalaf, 2012). Besides, Zhang and Ibrahim (2005) stated that it was uncertain if ridge regression provided better estimates than OLS regression during different applications. Therefore, a more thorough approach is using the *lmridge* package in R developed by Ullah et al. (2018) to estimate the biasing parameter because ridge regression is a multiple regression with no penalty. Ullah et al. (2018) stated that the *lmridge* package in R provided suitable tools for ridge regression analysis in R as compared to other packages.

During model building, most of the researchers in the agriculture field only consider the individual term without considering the interaction term between the variables. For example, Jamal and Rind (2007) did not include interaction terms in developing the forecast models for acreage and production of the wheat crop in their study. However, interaction terms should be included during model building to avoid bias. Therefore, Javaid et al. (2019a) also addressed the interaction terms in their regression model to examine the main factors with their interaction terms affecting the collector efficiency, and they found that the interaction terms had a significant effect in the best final model.

Eight selection criteria (8SC) are always used for model selection purpose. For instance, in the study of Abdullah et al. (2015), they found that the application of multiple regression with 8SC was able to model and forecast biomass and biofuel production. Besides, Abdullah et al. (2011) used the polynomial regression technique with 8SC to find out the best model to estimate the volumetric stem biomass. Javaid et al. (2019b) applied multiple regression with 8SC in their study on forecasting the moisture ratio removal during the seaweed drying process. Yahaya et al. (2012) selected the best model in estimating the electrical conductivity levels by using 8SC.

Fish drying data were fitted to the thin layer drying model by many researchers. For example, Guan et al. (2013) applied nine thin layer models and found out the Page model was able to predict and describe the drying process more accurately. Kituu et al. (2010) also applied a thin layer model in drying fish. However, the thin layer model is used to understand the drying behavior and does not involve model building. Besides, the thin layer drying model does not incorporate the interaction term in the drying model.

Furthermore, there is limited research conducted on the moisture content of drying fish and the factors affecting it with its interaction terms by using ridge regression with 8SC. Besides, in different applications, the performance of OLS regression and ridge regression may vary. Hence, OLS regression and ridge regression were conducted in this study. From all possible models, 8SC was applied for the model selection purpose to choose the best model to forecast the moisture content of drying fish.

MATERIALS AND METHODS

v-GHSD

The v-GHSD used in the fish drying process in this study consists of fans that back powered by solar panels. Besides, it also consists of a drying chamber, solar collector, v-aluminum roof, solar panel, and sensors using IoT for data collection every thirty minutes. The sensors are placed to measure the inlet and outlet temperature, inlet and outlet humidity, wind speed, and solar radiation. For this study, we looked at the effect of some factors and their interaction. Figure 2 shows the v-GHSD used in this study. Figure 3 shows the Chemical Fluid Dynamic (CFD) analysis using original data collected by using IoT and the parameters involved in this study.

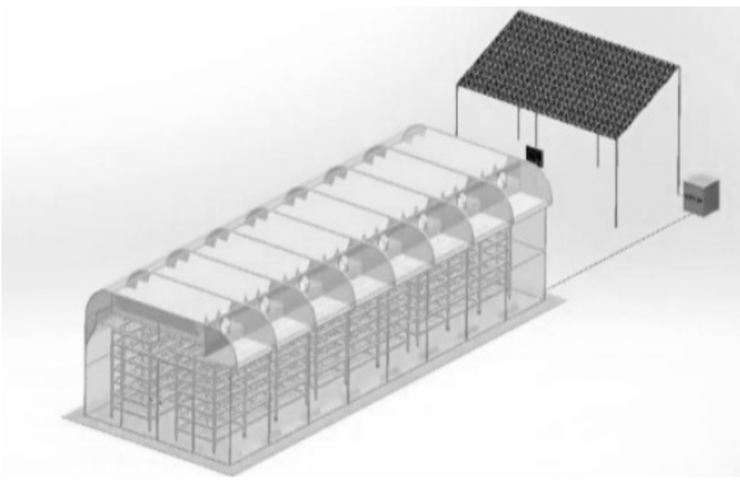

Figure 2. Simulation diagram of v-GHSD

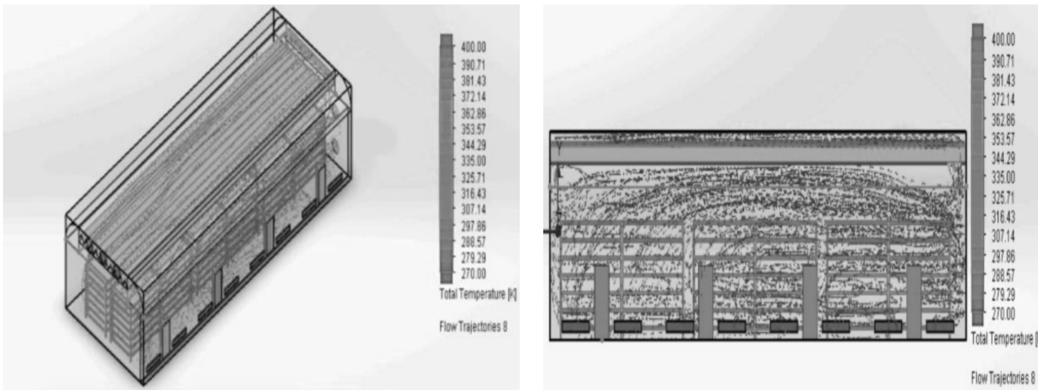

Figure 3. CFD Simulation diagram of v-GHSD

Model Development

Consider a multiple regression model (Equation 1),

$$y = X\beta + \varepsilon, \qquad (1)$$

where y is a n x 1 vector of response variables, X is known as the design matrix of order n x p, β is a p x 1 vector of unknown parameters and ε is a n x 1 vector of identically and independent distributed errors.

According to Gujarati (2004), the OLS estimator of β is obtained as in Equation 2

$$\hat{\beta} = (X'X)^{-1}X'y. \qquad (2)$$

In Equation 2, if the regressors are nearly dependent, matrix $X'X$ becomes ill conditioned. Hence, Hoerl and Kennard (1970) suggested ridge estimator as in Equation 3,

$$\hat{\beta}^{ridge} = (X'X + \lambda I)^{-1}X'y, \qquad (3)$$

where λ is a ridge parameter and l is an identity matrix. The ridge parameter, $\lambda > 0$ indicates the degree of shrinkage. Note that a value $\lambda = 0$ gives rise to OLS estimates.

Golub et al. (1979) proposed generalized cross-validation (GCV) as a method for choosing the ridge parameter (Equation 4).

$$GCV = \frac{SS_e}{\left(n - tr(H)\right)^2} \qquad (4)$$

where SS_e refers to the residual sum of squares of a model using the ridge coefficients and H refers to an augmented hat matrix (Equation 5),

$$H = X(X'X + \lambda I)^{-1}X'. \qquad (5)$$

We look for λ value that minimizes Equation 4. The ridge regression is carried out if the λ obtained is greater than zero for minimum GCV. If λ obtained is equal to zero, then ridge regression will be automatically equal to the OLS regression analysis. The *lmridge* package in R software was used in this study.

Phase 1– All Possible Models

Phase 1 involves computations of all possible models for the best model selection. According to Ali et al. (2017a), the formulae to compute the total number of all possible models are shown in Equation 6:

$$N = \sum_{j=1}^{k} j\left(k_{C_j}\right) \tag{6}$$

where N indicates the number of possible models, k indicates the total number of independent variables and j is 1, 2, ..., k. C shows the combinations for all possible models.

By using Equation 6, all possible models are computed.

Phase 2- Selected Models

Multicollinearity is checked among the variables by obtaining the correlation matrix for all factors. Only one highly correlated variable is removed from the analysis at a time. This procedure is performed until there is no collinear variable left in the model. However, for ridge regression, there is no need to check the problem of multicollinearity as it has the ability to deal with this problem.

Once the multicollinearity is checked among the variables in all possible models, a coefficient test is conducted for the OLS regression model after the model is free from the multicollinearity issue. For the ridge regression model, the coefficient test is conducted directly without checking the multicollinearity. The coefficient test is conducted in this phase to check the significance of the individual regression coefficient, β_j at the 5% level of significance. Adding an unimportant variable may make the model worse. The hypothesis statement of the coefficient test is shown as below:

$$H_0: \beta_j = 0,$$

$$H_1: \beta_j \neq 0.$$

where β_j is the coefficient of variable in the model for $j = 1,2, ..., k$. The test statistics of this test is (Equation 7)

$$t_0 = \frac{\hat{\beta}_j}{s(e\hat{\beta}_j)} \tag{7}$$

where $\hat{\beta}_j$ is the estimated regression coefficient of β_j and $s(e\hat{\beta}_j)$ is the standard error of $\hat{\beta}_j$.

Note that the null hypothesis is rejected if $|t_0| > t_{\frac{\alpha}{2},n-k-1}$. If the null hypothesis is rejected, then the selected parameter will be eliminated from the regression model. The selected model will be renamed as shown in Figure 4, where M denotes the model.

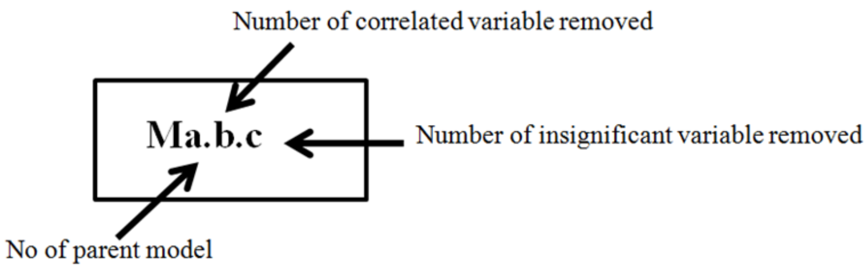

Figure 4. Model labeling in regression model

Phase 3 - The Best Model

Next, the selection of the best model from every selected model is conducted by using 8SC. According to Ali et al. (2017a, 2017b), the 8SC includes Akaike information criterion (AIC), RICE, Final prediction error (FPE), SCHWARZ, generalized cross-validation (GCV), sigma square (SGMASQ), Hannan-Quinn information criterion (HQ) and SHIBATA. The formulae of all of the model selection criteria are listed in Table 1. The most efficient model is selected based on the most number of the minimum value of the selection criteria.

Where SSE indicates the sum of squares error, $k+1$ indicates the number of estimated parameters and n indicates the sample size. According to Hajijubok and Gopal (2008), the condition that needs to be fulfilled when doing evaluation by using these model selection criteria is $2(k+1) < n$.

Phase 4 - Goodness of Fit

Five percent of the dataset reserved previously was used as test data to fit into the final best model chosen from phase 3. Then, residual analysis was conducted. The residual analysis is very important to check the randomness and normality of the residuals. In this study, a run test was used to check the randomness of the residuals, while the Kolmogorov-Smirnov test is used to check the normality assumption of the residuals. However, if the best model obtained from phase 3 is ridge regression model, then the normality of the residuals is not required because ridge regression does not require the residuals normality assumptions. Scatter plot and box plot of the residuals are used as supporting evidence of the goodness of fit test. Besides, the mean absolute percentage error (MAPE) is calculated as a measure of prediction accuracy (Ali et al., 2017a). The smaller the MAPE value the better, the higher the prediction accuracy. The formula of MAPE is shown in Equation 8:

$$MAPE = \frac{100}{N}\left(\frac{\sum_{i=1}^{j}|A_i - E_i|}{A_i}\right) \quad \text{for } i = 1,2,\dots,j \tag{8}$$

where
A = Actual value of dependent variable (y)
E = Expected value $(\hat{y})$
N = Number of reserved data.

Table 1
Formula used for 8SC

AIC: $\left(\frac{SSE}{n}\right)(e)^{2(k+1)/n}$ Akaike (1969)	RICE: $\left(\frac{SSE}{n}\right)\left[1-\left(\frac{2(k+1)}{n}\right)\right]^{-1}$ (Rice, 1984)
FPE: $\left(\frac{SSE}{n}\right)\frac{n+(k+1)}{n-(k+1)}$ (Akaike, 1974)	SCHWARZ: $\left(\frac{SSE}{n}\right)n^{(k+1)/n}$ (Schwarz, 1978)
GCV: $\left(\frac{SSE}{n}\right)\left[1-\left(\frac{k+1}{n}\right)\right]^{-2}$ (Golub et al., 1979)	SGMASQ: $\left(\frac{SSE}{n}\right)\left[1-\left(\frac{k+1}{n}\right)\right]^{-1}$ (Ramanatam, 2002)
HQ: $\left(\frac{SSE}{n}\right)(\ln n)^{2(k+1)/n}$ (Hannan & Quinn, 1979)	SHIBATA: $\left(\frac{SSE}{n}\right)\frac{n+2(k+1)}{n}$ (Shibata, 1981)

All the four phases are summarized in Figure 5.

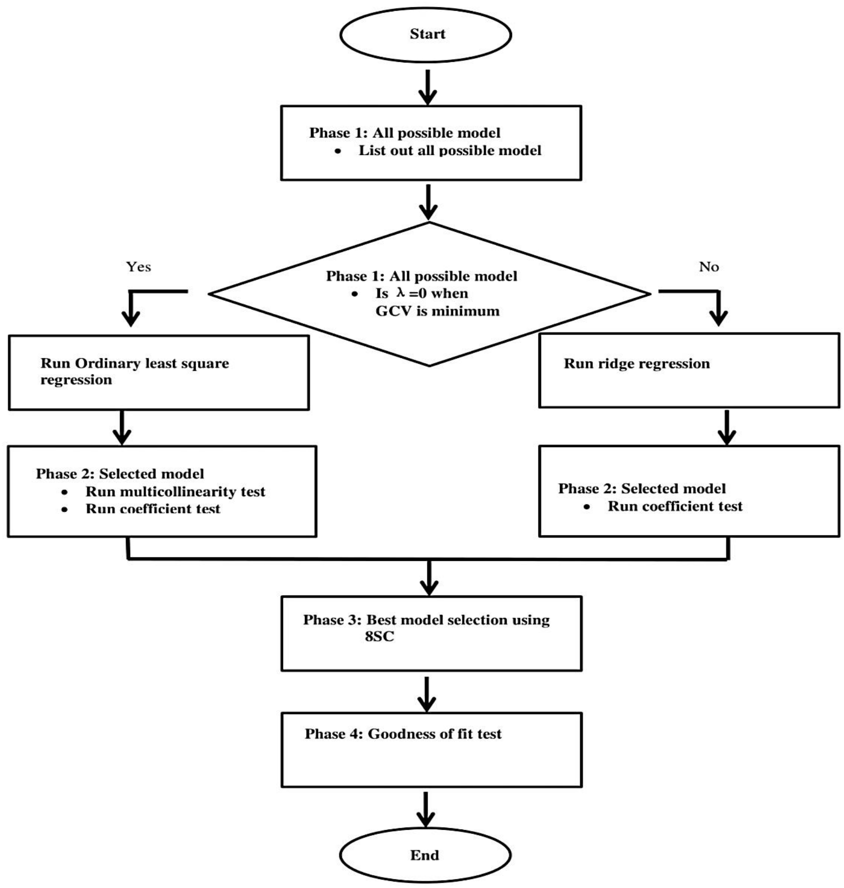

Figure 5. Flow Chart on the Procedures in Getting Best Model

RESULTS AND DISCUSSIONS

Data Collection and Procedure

In this study, the data were taken during the experiment drying process for drying fish by using v-GHSD at Selakan Island, Semporna. The fish was dried to thirty-five percent moisture content until it reached the EMC. The data collection started from 8th to 12th October 2019. The total number of data collected was 1914 and there were no missing data. Five percent of the dataset which is 96 data was reserved as test data. In this study, moisture content of fish (y)is the dependent variable, whereas the inlet temperature chamber (X_1), outlet temperature chamber (X_2), outlet humidity chamber (X_3), inlet humidity chamber (X_4) and solar radiation (X_5) are the independent variables. The five days drying data was collected for every thirty minutes.

Since five independent variables were used in this study, there were total 80 possible models until fourth order of interaction as shown in Table 2.

Table 2

All possible models

No of variables	Single	Interact				Total	Model Label
		1st Order	2nd Order	3rd Order	4th Order		
1	5	-	-	-	-	5	**M1-5**
2	10	10	-	-	-	20	**M6-25**
3	10	10	10	-	-	30	**M26-55**
4	5	5	5	5	-	20	**M56-75**
5	1	1	1	1	1	5	**M76-80**
Total Models	31	26	16	6	1	80	

The coefficient test is conducted, and a list of selected models with its ridge parameter λ and Error Sum of Squares (SSE) are obtained. Where k denotes the number of variables left in the model. The models with the same number of variables are kept in a single group. After grouping, the 69 models are left out of 80 possible models, and results are shown in Table 3. For example, M21.0.0 represents the original model. One variable is removed during the multicollinearity test so the model becomes M21.1.0 while no variable is removed from the coefficient test. So, the final model remains as M21.1.0.

Table 3

Selected Models by using OLS or Ridge Regression

Sr. NO	Selected models using OLS/Ridge	k	λ	SSE
1	*M1.0.0*	1	0.00000	407251.8445
2	*M2.0.0*	1	0.00000	382866.6136
3	*M3.0.0*	1	0.00000	415428.5487
4	*M4.0.0*	1	0.00000	496042.5512
5	*M5.0.0*	1	0.00000	346497.159
6	*M6.0.0=M16.1.0*	2	0.00800	381262.7479
7	*M7.0.0=M17.0.1*	2	0.00200	322025.8196
8	*M8.0.0*	2	0.00100	358834.0995

Table 3 *(Continued)*

Sr. NO	Selected models using OLS/Ridge	k	λ	SSE
9	M9.0.0	2	0.00500	342096.8842
10	M10.0.0=M20.1.0	2	0.00200	312301.581
11	*M11.0.0*	2	0.00000	303981.4905
12	*M12.0.0=M22.0.1*	2	0.00500	318511.0194
13	*M13.0.0*	2	0.00500	414295.4192
14	*M14.0.0*	2	0.00500	334854.2714
15	*M15.0.0*	2	0.00200	343805.3683
16	*M16.1.0*	2	0.00000	379272.0951
17	*M18.0.0*	3	0.00300	348307.4055
18	*M19.0.0*	3	0.01800	342107.8585
19	*M21.1.0*	2	0.00000	301316.8478
20	*M23.0.0*	3	0.00100	395358.3048
21	*M24.0.0*	3	0.01400	333336.4064
22	*M25.0.0*	3	0.00500	334033.4548
23	*M26.0.0*	3	0.01000	306848.4051
24	*M27.0.0*	3	0.00100	296660.9027
25	*M28.0.0*	3	0.00100	310557.9255
26	*M29.0.0=M59.0.1*	3	0.00100	288289.5733
27	*M30.0.0*	3	0.01000	315115.4032
28	*M31.0.0*	3	0.00600	325462.092
29	*M32.0.0*	3	0.00100	258597.1048
30	*M33.0.0=M57.0.1*	3	0.00900	294813.4643
31	*M34.0.0=M58.0.1*	3	0.00100	270406.179
32	*M35.0.0*	3	0.00600	333986.6613
33	*M36.3.0*	3	0.00000	304165.6933
34	*M37.2.1*	3	0.00000	289651.0102
35	*M38.3.0*	3	0.00000	315425.3665
36	*M39.0.1*	5	0.00900	287071.6748
37	*M40.2.0*	4	0.00000	294634.3668
38	*M41.1.1*	4	0.00000	319472.1287
39	*M42.2.0=M52.3.0*	4	0.00000	257707.6509
40	*M43.2.0*	4	0.00000	260872.4566
41	*M44.0.2*	4	0.00800	268781.3893
42	*M45.0.3*	3	0.00400	320402.8475

Table 3 *(Continued)*

Sr. NO	Selected models using OLS/Ridge	k	λ	SSE
43	M46.4.0	3	0.00000	304165.6933
44	M47.2.2	3	0.00000	289072.1514
45	M48.4.0	3	0.00000	314537.9494
46	M49.2.2	3	0.00000	287679.0882
47	M50.2.1	4	0.00000	291784.7212
48	M51.1.1	5	0.00000	318219.7034
49	M53.3.0	4	0.00000	264752.4059
50	M54.3.1	2	0.00000	281515.2994
51	M55.0.2	5	0.00600	317817.7528
52	M56.0.0/M76.0.1	4	0.00200	247253.6408
53	M60.0.0	4	0.00100	251008.3619
54	M61.4.2=M66.7.3=M71.8.3	4	0.00000	246598.7709
55	M62.5.0	5	0.00000	257482.0627
56	M63.4.2	4	0.00000	268756.3648
57	M64.2.2	6	0.00000	283819.5595
58	M65.3.3	4	0.00000	249022.3153
59	M67.9.1	4	0.00000	263008.9208
60	M68.8.2	4	0.00000	266633.687
61	M69.4.2	8	0.00000	280462.7962
62	M70.6.2	6	0.00000	247520.6769
63	M72.10.0	4	0.00000	266930.9018
64	M73.9.2	4	0.00000	268766.6487
65	M74.4.2	9	0.00000	276344.7077
66	M75.7.2	6	0.00000	248298.1572
67	M77.6.1	8	0.00000	244108.2359
68	M78.7.3=M79.17.4	11	0.00000	230561.7746
69	M80.18.3	9	0.00000	236260.0805

After the coefficient test, all of the best selected models, as shown in Table 4 are evaluated by using 8SC.

From the results in Table 4, M78.7.3 provides the minimum of all the 8SC value. Hence, M78.7.3 is obtained as the best model among all the selected models. Since M78.7.3 is with λ equal to 0, hence, this model is an OLS regression model. Furthermore, M56.0.0 with λ equal to 0.002 provides the minimum 8SC value for the ridge regression model. The best model M78.7.3 for OLS and M56.0.0 for ridge are shown as in Equation 9 and Equation 10 respectively. The coefficients are obtained using R software.

$$M78.7.3 = \hat{Y} = -105.3 + 5.007x_2 - 0.0515^{x_3} + 0.03444x_{14} + 0.0186x_{24} - 0.0007453x_{25} - 0.00279600x_{45} - 0.00133600x_{124} + 0.00004118x_{135} + 0.00003168x_{145} - 0.00011440x_{234} + 0.00002246x_{345} \quad (9)$$

Table 4

8SC for OLS/Ridge Selected Models

Selected models from OLS/Ridge	AIC	FPE	GCV	HQ	RICE	SCHWARZ	SGMASQ	SHIBATA
M1.0.0	224.5043	224.5043	224.5046	225.0066	224.5049	225.8682	224.2576	224.5038
M2.0.0	211.0616	211.0616	211.0618	211.5337	211.0621	212.3438	210.8296	211.0611
M3.0.0	229.0119	229.0119	229.0122	229.5242	229.0124	230.4031	228.7602	229.0113
M4.0.0	273.4517	273.4517	273.452	274.0634	273.4523	275.1129	273.1512	273.451
M5.0.0	191.0123	191.0123	191.0125	191.4396	191.0128	192.1727	190.8024	191.0118
M6.0.0=M16.1.0	210.4088	210.4088	210.4093	211.1152	210.4099	212.329	210.0621	210.4076
M7.0.0=M17.0.1	177.7175	177.7175	177.718	178.3142	177.7184	179.3394	177.4247	177.7165
M8.0.0	198.031	198.031	198.0315	198.6959	198.0321	199.8383	197.7047	198.0299
M9.0.0	188.7942	188.7942	188.7947	189.428	188.7952	190.5172	188.4831	188.7931
M10.0.0=M20.1.0	172.3509	172.3509	172.3514	172.9296	172.3519	173.9239	172.067	172.35
M11.0.0	167.7593	167.7593	167.7597	168.3225	167.7602	169.2903	167.4829	167.7584
M12.0.0=M22.0.1	175.7777	175.7777	175.7782	176.3679	175.7787	177.382	175.4882	175.7768
M13.0.0	228.6386	228.6386	228.6392	229.4063	228.6399	230.7252	228.2619	228.6374
M14.0.0	184.7972	184.7972	184.7977	185.4176	184.7982	186.4837	184.4927	184.7962
M15.0.0	189.737	189.737	189.7375	190.3741	189.7381	191.4686	189.4244	189.736
M16.1.0	209.3102	209.3102	209.3107	210.0129	209.3113	211.2204	208.9653	209.309
M18.0.0	192.4332	192.4332	192.4341	193.2951	192.435	194.7783	192.0107	192.4313
M19.0.0	189.008	189.008	189.0089	189.8546	189.0099	191.3115	188.5931	189.0062
M21.1.0	166.2887	166.2887	166.2892	166.8471	166.2897	167.8064	166.0148	166.2878

Table 4 *(Continued)*

Selected models from OLS/Ridge	AIC	FPE	GCV	HQ	RICE	SCHWARZ	SGMASQ	SHIBATA
M23.0.0	218.4279	218.4279	218.4289	219.4063	218.43	221.0899	217.9483	218.4258
M24.0.0	184.162	184.162	184.1629	184.9869	184.1638	186.4063	183.7577	184.1602
M25.0.0	184.5471	184.5471	184.548	185.3737	184.5489	186.7961	184.1419	184.5453
M26.0.0	169.5279	169.5279	169.5287	170.2872	169.5295	171.5939	169.1557	169.5262
M27.0.0	163.8995	163.8995	163.9003	164.6336	163.9011	165.8969	163.5396	163.8979
M28.0.0	171.5773	171.5773	171.5781	172.3458	171.579	173.6683	171.2006	171.5756
M29.0.0=M59.0.1	159.2745	159.2745	159.2752	159.9879	159.276	161.2155	158.9248	159.2729
M30.0.0	174.0952	174.0952	174.0961	174.875	174.0969	176.2169	173.713	174.0935
M31.0.0	179.8116	179.8116	179.8124	180.617	179.8133	182.0029	179.4168	179.8098
M32.0.0	142.8699	142.8699	142.8706	143.5099	142.8713	144.6111	142.5563	142.8686
M33.0.0=M57.0.1	162.8788	162.8788	162.8796	163.6084	162.8804	164.8638	162.5212	162.8772
M34.0.0=M58.0.1	149.3942	149.3942	149.395	150.0634	149.3957	151.2149	149.0663	149.3928
M35.0.0	184.5212	184.5212	184.5221	185.3477	184.523	186.77	184.1161	184.5194
M36.3.0	168.0457	168.0457	168.0465	168.7984	168.0473	170.0937	167.6768	168.0441
M37.2.1	160.0266	160.0266	160.0274	160.7434	160.0282	161.9769	159.6753	160.0251
M38.3.0	174.2665	174.2665	174.2673	175.047	174.2682	176.3902	173.8839	174.2648
M39.0.1	158.9509	158.9509	158.9527	160.0201	158.9544	161.8655	158.4281	158.9475
M40.2.0	162.959	162.959	162.9602	163.8719	162.9615	165.4453	162.5121	162.9566
M41.1.1	176.6965	176.6965	176.6978	177.6864	176.6992	179.3923	176.2119	176.6938
M42.2.0=M52.3.0	142.5353	142.5353	142.5363	143.3338	142.5374	144.7099	142.1443	142.5331

Table 4 *(Continued)*

Selected models from OLS/ Ridge	AIC	FPE	GCV	HQ	RICE	SCHWARZ	SGMASQ	SHIBATA
M43.2.0	144.2857	144.2857	144.2868	145.094	144.2879	146.487	143.8899	144.2835
M44.0.2	148.66	148.66	148.6611	149.4928	148.6623	150.9281	148.2523	148.6578
M45.0.3	177.0164	177.0164	177.0173	177.8093	177.0181	179.1737	176.6278	177.0147
M46.4.0	168.0457	168.0457	168.0465	168.7984	168.0473	170.0937	167.6768	168.0441
M47.2.2	159.7068	159.7068	159.7076	160.4222	159.7084	161.6532	159.3562	159.7053
M48.4.0	173.7762	173.7762	173.777	174.5546	173.7779	175.894	173.3947	173.7745
M49.2.2	158.9372	158.9372	158.9379	159.6491	158.9387	160.8741	158.5883	158.9356
M50.2.1	161.3829	161.3829	161.3841	162.287	161.3854	163.8451	160.9403	161.3805
M51.1.1	176.1975	176.1975	176.1995	177.3827	176.2014	179.4283	175.6179	176.1937
M53.3.0	146.4316	146.4316	146.4327	147.252	146.4339	148.6657	146.03	146.4294
M54.3.1	155.3608	155.3608	155.3612	155.8824	155.3616	156.7787	155.1048	155.36
M55.2.1	175.975	175.975	175.9769	177.1586	175.9788	179.2017	175.3961	175.9712
M56.0.0/M76.0.1	136.7533	136.7533	136.7543	137.5194	136.7553	138.8397	136.3782	136.7512
M60.0.0	138.83	138.83	138.831	139.6077	138.8321	140.9481	138.4492	138.8279
M61.4.2=M66.7.3=M71.8.3	136.3911	136.3911	136.3921	137.1551	136.3931	138.472	136.017	136.389
M62.5.0	142.5672	142.5672	142.5688	143.5262	142.5704	145.1814	142.0983	142.5641
M63.4.2	148.6462	148.6462	148.6473	149.4789	148.6484	150.914	148.2385	148.6439

Table 4 *(Continued)*

Selected models from OLS/Ridge	AIC	FPE	GCV	HQ	RICE	SCHWARZ	SGMASQ	SHIBATA
M64.2.2	157.3232	157.3232	157.3256	158.5585	157.3279	160.6938	156.7198	157.3186
M65.3.3	137.7315	137.7315	137.7325	138.5031	137.7336	139.8328	137.3537	137.7294
M67.9.1	145.4673	145.4673	145.4684	146.2823	145.4695	147.6867	145.0684	145.4651
M68.8.2	147.4721	147.4721	147.4733	148.2983	147.4744	149.7221	147.0677	147.4699
M69.4.2	155.805	155.805	155.8088	157.3796	155.8127	160.1098	155.0375	155.7974
M70.6.2	137.2025	137.2025	137.2045	138.2798	137.2066	140.142	136.6762	137.1984
M72.10.0	147.6365	147.6365	147.6377	148.4636	147.6388	149.889	147.2316	147.6343
M73.9.2	148.6519	148.6519	148.653	149.4846	148.6541	150.9198	148.2442	148.6496
M74.4.2	153.6862	153.6863	153.6909	155.413	153.6956	158.4115	152.8455	153.677
M75.7.2	137.6335	137.6335	137.6355	138.7141	137.6376	140.5822	137.1056	137.6294
M77.6.1	135.609	135.609	135.6123	136.9795	135.6157	139.3558	134.941	135.6024
M78.7.3=M79.17.4	128.507	128.507	128.5126	130.2416	128.5183	133.2628	127.6643	128.4959
M80.18.3	131.3936	131.3936	131.3976	132.8699	131.4016	135.4335	130.6748	131.3857

$$M56.0 = -66.2806 + 0.3273x_1 + 2.8945\,x_2 - 0.0317x_3 + 0.4450x_4 \quad (10)$$

For model M78.7.3, eleven variables were retained in the model, including the interaction terms. The signs of the coefficient show the type of relationship of the independent variable with the dependent factor. The coefficients that are far away from the zero mean that they are the strongest factors in the analysis. From the results, the significance of the variables with the interaction term shows that the interaction terms are very important and cannot be ignored. For model M56.0.0, four variables are remained in the model without including the interaction term. For both of the models, MAPE was computed by using formulae as stated in Equation 8. The MAPE value for M78.7.3 is 15.7342. The MAPE value for M56.0.0 is 17.4054. Both of the MAPE value is less than 20 and indicates both models can be used to forecast the moisture content of the fish.

```
        Runs Test

data:  std_res$stdres
statistic = 0.59445, runs = 971, n1 = 957, n2 = 957, n = 1914, p-value
= 0.5522
alternative hypothesis: nonrandomness
```

Figure 6. Run test for standardized residuals M78.7.3

```
        Runs Test

data:  std_res$stdres
statistic = 1.2346, runs = 985, n1 = 957, n2 = 957, n = 1914, p-value = 0.217
alternative hypothesis: nonrandomness
```

Figure 7. Run test for standardized residuals M56.0.0

```
        One-sample Kolmogorov-Smirnov test

data:  std_res$stdres
D = 0.070452, p-value = 1.121e-08
alternative hypothesis: two-sided
```

Figure 8. Kolmogorov-Smirnov test for standardized residuals M78.7.3

To test the randomness of the standardized residuals, a run test was conducted. From the results as shown in Figure 6, the run test p-value was equal to 0.5522 for M78.7.3. From the results as shown in Figure 7, the run test p-value was equal to 0.217 for M56.0.0. Since the p-value of the run test of both models is more than 0.05, hence, the standardized residuals are random. Furthermore, Kolmogorov-Smirnov test was conducted for M78.7.3 to test the normality assumptions of residuals. The results are shown in Figure 8. The p value obtained from the Kolmogorov-Smirnov test for M78.7.3 is less than 0.05. Therefore, the residuals are not normally distributed.

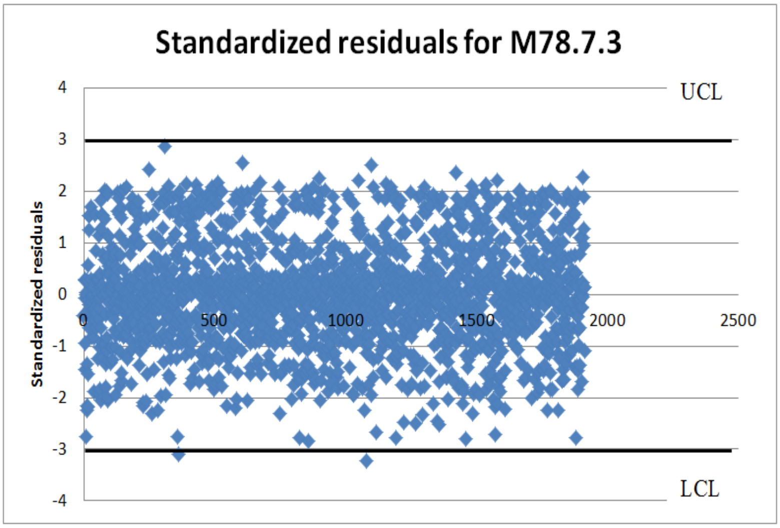

Figure 9. Scatterplot of standardized residuals for OLS regression M78.7.3

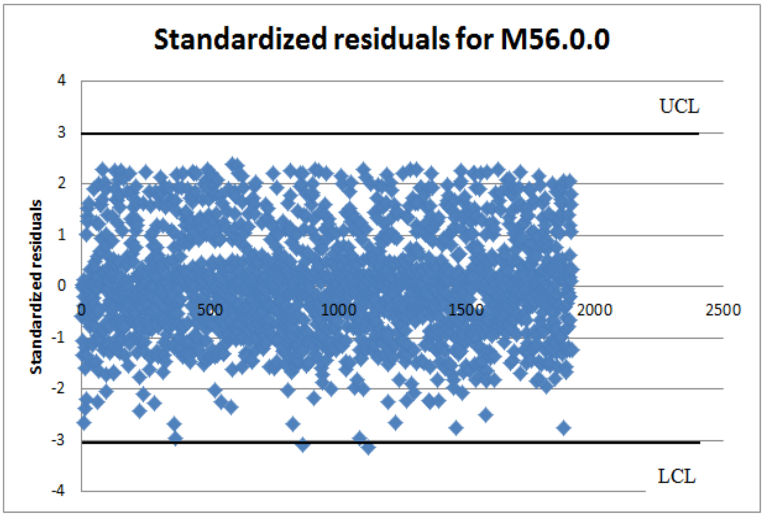

Figure 10. Scatterplot of standardized residual for Ridge regression M56.0.0

Outliers outside the 3-sigma limit can be observed from Figure 9 and 10. UCL and LCL represent the upper-class limit and lower-class limit respectively. The percentage of outliers is obtained based on the number of observations outside the 3-sigma limit. Table 5 shows the percentage of outliers outside 3-sigma limit for M78.7.3 and M56.0.0.

Table 5

Percentage of outliers outside 3-sigma limits

Selected model	Method	$\mu \pm 3\sigma$
M78.7.3	OLS	0.11%
M56.0.0	Ridge	0.11%

There are a total of 0.11% of outliers for both of the OLS and the ridge model. Apart from standardized residual plots, a box plot is able to provide a clear graphical representation by labeling outliers (Ramachandran & Tsokos, 2014). Hence, box plot of both models are observed as shown in Figure 11 and 12.

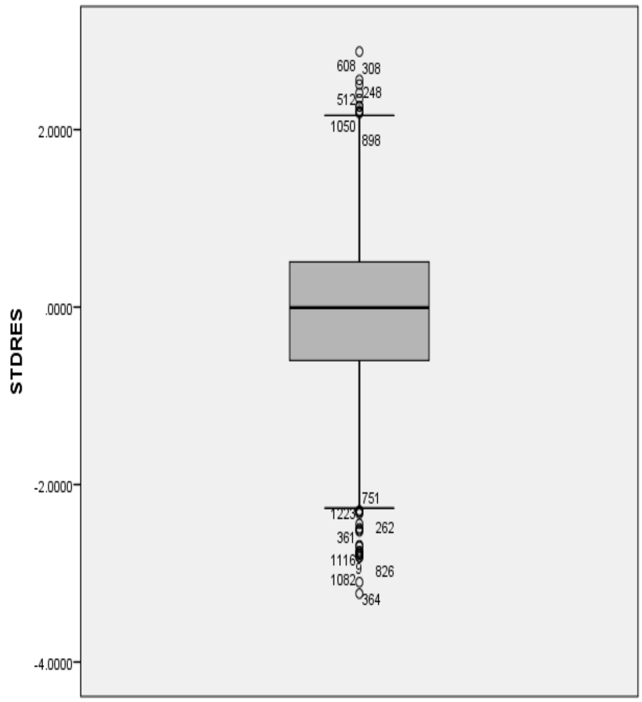

Figure 11. Box plot for OLS regression M78.7.3

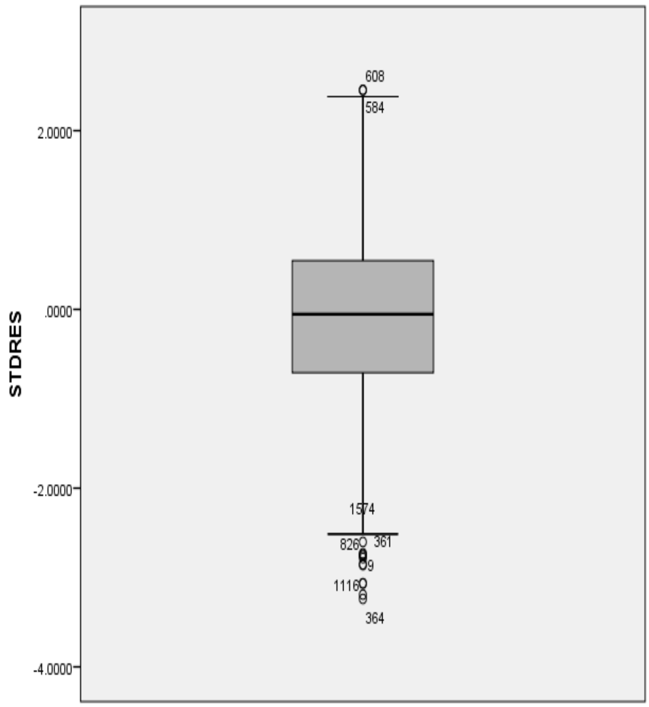

Figure 12. Box plot for Ridge regression M56.0.0

From Figure 11 and 12, the outliers in the dataset can be observed. There are more outliers for M78.7.3 as compared to M56.0.0. Deleting the outliers is not always the best option in the real life dataset. So, the results obtained from OLS cannot be trusted for a better forecast in the presence of outliers. On the other hand, ridge regression has the ability to deal in the presence of outliers (Steece, 1986). So, the ridge regression can be trusted to forecast the moisture content of the fish. Although the MAPE for OLS regression is less than the MAPE for the ridge regression, but due to the non- normality of the residuals and the presence of outliers, OLS cannot be trusted for a better forecast. On the other hand, ridge regression does not need any kind of normality assumptions.

CONCLUSIONS

In a nutshell, the best OLS model obtained to forecast the moisture content of fish was M78.7.3 with a total of 11 independent variables in this model after checking the multicollinearity and conducting a coefficient test. Furthermore, the best ridge model obtained to forecast the moisture content of fish was M56.0.0 with ridge parameter 0.002 and a total of 4 independent variables in this model after the conduct coefficient test. However, more outliers are detected for OLS model M78.7.3 as compared to the ridge model M56.0.0. The MAPE value of both of the models shows satisfying results. For OLS model M78.7.3, the MAPE value is 15.7342. The MAPE value for ridge model M56.0.0 is 17.4054. Due to non-normality of the residuals, and presence of outliers in the dataset, ridge regression is preferred for the best forecast with the MAPE of 17.4054. So, the moisture content of fish can forecast with the crucial factors as inlet temperature chamber, outlet temperature chamber, outlet humidity chamber and inlet humidity chamber. Since the MAPE is less than 20, so it will provide a good forecast. This paper only addressed multicollinearity and outliers by assuming no autocorrelated errors. We will consider autocorrelated errors in future study.

ACKNOWLEDGEMENT

The authors would like to extend their greatest gratitude to Universiti Sains Malaysia for funding this study under the Short term Grant Scheme (304.PMATHS.6315132).

REFERENCES

Abdullah, N., Jubok, Z. H., & Ahmed, A. (2011). Improved stem volume estimation using P-Value approach in polynomial regression models. *Research Journal of Forestry, 5*(2), 50-65.

Abdullah, N., Lee, C. L., & Jubok, Z. H. (2015). Factors on palm oil fruit bunches production volume for biomass fuel and biofuel during cogeneration processes. *Journal of the Japan Institute of Energy, 94*(12), 1428-1439.

Akaike, H. (1969). Fitting autoregressive models for prediction. *Annals of the Institute of Statistical Mathematics, 21*(1), 243-247.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*(6), 716-723.

Alfiya, P., Murali, S., Delfiya, D. A., & Samuel, M. P. (2018). Empirical modelling of drying characteristics of elongate glassy perchlet (Chanda nama)(Hamilton, 1822) in solar hybrid dryer. *Fishery Technology, 55*(2), 138-142.

Ali, M. K. M., Fudholi, A., Muthuvalu, M., Sulaiman, J., & Yasir, S. M. (2017a, December 4-7). Implications of drying temperature and humidity on the drying kinetics of seaweed. In *Proceedings of the 13th IMT-GT International Conference on Mathematics, Statistics and their Applications (ICMSA2017)*. Kedah, Malaysia.

Ali, M. K. M., Fudholi, A., Muthuvalu, M., Sulaiman, J., Yasir, S. M., & Hurtado, A. Q. (2017b). Post-harvest handling of eucheumatoid seaweeds. In *Tropical seaweed farming trends, problems and opportunities* (pp. 131-145). Cham, Switzerland: Springer.

Bodirsky, B. L., Rolinski, S., Biewald, A., Weindl, I., Popp, A., & Lotze-Campen, H. (2015). Global food demand scenarios for the 21st century. *PLoS One, 10*(11), 1-27.

Chatterjee, S., & Hadi, A. S. (2015). *Regression analysis by example*. Hoboken, New Jersey: John Wiley & Sons.

Delaney, N. J., & Chatterjee, S. (1986). Use of the bootstrap and cross-validation in ridge regression. *Journal of Business and Economic Statistics, 4*(2), 255-262.

Ertekin, C., & Yaldiz, O. (2004). Drying of eggplant and selection of a suitable thin layer drying model. *Journal of Food Engineering, 63*(3), 349-359.

FAO. (1996). *The state of food and agriculture 1996*. Rome, Italy: Food & Agriculture Org.

Golub, G. H., Heath, M., & Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics, 21*(2), 215-223.

Guan, Z., Wang, X., Li, M., & Jiang, X. (2013). Mathematical modeling on hot air drying of thin layer fresh tilapia fillets. *Polish Journal of Food and Nutrition Sciences, 63*(1), 25-33.

Gujarati, D. N. (2004). *Basic econometrics* (4th Ed.). New York, USA: The McGraw-Hill Companies.

Hajijubok, Z., & Gopal , P. K. (2008). Procedure in getting best model using multiple regression. *Journal of Borneo Science, 23*, 47-63.

Hannan, E. J., & Quinn, B. G. (1979). The determination of the order of an autoregression. *Journal of the Royal Statistical Society: Series B (Methodological), 41*(2), 190-195.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics, 12*(1), 55-67.

Hossain, M., & Bala, B. (2007). Drying of hot chilli using solar tunnel drier. *Solar Energy, 81*(1), 85-92.

Jamal, N., & Rind, M. Q. (2007). Ridge regression: A tool to forecast wheat area and production. *Pakistan Journal of Statistics and Operation Research, 3*(2), 125-134.

Javaid, A., Ismail, M., & Ali, M. K. M. (2020). Efficient model selection of collector efficiency in solar dryer using hybrid of LASSO and robust regression. *Pertanika Journal of Science and Technology, 28*(1), 193-210.

Javaid, A., Ismail, M. T., & Ali, M. K. M. (2019a). Model selection for collector efficiency of seaweed drier by using LASSO and multiple regression analysis using 8sc. In *Proceedings of the International Conference on Mathematical Sciences and Technology 2018 (MATHTECH2018)* (pp. 1-9). New York, NY: AIP Publishing LLC.

Javaid, A., Muthuvalu, M. S., Sulaiman, J., Ismail, M. T., & Ali, M. K. M. (2019b). Forecast the moisture ratio removal during seaweed drying process using solar drier. In *Proceedings of the International Conference on Mathematical Sciences and Technology 2018 (MATHTECH2018)* (pp. 1-8). New York, NY: AIP Publishing LLC.

Kennard, R. W. (1971). A note on the Cp statistic. *Technometrics, 13*(4), 899-900.

Khalaf, G. (2012). A proposed ridge parameter to improve the least square estimator. *Journal of Modern Applied Statistical Methods, 11*(2), 443-449.

Kituu, G. M., Shitanda, D., Kanali, C., Mailutha, J., Njoroge, C., Wainaina, J., & Silayo, V. (2010). Thin layer drying model for simulating the drying of Tilapia fish (Oreochromis niloticus) in a solar tunnel dryer. *Journal of Food Engineering, 98*(3), 325-331.

Krokida, M. K., Karathanos, V., Maroulis, Z., & Marinos-Kouris, D. (2003). Drying kinetics of some vegetables. *Journal of Food Engineering, 59*(4), 391-403.

Mahajan, V., Jain, A. K., & Bergier, M. (1977). Parameter estimation in marketing models in the presence of multicollinearity: An application of ridge regression. *Journal of Marketing Research, 14*(4), 586-591.

Rajarathinam, A., & Vinoth, B. (2014). Outlier detection in simple linear regression models and robust regression–A case study on wheat production data. *International Journal of Scientific Research, 3*(2), 531-536.

Ramachandran, K. M., & Tsokos, C. P. (2014). *Mathematical statistics with applications in R* (2nd Ed.). Oxford, UK: Elsevier.

Ramanatam, R. (2002). *Introductory econometrics with application* (5th Ed.). South Western, USA: Harcourt College Publishers.

Rice, J. (1984). Bandwidth choice for nonparametric regression. *The Annals of Statistics, 12*(4), 1215-1230.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*(2), 461-464.

Shibata, R. (1981). An optimal selection of regression variables. *Biometrika, 68*(1), 45-54.

Silva, B. G., Fileti, A. M. F., Foglio, M. A., Rosa, P. D. T. V., & Taranto, O. P. (2017). Effects of different drying conditions on key quality parameters of pink peppercorns (*Schinus terebinthifolius* Raddi). *Journal of Food Quality, 2017*, 1-12.

Steece, B. M. (1986). Regressor space outliers in ridge regression. *Journal of Communications in Statistics*, *15*(12), 3599-3605.

Stiling, J., Li, S., Stroeve, P., Thompson, J., Mjawa, B., Kornbluth, K., & Barrett, D. M. (2012). Performance evaluation of an enhanced fruit solar dryer using concentrating panels. *Energy for Sustainable Development, 16*(2), 224-230.

Tiwari, A. (2016). A review on solar drying of agricultural produce. *Journal of Food Processing and Technology, 7*(9), 1-12.

Ullah, M. I., Aslam, M., & Altaf, S. (2018). lmridge: A comprehensive R package for ridge regression. *The R Journal, 10*(2), 326-346.

Wen, Y. W., Tsai, Y. W., Wu, D. B. C., & Chen, P. F. (2013). The impact of outliers on net-benefit regression model in cost-effectiveness analysis. *PLoS One, 8*(6), 1-9.

Yahaya, A. H., Abdullah, N., & Zainodin, H. (2012). Multiple regression models up to first-order interaction on hydrochemistry properties. *Asian Journal of Mathematics and Statistics, 5*(4), 121-131.

Zhang, J., & Ibrahim, M. (2005). A simulation study on SPSS ridge regression and ordinary least squares regression procedures for multicollinearity data. *Journal of Applied Statistics, 32*(6), 571-588.