Data Screening Technique for Developing a Crude Oil Price Forecasting Models for Petroleum Industry

Abdul Talib Bon
Faculty of Technology Management and Business
Universiti Tun Hussein Onn Malaysia
Parit Raja, Batu Pahat
86400, Johor, Malaysia
talib@uthm.edu.my

Nuhu Isah
Faculty of Technology Management and Business
Universiti Tun Hussein Onn Malaysia
Parit Raja, Batu Pahat
86400, Johor, Malaysia
nuhuisah33g@gmail.com

Abstract

Forecasting studies of crude oil price depends heavily on historical data. These data should be stationary, consistent, and homogeneous when they are used for frequency analyses or to simulate a model. To determine whether the data meet these criteria, the researcher needs a simple but efficient screening procedure. Such a procedure is described in this paper. A time series of crude oil price data is volatile if its statistical properties are unaffected by choice of time origin. The basic data-screening procedure presented in this research is based upon split-record tests for stability of the variance and mean of such a time series. Although the stability of these two properties indicates only a weak form of volatility, this is enough to identify a non-volatile time series or to select those parts of a time series that are acceptable for use. We employed statistical package for social science to simulate the stated screening procedure to find out how good the data is for further forecasting procedures. It was found out that, after employing all the techniques, the data is found to be rich for forecasting of crude oil price for petroleum industry in Malaysia.

Keywords
Data Screening, Crude Oil, Price, petroleum industry, Malaysia.

1. Introduction
A time series of crude oil price data may exhibit jumps and trends owing to what call inconsistency and non-homogeneity. Inconsistency is a change in the amount of systematic error associated with the recording of data. It can arise from the use of different instruments and methods of observation. Non-homogeneity is a change in the statistical properties of the time series. Its causes can be either natural or man-made. These include alterations to economic activities (De Vries et al., 2016). The tests for stability of variance and mean verify not only the volatility of a time series but also its consistency and homogeneity. In the basic data-screening procedure, these two tests are reinforced by a third one, for the absence of a trend. Because all three tests are performed on individual time series that are not compared with similar series, their results indicate the presence (or absence) of absolute consistency and homogeneity (Franklin et al., 2017).

We applied the basic data-screening procedure to time series of crude oil price data that are summated over a year; we assume that, if the data are acceptable at this level of aggregation, they will be equally acceptable at lower levels that cover, say, a month or a day. Nevertheless, the independence (and acceptability) of a time series depends on both the level of aggregation and the separation in time of the data points. Of these two, separation in time is the easier to verify. For example, separation in time of the data points, so that successive crude oil price events are not associated with the related economic situation, is an obvious prerequisite to a successful frequency analysis, at whatever level of aggregation (Cassim et al., 2017).

A plot of progressive departures from the mean can help the researcher to pinpoint moments of change more accurately. Accordingly, we compute these departures and interpret the resulting plot. After ascertaining the absolute consistency and volatility of the data series, one can use double-mass analysis to test its relative consistency and volatility (Kim et al., 2015). The basic data-screening procedure, when applied to a time series of proportionality factors, before and after a suspected breakpoint in a double-mass line, is a good alternative to analysis of variance. We present a complete equation to perform all the computations.

2. Data Screening Procedure

The data screening procedure consists of four principal steps. These are:

- Do a rough screening of the data and compute or verify the totals for the crude oil price time series.
- Plot these totals according to the chosen time step (e.g. days, month, year) and note any trends or discontinuities.
- Test the time series for the absence of trend with Spearman's rank-correlation method.
- Apply the F-test for the stability of variance and the t-test for stability of mean to split, non-overlapping, sub-sets of the time series.

These steps form what we call the 'basic procedure'. If necessary, one can expand the basic procedure to include two additional steps. These are:

- Test the time series for the absence of persistence by computing the first serial-correlation coefficient.
- Test the time series for relative consistency and homogeneity with double-mass analysis.

Together, the two sets of steps form the complete data-screening procedure, which is illustrated in the flowchart below.

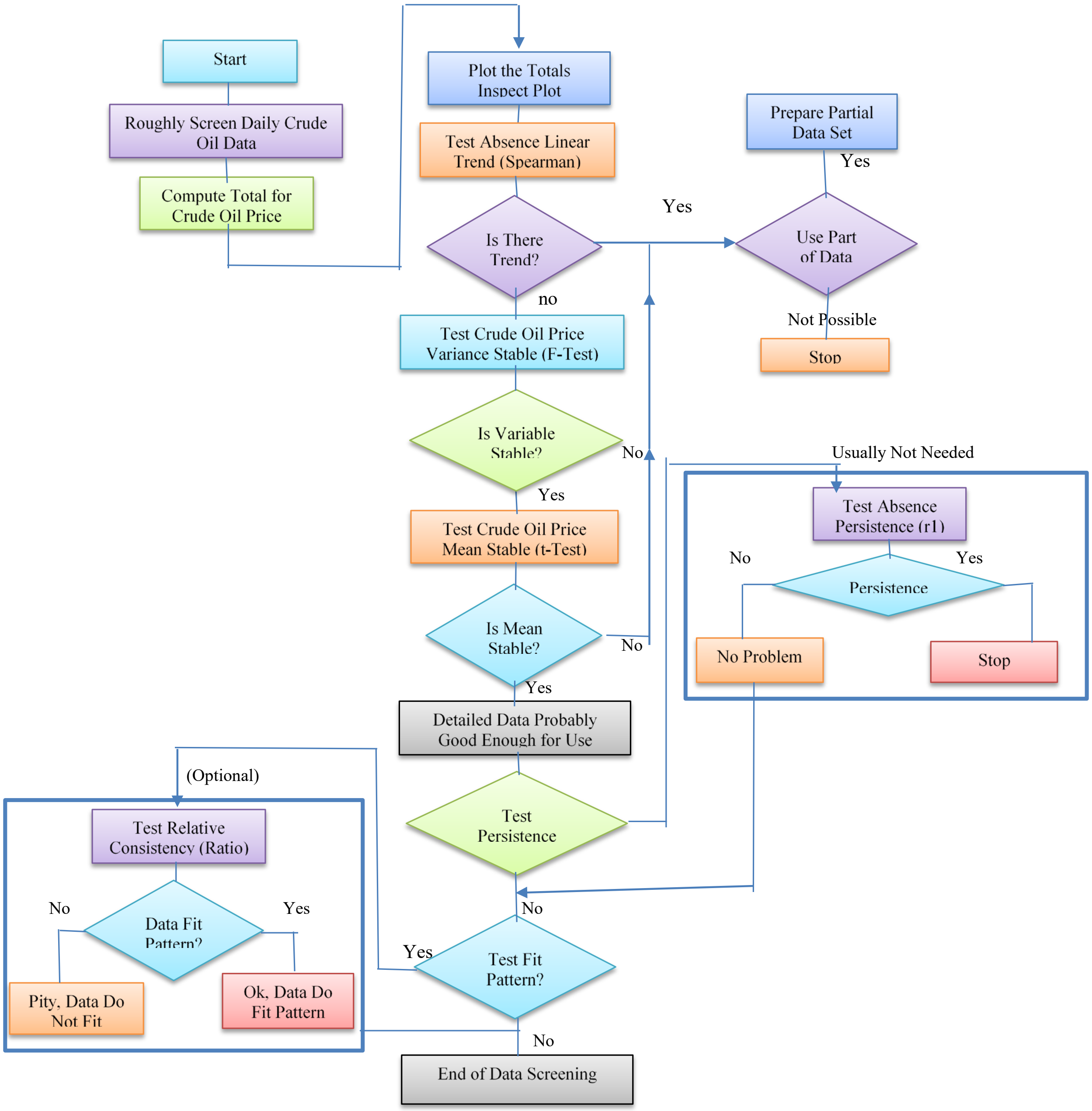

Figure 1: Data Screening Flow-Chart

3. The Procedure for Data Screening

3.1 Rough Screening of the Data

The basic procedure begins with an initial, rough screening of the data. For crude oil price totals, we tabulate daily crude oil price from petroleum industry Malaysia for 100 consecutive days. This will allow visual detection of whether the observations have been consistently or accidentally credited to the wrong day, whether they show gross errors (e.g. from weekly readings instead of daily ones), or whether they contain misplaced decimal points (Kondili et al., 2018). In most cases, it is convenient and perfectly acceptable to use daily, monthly or yearly totals as long as by 'days, months or years' means 'crude oil price by days, by months or by years'.

3.2 Plotting the Data

After doing a rough screening of the data, the researcher plots them on arithmetic or semi-logarithmic paper graph. The figure below shows a time series of a daily crude oil price totals from petroleum industry in Malaysia from January to May 2017 making 100 days. Note that it does not show any obvious trends or discontinuities.

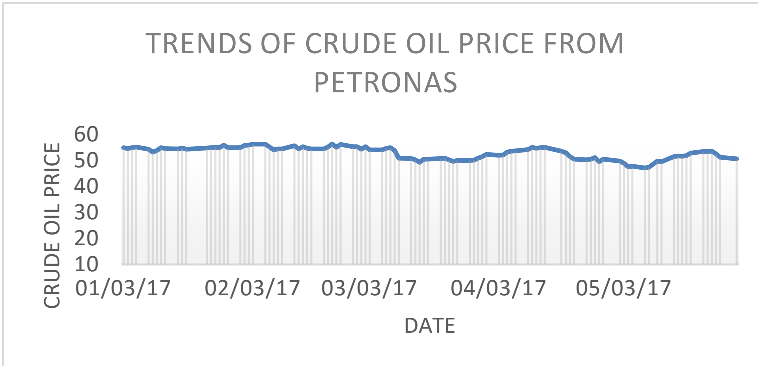

Figure 2: Actual Movement of Crude Oil Price of PETROLEUM INDUSTRY Malaysia from January to May 2017

3.3 Spearman's Rank Correlation Method

After plotting a time series, the researcher makes sure that there is no correlation between the order in which the data have been collected and the increase (or decrease) in the magnitude of those data. To verify the absence of a trend, the researcher employed Spearman's rank-correlation method as stated in the flowchart. The method is based on the Spearman rank-correlation coefficient, R_{sp} which is defined as:

$$R_{sp} = 1 - \frac{6 * \sum_{i-1}^{n} D_i * D_i}{n * (n * n - 1)}$$

where n is the total number of data, D is the difference, and i is the chronological order number, the difference between rankings is computed with:

$$D_i = K_{Xi} - K_{Yi}$$

Where K_{Xi} is the rank of the variable, x, which is the chronological order number of the observations, the series of observations, Y_i, is transformed to its rank equivalent, K_{Yi}, by assigning the chronological order number of an observation in the original series to the corresponding order number in the ranked series, Y. If there are ties, i.e. two or more ranked observations, Y, with the same value, the convention is to take K_X as the average rank. One can test the null hypothesis, H_O:$R_{sp} = 0$ (there is no trend), against the alternate hypothesis, H_1, R_{sp}, $<> 0$ (there is a trend), with the test statistic:

$$t_t = R_{sp} \left[\frac{n-2}{1 - R_{sp} * R_{sp}} \right]^{0.5}$$

Where t_t has Student's t-distribution with $v = n - 2$ degrees of freedom. Student's t-distribution is symmetrical around $t = 0$. We used a table of the percentile points of the t-distribution for a significance level of 5 per cent (two-tailed). One should, therefore, keep in mind that $t\{v, p\} = -t\{v, 1-p\}$ when using an incomplete table in some textbooks. At a significance level of 5 per cent (two-tailed), the two-sided critical region, $U \ of \ t_t$ is bounded by:

$$\{-\infty, t\{v, 2.5\%\}\} \ U\{t\{v, 97.5\%\}, +\infty\}$$

And the null hypothesis is accepted if *t*, is not contained in the critical region. In other words, the time series has no trend if:

$$t\{v, 2.5\%\} < t_t < t\{v, 97.5\%\}$$

If the time series does have a trend, the data cannot be used for frequency analyses or modelling. Removal of the trend is justified only if the physical processes underlying it are fully understood, which is rarely the case (Witten et al., 2016).

However, correlation is a statistical technique used in this research to determine the fitness of data. It shows whether and how strongly pairs of variables are related. The main result of a correlation is called the correlation coefficient (or "r") that ranges from -1.0 to +1.0.

Table 1: Spearman Correlation Test for the Crude Oil Price Data
from January to May 2017
Source: (Research Analysis 2018)

Correlations				
			Date	COP
Spearman's rho	Date	Correlation Coefficient	1.000	-.076
		Sig. (2-tailed)	.	.290
		N	100	100
	COP	Correlation Coefficient	-.076	1.000
		Sig. (2-tailed)	.290	.
		N	100	100

The table above shows the result of data fitness of crude oil price using Spearman correlation at a strength (coefficient) of -0.076 at $p = 0.076$. This means the data is fit for further analysis.

Spearman's rho correlation coefficient is significant at $p < 0.01$and 0.05 levels (2-tailed). Values close to +1 or -1 reveal the variables are fit to use while values near 0 either + or – indicate weak fitness of the data. The coefficient values of daily crude oil prices from January to May 2017 is -0.076 which is not significant at $p < 0.076$ levels (2 tailed). Value of daily crude oil prices shows strong fitness for data analysis using Spearman correlation.

3.4 The F-Test For Stability of Variance

In addition to testing the time series for the absence of a trend, we test it for the stability of variance and mean. The test for stability of variance is done first. There are two reasons for this sequence: firstly, instability of the variance implies that the time series is not stationary and, thus, not suitable for further use; secondly, the test for stability of mean is much simpler if one can use a pooled estimate of the variances of the two sub-sets (DeSimone et al., 2015).

The test statistic is the ratio of the variances of two split, non-overlapping, sub-sets of the time series. The distribution of the variance-ratio of samples from a normal distribution is known as the F, or Fisher, distribution (Geuder et al., 2015). Even if the samples are not from a normal distribution, the F-test will give an acceptable indication of the stability of variance (Aliyu et al., 2014).
Thus, the test statistic reads:

$$F_t = \frac{\sigma^{\frac{1}{2}}}{\sigma^{\frac{2}{2}}} = \frac{S^{\frac{2}{1}}}{S^{\frac{2}{2}}}$$

Where S^2 is variance? Note that, to compute F_t it is irrelevant whether one uses the sample standard deviation, S or the population standard deviation, σ. The researcher presents two convenient formulae for computing the sample standard deviation,, namely:

$$S = \left[\frac{\sum_{i=1}^{n}\left(X_i^{\,2}\right) - \left(\sum_{i=1}^{n}\left(X_i\right)^{\frac{2}{n}}\right)}{n-1} \right]^{0.5}$$

And

$$S = \left[\frac{\sum_{i=1}^{n}\left(X_I^2\right) - n * \overline{X}^2}{n-1} \right]^{0.5}$$

where X_i is the observation, n is the total number of data in the sample, and $\overline{X}$ is the mean of the data, the null hypothesis for the test, H_0: $S_1^2 = S_2^2$ is the equality of the variances; the alternate hypothesis is H_l: $S_1^2 <> S_2^2$. The rejection region, U, is bounded by:

$$\{0, F\{V_1, V_2, 2.5\%\}\} U \{F\{V_1, V_2, 97.5\%\}, +\infty\}$$

Where $V_1 = n_1 - 1$ is the number of degrees of freedom for the numerator, $V_2 = n_2 - 1$ is the number of degrees of freedom for the denominator, and n_1 n_2 are the number of data in each sub-set. In other words, the variance of the time series is stable, and one can use the sample standard deviation, S as an estimate of the population standard deviation, σ, if:

$$F\{V_1, V_2, 2.5\%\} < F_t < F\{V_1, V_2, 97.5\%\}$$

The F-distribution is not symmetrical for V_1 and V_2. One should, therefore, enter tables properly, usually by taking V_1 horizontally and V_2 vertically. The table of the F-distribution F $\{V_1, V_2, P\}$ for the 5-per-cent level of significance (two-tailed) was used in this research. The table below shows the result of the F test (Analysis of Variance) to figure out the fitness of the crude oil price data that would be used in this research.

Table 2: F Test for the Crude Oil Price Data from January to May 2017
Source: (Research Analysis 2018)

ANOVA					
Date					
	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	9526362327301 226.000	83	5506567819249 2.630	3.054	.002
Within Groups	3967209446400 00.000	22	1803277021090 9.090		
Total	9923083271941 226.000	100			

3.5 The T-Test For Stability of Mean

The t-test for the stability of mean involves computing and then comparing the means of two or three non-overlapping sub-sets of the time series (the same subsets from the F-test for stability of variance). A suitable statistic for testing the null hypothesis, H_0: $\overline{X_1} = \overline{X}_2$, against the alternate hypothesis, H_1: $\overline{X}_1 <> \overline{X}_2$, is:

$$t_t = \frac{\overline{X}_1 - \overline{X}_2}{\left[\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 + 2} * \left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right]^{0.5}}$$

Where n is the number of data in the sub-set, $\overline{X}$ the mean of the sub-set, and S^2 its variance. The test statistic t is valid for small samples with unknown variances. These variances can, however, differ only because of sampling variability if the t-test is applied in this form. This means that the variances of the sub-sets should not differ statistically: hence the requirement that the time series must be tested for stability of variance before it is tested for stability of mean. In samples from a normal distribution, t_t has a Student t-distribution. The requirement for normality is much less stringent for the t-test than for the F-test. One can apply the t-test to data that belong to any frequency distribution, but the length of the sub-sets should be equal if the distribution is skewed. One can avoid problems from a possibly skewed, underlying distribution by making the lengths of the sub-sets equal, or approximately so. For t_t the two-sided critical region, U, is:

$$\{-\infty, t\{v, 2.5\%\}\} \text{ U } \{t\{v, 97.5\%\}, +\infty\}$$

With $v = n_1 - 1 + n_2 - 1$ degrees of freedom, i.e. the total number of data minus 2. If t_1 is not in the critical region, the null hypothesis, H_0: $\overline{X}_1 = \overline{X}_2$, is accepted instead of the alternate hypothesis, H_1: $\overline{X}_1 <> \overline{X}_2$. In other words, the mean of the time series is considered stable if:

$$t\{v, 2.5\%\} < t_1 < t\{v, 97.5\%\}$$

Table 3: One-Sample T-Test for the Crude Oil Price Data from January to May 2017
Source: (Research Analysis 2018)

One-Sample T-Test						
	Test Value = 0					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
COP	215.672	100	.000	52.04821	51.5723	52.5242

4. Test for Absence of Persistence
4.1 The Serial Correlation Coefficient

As we know that time series of daily, monthly and yearly are usually independent. The serial-correlation coefficient can help to verify the independence of a time series. If a time series is completely random, the population auto-correlation function will be zero for all lags other than zero, and the sample serial-correlation coefficients will deviate slightly from zero only because of sampling effects. For our purposes, it is usually sufficient to compute the lag 1 serial-correlation coefficient, i.e. the correlation between adjacent observations in a time series. Here, we define the lag 1 serial correlation coefficient, R_1, according to Box and Jenkins (1970). This read:

$$r_1 = \frac{\sum_{i=1}^{n-1}(X_i - \overline{X}) * (X_{i+1} - \overline{X})}{\sum_{i=1}^{n}(X_i - \overline{X})}$$

where X_i is an observation, X_{i+1} is the following observation, $\overline{X}$ is the mean of the time series, and n is the number of data.

After computing r_1, we can test the hypothesis H_0: r_1 = 0 (that there is no correlation between two consecutive observations) against the alternate hypothesis, H_1: , < > 0. Anderson (1942) defines the critical region, U, at the 5-per-cent level of significance as:

$$\left\{-1, \left(-1-1.96(n-2)^{0.5}\right)/(n-1)\right\} \; U \left\{\left(-1+1.96(N-2)^{0.5}\right)/(N-1), +1\right\}$$

Table 4: Serial Correlation for the Crude Oil Price Data from
January to May 2017
Source: (Research Analysis 2018)

Model Summary					
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.086[a]	.007	.002	3.37478	.060
a. Predictors: (Constant), Date					
b. Dependent Variable: COP					

5. Test Fit Pattern
5.1 Double-Mass Analysis

The double-mass analysis assumes a linear relation between time series of crude oil price data. Crude oil price data are usually linear. The term ‘double-mass curve’ is commonly used in the literature. The researcher used the term ‘double-mass line’ instead, to stress the assumed linear relation between the data sets. Non-linear relations fall outside the scope. A linear relation between two variables that include the pair x = O and y = O can be expressed as:

$$y=b*x$$
where b is a proportionality factor

If *y*, is the time series to be tested, *x*, the time series of the pattern, and *i* = *O*, ..., n (the number of data pairs and the index of the time steps), then the plot of *Y*, = *C*(*y*,) (the mass of *y*) against *XI* = *C*(*x*,) (the mass of *x*) will result in a broken line through the origin, with an average slope *b*,, = *YJX*,. The line passes through the origin because the sum of the data at time zero is zero for both *X* and *Y*. Defining the average slope as the slope of the line through the points *0,O* and *Y, X*, will give a good enough estimate of the true mean of the proportionality factors.

The plotted points will never fall exactly on the average line. If there is a trend away from the line during a certain period, then an opposite trend will necessarily materialize during a following period to realize the average slope for the whole period. Analyzing persistent trends away from the average slope, one sees that break points between two periods with apparently different slopes indicate the moment at which the linear relationship changes between the means of two parts of the time series (Saint-Maurice et al., 2014). This is a break that, if significant, indicates a real change. Double-mass analysis is used not only to verify the relative consistency of a time series, but also to find correction factors for errors and fill in gaps (Saidin, 2014). Furthermore, at its best, double-mass analysis preserves the mean and not the standard deviation of the time series, unless a proportional error has been made.

It is generally acknowledged that C.F. Merriam was the first to use double-mass analysis to test a time series for relative consistency (Guzman et al., 2014). In a paper published in 1937, Merriam compares two tests for relative consistency, namely the plotting of cumulative departures from the mean and the cumulative plotting of one time series against another, i.e. double-mass analysis (Shehu et al., 2014).

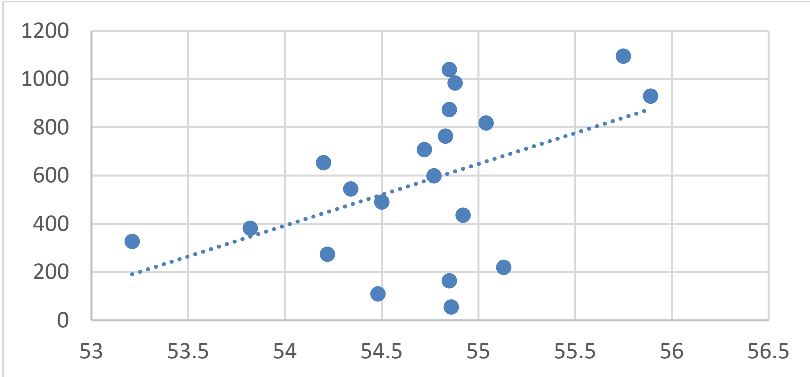

Figure 3: Cumulative Double-Mass Plot
Source: (Research Analysis 2018)

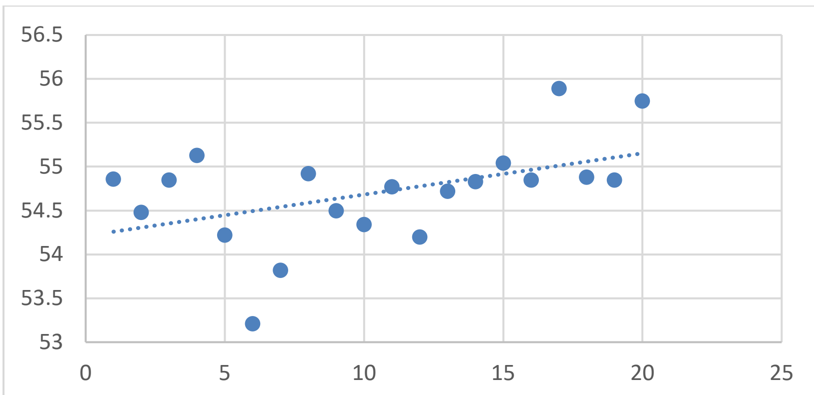

Figure 4: Simplified Double-Mass Plot
Source: (Research Analysis 2018)

We used the crude oil price data to generate the double-mass analysis figures above. The cumulative sums, *Xi* and *Yi,* are plotted against each other, and a line of best fit is drawn freehand. A simplification is to plot *Yi* against the year numbers, which is certainly permissible if successive values of *Xi* are not too different. Gorondutse et al., (2014) recommend plotting the cumulative differences between the sums at the against the corresponding values of the average line, i.e. plotting *Yi* - b, **Xi* against *Xi,* or against the time-step index (bottom plot).
In this plot of cumulative differences, also called a residual-mass plot, the maximum and minimum values correspond to break points in the original double-mass line. In our findings, an obvious and possibly significant break point is at 54.5 of *y*-axis to 600 of *x*-axis for cumulative double-mass plot, while it is slightly identified at 11 point of *y*-axis to 554.4 of *x*-axis.

Summary

This paper explained in detail the data screening process, it starts with correlations test, followed by F-test for the stability of variance, T-test for stability of mean, Test for absence of persistence, serial-correlation coefficient, test fit pattern as well as double mass analysis. After all the analysis and procedures followed based on the flowchart, it is found out that the data is fit to continue with the analysis.

Acknowledgement

This article is supported and sponsored by the office for research, innovation, commercialization and consultancy management, Universiti Tun Hussein Onn Malaysia under the GRANT GPPS, Vot No. U731. My appreciation goes to my supervisor for his guidance throughout this research paper.

References

De Vries, H., Wagelmans, A. P., Hasker, E., Lumbala, C., Lutumba, P., De Vlas, S. J., & van de Klundert, J. (2016). Forecasting human African trypanosomiasis prevalences from population screening data using continuous time models. *PLoS computational biology*, *12*(9), e1005103.

Franklin, J., Serra-Diaz, J. M., Syphard, A. D., & Regan, H. M. (2017). Big data for forecasting the impacts of global change on plant communities. *Global Ecology and Biogeography*, *26*(1), 6-17.

Cassim, N., Coetzee, L. M., Schnippel, K., & Glencross, D. K. (2017). Estimating the cost-per-result of a national reflexed Cryptococcal antigenaemia screening program: Forecasting the impact of potential HIV guideline changes and treatment goals. *PloS one*, *12*(8), e0182154.

Kim, T., Hong, J., & Kang, P. (2015). Box office forecasting using machine learning algorithms based on SNS data. *International Journal of Forecasting*, *31*(2), 364-390.

Kondili, L. A., Robbins, S., Blach, S., Gamkrelidze, I., Zignego, A. L., Brunetto, M. R., ... & Santantonio, T. A. (2018). Forecasting Hepatitis C liver disease burden on real life data. Does the hidden iceberg matter to reach the elimination goals? *Liver International*.

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

DeSimone, J. A., Harms, P. D., & DeSimone, A. J. (2015). Best practice recommendations for data screening. *Journal of Organizational Behavior*, *36*(2), 171-181.

Geuder, N., Wolfertstetter, F., Wilbert, S., Schüler, D., Affolter, R., Kraas, B., ... & Espinar, B. (2015). Screening and flagging of solar irradiation and ancillary meteorological data. *Energy Procedia*, *69*, 1989-1998.

Aliyu, A. A., Rosmain, T., & Takala, J. (2014). Online banking and customer service delivery in Malaysia: data screening and preliminary findings. *Procedia-Social and Behavioral Sciences*, *129*, 562-570.

Saint-Maurice, P. F., Welk, G. J., Bai, Y., & Allums-Featherston, K. (2014). Comparison of data screening methods for evaluating school-level fitness patterns in youth: findings from the NFL PLAY 60 FitnessGram Partnership Project. *Open Journal of Preventive Medicine*, *4*(11), 876.

Saidin, S. Z. (2014). Internal audit effectiveness: Data screening and preliminary analysis. *Asian Social Science*, *10*(10), 76.

Guzman, J. A., Chu, M. L., Starks, P. J., Moriasi, D. N., Steiner, J. L., Fiebrich, C. A., & McCombs, A. G. (2014). Upper Washita River experimental watersheds: Data screening procedure for data quality assurance. *Journal of environmental quality*, *43*(4), 1250-1261.

Shehu, A. M., & Mahmood, R. (2014). Market orientation, Knowledge management and Entrepreneurial orientation as predictors of SME performance: Data screening and Preliminary Analysis. In *Information and Knowledge Management* (Vol. 4, No. 7, pp. 12-23).

Gorondutse, A. H., & Hilman, H. (2014). Effect of business social responsibility (BSR) on performance of SMEs: Data screening and preliminary analysis. *Asian Social Science*, *10*(8), 103.

Biographies

Dr. Abdul Talib Bon is Professor of Technology Management in the Department of Production and Operations Management at the Universiti Tun Hussein Onn Malaysia. He has a PhD in Computer Science, which he obtained from the Universite de La Rochelle, France. His doctoral thesis was on topic Process Quality Improvement on Beltline Moulding Manufacturing. He studied Business Administration in the Universiti Kebangsaan Malaysia for which he was awarded the MBA. He's bachelor degree and diploma in Mechanical Engineering which his obtained from the Universiti Teknologi Malaysia. He received his postgraduate certificate in Mechatronics and Robotics from Carlisle, United Kingdom. He had published more 150 International Proceedings and International Journals and 8 books. His research interests include manufacturing, forecasting, simulation, optimization, TQM and Green Supply Chain. He is a member of IEOM, IIE, IIF, TAM, MIM and council member's of MSORSM.

Nuhu Isah is from Katsina state and was born in 1985 at Katsina state of Nigeria. He attended Musa Yar'adua Qur'anic Model Primary School, then to Sir Usman Nagogo College of Arabic and Islamic Studies and finished his secondary school at Kiddies International College Katsina, Katsina State of Nigeria. He also attended Ahmadu Bello University Zaria, Nigeria where he obtained B.Sc. Business Administration in 2011. He started working with Unity Bank Nigeria Plc from 2011 to 2013 where he worked as recruitment assistant officer in Human Capital Management Department and then transferred to branch as teller in Operations Department. Thereafter, he applied and later gains admission for Master's Degree in Technology Management at University Tun Hussein Onn Malaysia (UTHM). After his MSc in 2015, he immediately continues his PhD in the same university at the same faculty. The author published 1 book and more than 15 International Proceedings and Journals.