ISSN: 2278 – 7798
International Journal of Science, Engineering and Technology Research (IJSETR)
Volume 2, No 6, June 2013

Regression and Time Series Analysis of Petroleum Product Sales in Masters Energy oil and Gas

[1]Ezeliora Chukwuemeka Daniel
[1]Department of Industrial and Production Engineering, Nnamdi Azikiwe University Awka, Anambra State, Nigeria Mobile: 2348060480087,
[2]Chinwuko Emmanuel Chuka
[2]Department of Industrial and Production Engineering, Nnamdi Azikiwe University Awka, Anambra State, Nigeria Mobile: 2348037815808,
[3]Ogunoh Arinze Victor
[3]Department of Industrial and Production Engineering, Nnamdi Azikiwe University Awka, Anambra State, Nigeria Mobile: 2348030616723,
[4] Nwosu Moses C
[4]Department of Industrial and Production Engineering, Nnamdi Azikiwe University Awka, Anambra State, Nigeria Mobile: 2348068261242

Abstract

This research work focuses on the regression analysis and time series analysis of the Petroleum Product Sales in Masters Energy oil and Gas. The application of multiple regression model were made to show the effect of environmental factors in the petroleum product sales. In the regression analysis, the p-value shows how the environmental factors (that is the independent variables) affect the model and how important is any of the environmental factors. The time series analysis were also used to show the trend influence in the data, detrend influence in the data, seasonal variation in the data and deseasoning of the data. Furthermore, trend estimation model was also used to forecast the product sales of the data. In conclusion, it was observed that the environmental factors have little effect on the petroleum product sales.

Key words: Regression, Time Series, Trend, Seasonal index, Forecasting, Deseason, Detrend and Petroleum

1. INTRODUCTION

Forecasting is an integral part of a marketing manager's role. Sales forecasts are important for understanding market share and the competition, future production needs, and the determinants of sales, including promotions, pricing, advertising and distribution[2][5].

Market researchers and business analysts are often faced with the task of predicting sales. One approach to prediction is to use cross-sectional data, working with one point in time or aggregating over several time periods. We search for variables that relate to sales and use those variables as explanatory variables in our models. Another approach is to work with time series data, aggregating across sales territories or accounts. We use past and current sales as a predictor of future sales as we search for explanatory variables that relate to sales [3] [4].

Cross-sectional and simple time series approaches do not make full use of data available to sales and marketing managers. Typical sales data have a hierarchical structure. They are longitudinal or panel data, having both cross-sectional and time series characteristics. They are cross-sectional because they include observations from many cases, sales across stores, territories or accounts; we say that these data are differentiated across "space." They are time series data because they represent many points in time [6][7].

2. Time Series Analysis and Forecasting

A time series is a chronological sequence of observations on a particular variable. Usually the observations are taken at regular intervals (days, months, years), but the sampling could be irregular. Common examples of time series are the Dow Jones Industrial Average, Gross Domestic Product, unemployment rate, and airline passenger loads. A time series analysis consists of two steps: (1) building a model that represents a time series, and (2) using the model to predict (forecast) future values [8]

. If a time series has a regular pattern, then a value of the series should be a function of previous values. If Y is the target value that we are trying to model and predict, and Y_t is the value of Y at time t, then the goal is to create a model of the form:

$Y_t = f(Y_{t-1}, Y_{t-2}, Y_{t-3}, \ldots, Y_{t-n}) + e_t$

Where Y_{t-1} is the value of Y for the previous observation, Y_{t-2} is the value two observations ago, etc., and e_t represents noise that does not follow a predictable pattern (this is called a random shock). Values of variables occurring prior to the current observation are called lag values. If a time series follows a repeating pattern, then the value of Y_t is usually highly correlated with $Y_{t\text{-cycle}}$ where cycle is the number of observations in the regular cycle. For example, monthly observations with an annual cycle often can be modeled by

$Y_t = f(Y_{t-12})$

The goal of building a time series model is the same as the goal for other types of predictive models which is to create a model such that the error between the predicted value of the target variable and the actual value is as small as possible. The primary difference between time series models and other types of models is that lag values of the target variable are used as predictor variables, whereas traditional models use other variables as predictors, and the concept of a lag value doesn't apply because the observations don't represent a chronological sequence.

ARMA and modern types of models

Traditional time series analysis uses Box-Jenkins ARMA (Auto-Regressive Moving Average) models. An ARMA model predicts the value of the target variable as a linear function of lag values (this is the auto-regressive part) plus an effect from recent random shock values (this is the moving average part). While ARMA models are widely used, they are limited by the linear basis function.

In contrast to ARMA models, DTREG can create models for time series using neural networks, gene expression programs, support vector machines and other types of functions that can model nonlinear relationships. So, with a DTREG model, the function f(.) in

$Y_t = f(Y_{t-1}, Y_{t-2}, Y_{t-3}, \ldots, Y_{t-n}) + e_t$

can be a neural network, gene expression program or other type of general model. This makes it possible for DTREG to model time series that cannot be handled well by ARMA models.

Setting up a time series analysis

Input variables

When building a normal (not time series) model, the input must consist of values for one target variable and one or more predictor variables. When building a time series model, the input can consist of values for only a single variable – the target variable whose values are to be modeled and forecast. Here is an example of an input data set:

The time between observations must be constant (a day, month, year, etc.). If there are missing values, you must provide a row with a missing value indicator for the target variable like this:

For financial data like the DJIA where there are never any values for weekend days, it is not necessary to provide missing values for weekend days. However, if there is odd missing days such as holidays, then those days must be specified as missing values. It is also desirable to put in missing values for February 29 on non-leap years so that all years have 366 observations.

Lag variables

A lag variable has the value of some other variable as it occurred some number of periods earlier. For example, here is a set of values for a variable Y, its first lag and its second lag:

Notice that lag values for observations before the beginning of the series are unknown.

DTREG provides automatic generation of lag variables. On the Time Series Property page you can select which variables are to have lag variables generated and how far back the lag values are to run. You can also create variables for moving averages, linear trends and slopes of previous observations.

On the Variables property page, you can select which generated variables you want to use as predictors for the model. While it is tempting to generate lots of variables and use all of them in the model, sometimes better models can be generated using only lag values that are multiples of the series' cycle period. The autocorrelation table (see below) provides information that helps to determine how many lag values are needed. Moving average, trend and slope variables may detract from the model, so you should always try building a model using only lag variables.

Intervention variables

An exceptional event occurring during a time series is known as an intervention. Examples of interventions are a change in interest rates, a terrorist act or a labor strike. Such events perturb the time series in ways that cannot be explained by previous (lag) observations.

DTREG allows you to specify additional predictor variables other than the target variable. You could have a variable for the interest rate, the gross domestic product, inflation rate, etc. You also could provide a variable with values of 0 for all rows up to the start of a labor strike, then 1 for rows during a strike, then decreasing values following the end of a strike. These variables are called intervention variables; they are specified and used as ordinary predictor variables. DTREG can generate lag values for intervention variables just as for the target variable [8].

Trend removal and stationary time series

A time series is said to be stationary if both its mean (the value about which it is oscillating), and its variance (amplitude) remain constant through time. Classical Box-Jenkins ARMA models only work satisfactorily with stationary time series, so for those types of models it is essential to perform transformations on the series to make it stationary. The models developed by DTREG are less sensitive to non-stationary time series than ARMA models, but they usually benefit by making the series stationary before building the model. DTREG includes facilities for removing trends from time series and adjusting the amplitude.

4. Selecting the type of model for a time series

DTREG allows you to use the following types of models for time series:

1. Decision tree
2. TreeBoost (boosted series of decision trees)
3. Multilayer perceptron neural network
4. General regression neural network (GRNN)
5. RBF neural network
6. Cascade correlation network
7. Support vector machine (SVM)
8. Gene expression programming (GEP)

Experiments have shown that decision trees usually do not work well because they do a poor job of predicting continuous values. Gene expression programming (GEP) is an excellent method for time series because the functions generated are very general, and they can account for trends and variance changes. General regression neural networks (GRNN) also perform very well in tests. Multilayer perceptrons sometimes work very well, but they are more temperamental to train. So the best recommendation is to always try GEP and GRNN models, and then try other types of models if you have time. If you use a GEP model, it is best to enable the feature to allow it to evolve numeric constants.

5. Evaluating the forecasting accuracy of a model

Before you bet your life savings on the forecasts of a model, it is nice to do some tests to evaluate the predictive accuracy of the model. DTREG includes a built-in validation system that builds a model using the first observations in the series and then evaluates (validates) the model by comparing its forecast to the remaining observations at the end of the series.

Time series validation is enabled on the Time Series property page (see above).

Specify the number of observations at the end of the series that you want to use for the validation. DTREG will build a model using only the observations prior to these held-out observations. It will then use that model to forecast values for the observations that were held out, and it will produce a report and chart showing the quality of the forecast. Here is an example of a chart showing the actual values with black squares and the validation forecast values with open circles:

If you compare validation results from DTREG with other programs, you need to check how the other programs compute the predicted values. Some programs use actual (known) lag values when generating the predictions; this gives an unrealistically accurate prediction. DTREG uses the lag values for predicted values when forecasting: this makes validation operate like real forecasting where lag values must be based on predicted values rather than known values [1].

6. Patterns that may be present in a time series

Trend: Data exhibit a steady growth or decline over time.

Seasonality: Data exhibit upward and downward swings in a short to intermediate time frame (most notably during a year).

Cycles: Data exhibit upward and downward swings in over a very long time frame.

Random variations: Erratic and unpredictable variation in the data over time with no discernable pattern [1].

7. R**esearch Method Used**

The data were collected for a period of three years product sales. The data were modeled and analyzed using regression and time series analysis. The regression model shows the effect of the environmental factors in its model and in the forecasting results. However, time series analyses were also used to show the seasonal and deseasonal influence and also trend and detrend influence.

Table 1: Presentation of Petroleum product Sales Data in Masters Energy Oil and Gas

Year	Month	Time	Humidity (%)	Rainfall (mm)	Temp (C)	Petroleum
2010	Jan	1	28.4	76	11.2	132000
	Feb	2	30.0	80	0.7	99000
	Mar	3	30.7	76	50.6	132000
	April	4	29.2	81	121.1	66000
	May	5	28.3	81	233.1	132000
	June	6	27.6	85	248.3	90000
	July	7	26.7	89	386.6	66000
	Aug	8	26.7	87	283.3	219000
	Sept	9	26.9	86	249.4	138000
	Oct	10	26.9	85	376.0	231000
	Nov	11	27.8	73	96.3	132000
	Dec	12	28.8	76	Nil	99000
2011	Jan	13	29.0	72	0.2	6600
	Feb	14	30.4	72	66.8	143000
	Mar	15	30.8	73	13.6	99000
	April	16	29.7	79	203.5	231000
	May	17	28.7	81	161.3	99000
	June	18	27.7	84	191.9	132000
	July	19	27.0	84	13.2	99000
	Aug	20	27.2	86	299.9	232000
	Sept	21	27.0	87	306.4	99000
	Oct	22	27.3	85	166.9	68000
	Nov	23	28.6	82	412.0	132000
	Dec	24	28.3	71	Nil	165000
2012	Jan	25	28.4	76	11.2	132000
	Feb	26	30.0	80	0.7	198000
	Mar	27	30.7	76	50.6	240000
	April	28	29.2	81	121.1	99000
	May	29	28.3	81	233.1	240000
	June	30	27.6	85	248.3	66000
	July	31	26.7	89	386.6	132000
	Aug	32	26.7	87	283.3	198000
	Sept	33	26.9	86	249.4	231000
	Oct	34	26.9	85	376.0	99000
	Nov	35	27.8	73	96.3	0
	Dec	36	28.8	76	Nil	0

General Regression Analysis: PETROLEUM versus Temp (C), Humidity (%), ...

```
Regression Equation

PETROLEUM  =  -571441 + 16956.5 Temp (C) + 2385.16 Humidity (%) + 156.584
              Rainfall (mm) + 193.726 MONTH/TIME
```

Coefficients

Term	Coef	SE Coef	T	P	95% CI	VIF
Constant	-571441	566377	-1.00894	0.321	(-1726574, 583692)	
Temp (C)	16957	12970	1.30734	0.201	(-9497, 43410)	2.17436
Humidity (%)	2385	3784	0.63029	0.533	(-5333, 10103)	3.23077
Rainfall (mm)	157	139	1.12638	0.269	(-127, 440)	2.82573
MONTH/TIME	194	1105	0.17537	0.862	(-2059, 2447)	1.05142

Summary of Model

S = 67147.5 R-Sq = 11.71% R-Sq(adj) = 0.32%

Analysis of Variance

Source	DF	Seq SS	Adj SS	Adj MS	F	P
Regression	4	1.85408E+10	1.85408E+10	4635198641	1.02804	0.408355
Temp (C)	1	1.03288E+08	7.70611E+09	7706108710	1.70913	0.200712
Humidity (%)	1	1.24399E+10	1.79121E+09	1791210280	0.39727	0.533121
Rainfall (mm)	1	5.85892E+09	5.72043E+09	5720433547	1.26873	0.268649
MONTH/TIME	1	1.38672E+08	1.38672E+08	138671792	0.03076	0.861927
Error	31	1.39772E+11	1.39772E+11	4508788527		
Total	35	1.58313E+11				

Fits and Diagnostics for All Observations

Obs	PETROLEUM	Fit	SE Fit	Residual	St Resid
1	132000	93345	27890.6	38655	0.63285
2	99000	128565	30952.0	-29565	-0.49616
3	132000	138901	28143.7	-6901	-0.11320
4	66000	136625	21632.1	-70625	-1.11103
5	132000	139096	20979.6	-7096	-0.11124
6	90000	139340	20338.7	-49340	-0.77103
7	66000	155469	26615.6	-89469	-1.45131
8	219000	134718	23157.6	84282	1.33722
9	138000	130609	21329.4	7391	0.11608
10	231000	148241	25895.8	82759	1.33582
11	132000	91278	32627.5	40722	0.69388
12	99000	100504	19655.3	-1504	-0.02343
13	6600	94580	24284.8	-87980	-1.40538
14	143000	128941	26356.5	14059	0.22764
15	99000	129973	25256.9	-30973	-0.49782
16	231000	155560	22781.1	75440	1.19433
17	99000	136960	12691.3	-37960	-0.57570
18	132000	132144	14275.7	-144	-0.00220
19	99000	92487	36004.3	6513	0.11491
20	232000	145735	16288.2	86265	1.32426
21	99000	145940	17603.2	-46940	-0.72440
22	68000	124607	19525.6	-56607	-0.88110
23	132000	178068	35139.4	-46068	-0.80511
24	165000	82425	30986.0	82575	1.38617
25	132000	97994	21061.3	34006	0.53335
26	198000	133215	31145.8	64785	1.08906
27	240000	143551	28983.1	96449	1.59235
28	99000	141275	21518.6	-42275	-0.66463
29	240000	143745	18165.9	96255	1.48901
30	66000	143990	18816.2	-77990	-1.20995
31	132000	160119	24831.4	-28119	-0.45071
32	198000	139367	22561.5	58633	0.92709
33	231000	135259	22484.5	95741	1.51319
34	99000	152891	26433.1	-53891	-0.87307
35	0	95927	33730.3	-95927	-1.65218
36	0	105154	27403.2	-105154	-1.71536

Table 2: Forecasting results for Petroleum product Sales using Regression model

Year	Month	Temp (C)	Humidity (%)	Rainfall (mm)	Time	Petroleum
2013	Jan	28.8	76	0	37	105346.2
	Feb	28.8	76	0	38	105539.9
	Mar	28.8	76	0	39	105733.7
	April	28.8	76	0	40	105927.4
	May	28.8	76	0	41	106121.1
	June	28.8	76	0	42	106314.9
	July	28.8	76	0	43	106508.6
	Aug	28.8	76	0	44	106702.3
	Sept	28.8	76	0	45	106896
	Oct	28.8	76	0	46	107089.8
	Nov	28.8	76	0	47	107283.5
	Dec	28.8	76	0	48	107477.2
2014	Jan	28.8	76	0	49	107670.9
	Feb	28.8	76	0	50	107864.7
	Mar	28.8	76	0	51	108058.4
	April	28.8	76	0	52	108252.1
	May	28.8	76	0	53	108445.8
	June	28.8	76	0	54	108639.6
	July	28.8	76	0	55	108833.3
	Aug	28.8	76	0	56	109027
	Sept	28.8	76	0	57	109220.7
	Oct	28.8	76	0	58	109414.5
	Nov	28.8	76	0	59	109608.2
	Dec	28.8	76	0	60	109801.9
2015	Jan	28.8	76	0	61	109995.6
	Feb	28.8	76	0	62	110189.4
	Mar	28.8	76	0	63	110383.1
	April	28.8	76	0	64	110576.8
	May	28.8	76	0	65	110770.6
	June	28.8	76	0	66	110964.3
	July	28.8	76	0	67	111158
	Aug	28.8	76	0	68	111351.7
	Sept	28.8	76	0	69	111545.5
	Oct	28.8	76	0	70	111739.2
	Nov	28.8	76	0	71	111932.9
	Dec	28.8	76	0	72	112126.6

Table 3: Forecasting Results of Petroleum Product Sales

year	month	time	Regression
2013	Jan	37	105346.2
	Feb	38	105539.9
	Mar	39	105733.7
	April	40	105927.4
	May	41	106121.1
	June	42	106314.9
	July	43	106508.6
	Aug	44	106702.3
	Sept	45	106896
	Oct	46	107089.8
	Nov	47	107283.5
	Dec	48	107477.2
2014	Jan	49	107670.9
	Feb	50	107864.7
	Mar	51	108058.4
	April	52	108252.1
	May	53	108445.8
	June	54	108639.6
	July	55	108833.3
	Aug	56	109027
	Sept	57	109220.7
	Oct	58	109414.5
	Nov	59	109608.2
	Dec	60	109801.9
2015	Jan	61	109995.6
	Feb	62	110189.4
	Mar	63	110383.1
	April	64	110576.8
	May	65	110770.6
	June	66	110964.3
	July	67	111158
	Aug	68	111351.7
	Sept	69	111545.5
	Oct	70	111739.2
	Nov	71	111932.9
	Dec	72	112126.6

Time Series Analysis for PETROLEUM

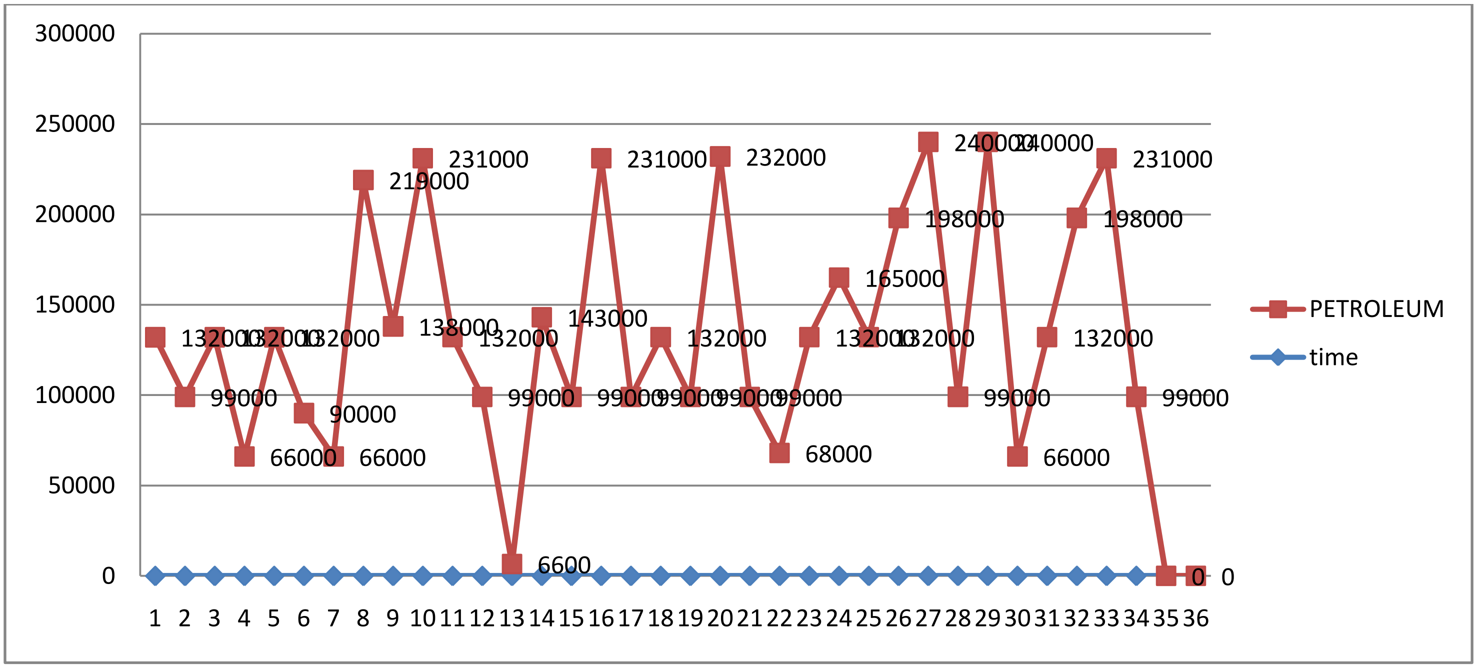

Figure 1: Time Series Analysis of Petroleum

Time Series Decomposition for PETROLEUM

```
Data      PETROLEUM
Length    36
NMissing  0

Fitted Trend Equation
Yt = 137332 - 168.076*t

Seasonal Indices

Period    Index
     1  0.45525
     2  1.15778
     3  1.11985
     4  1.18814
     5  1.15233
     6  0.74764
     7  0.63005
     8  1.70909
     9  0.88925
    10  1.12202
    11  0.92729
    12  0.90131

Accuracy Measures
MAPE           59
MAD         48843
MSD    3496371372
```

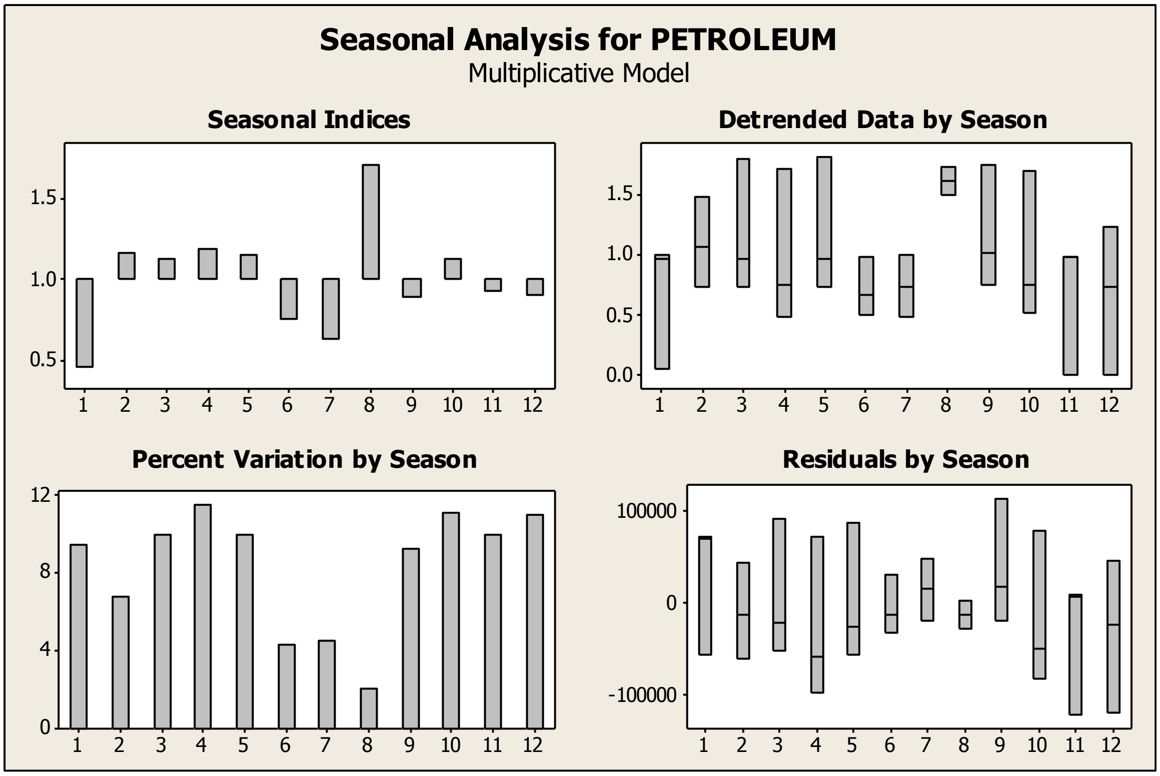

Figure 2: Decomposition - Seasonal Analysis for PETROLEUM

Figure 2 shows a seasonal analysis of the data for the petroleum product sales. The chart also shows the movement of the data periodically. From the seasonal indices plot, it shows that the influence of the season was more within the month of August, this is due to high increase in product sales and also slightly influence within the month of February over the period of three years data. This is due to little or no product sales during those seasons. However, detrend data by season was the plot of the seasonal influence of the absolute changes in values and removal of the effects of accumulating data sets from a trend by season. It was also observed that the seasonal influence of detrend data were high in the month of September and low in the months of January, November and December. Furthermore, the percent variation of detrend data by season were also plotted to show the percent variation of the data seasonally and it was noted that it was high in the month of April with about 12% variation. The residual plot by season was the error plot by season and was also high in the month of September.

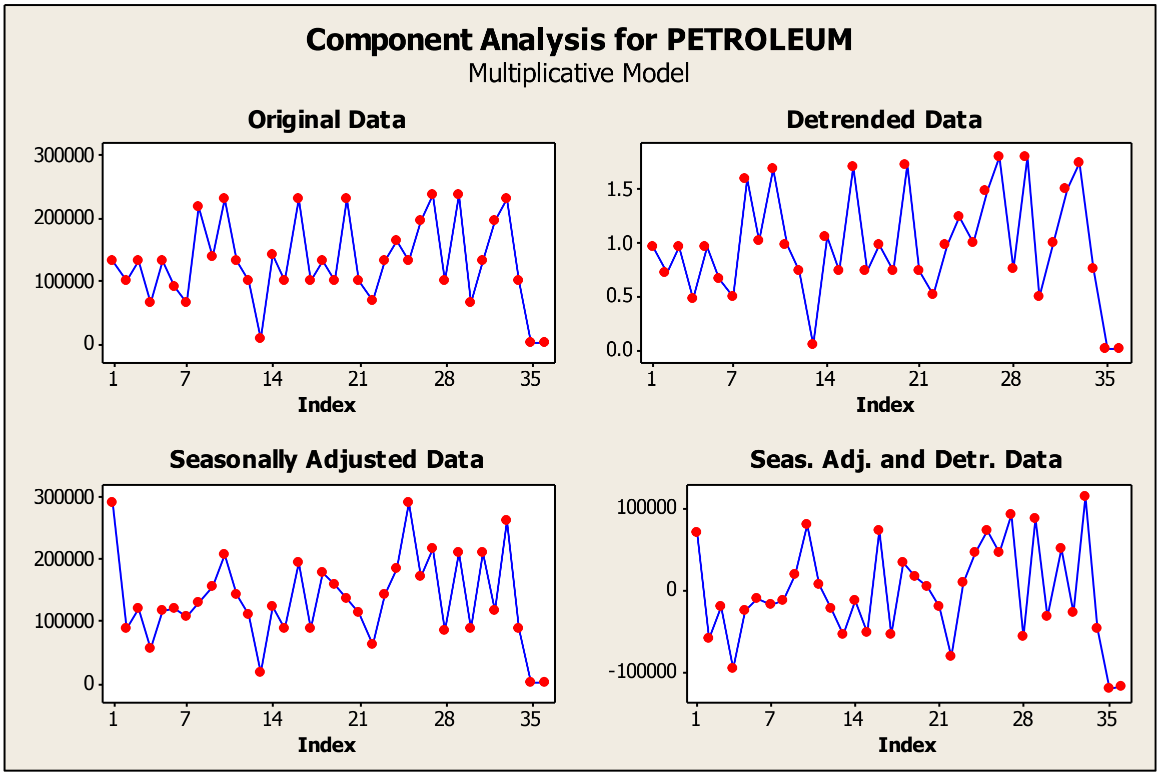

Figure 3: Decomposition - Component Analysis for PETROLEUM

The decomposition component analysis plot was used to show the time series data plot of the values for the 36 months period of time. The detrend data plot was the plot of the data after the removal of the effects of accumulating data sets from a trend and to show only the absolute changes in values to allow potential cyclical patterns to be identified. Seasonal adjusted data plot was the plot of the data after the removal of the seasonal effects on the data. Seasonal adjusted and detrend data plot were used to show the data plot when they have already removed the seasonal effects and also remove the effects of accumulating data sets from a trend, to show only the absolute changes in values.

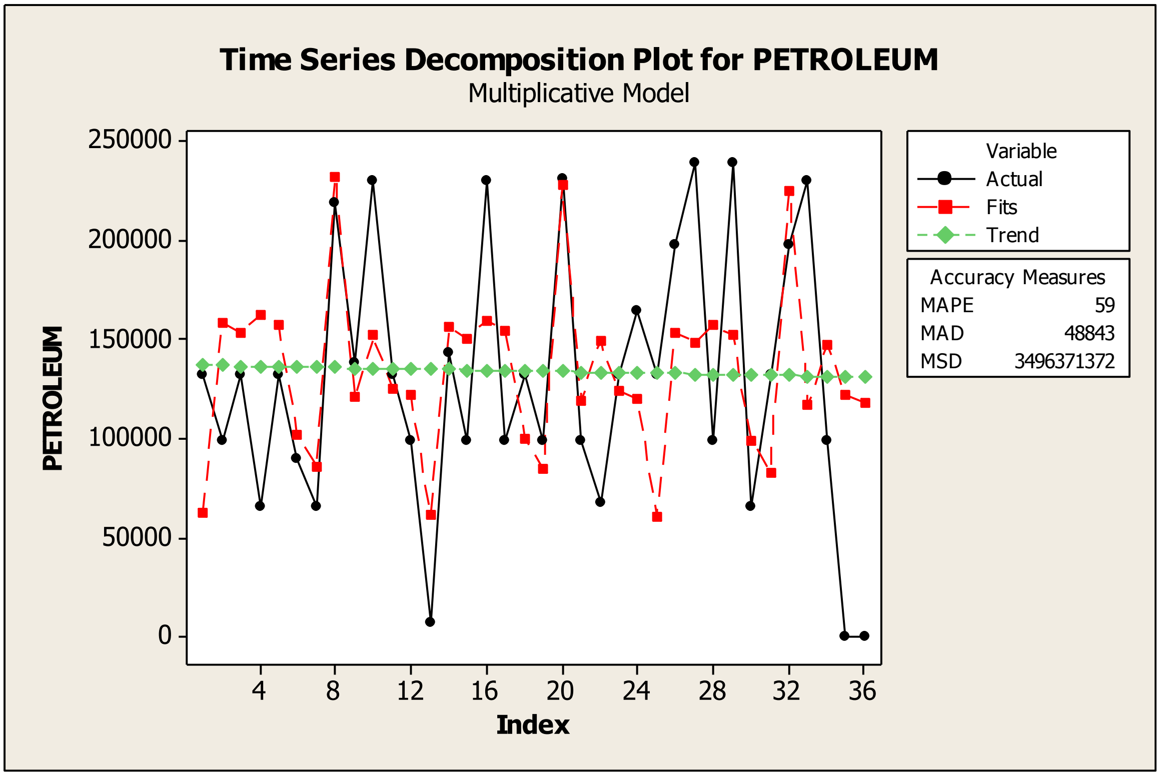

Figure 4: Time Series Decomposition Plot for PETROLEUM

Figure 4 shows a trend forecasting analysis of the data for the petroleum product sales. The chart also shows the movement of the data periodically. The trend line shows that the data contains seasonal influence and trend influence that is responsible for slightly reducing the yield by value at the trend component per month. The accuracy measures of the data were evaluated as shown. The mean absolute deviation is the measure of the average of the forecasting errors in the data and this was shown in the chart as 48843. It expresses accuracy in the same units as the data. However, the mean absolute percent error measures the percentage average of the sum of the forecasting errors divided by the sum of the values of the data and this is 59 percent in the chart. Furthermore, the mean squared deviation is the average squared deviation of the forecasting error in the data.

In conclusion, it shows that time series analysis explain the analysis of trend and seasonal influence on the data. It was also used to forecast the future product sales of the case study company. While the regression analysis was used to show the effect of the external (i.e. independent variables) factors in the forecast.

Reference

1. Abraham B. & Ledolter J., 1983, *Statistical Methods for Forecasting*, John Wiley & Sons
2. Ailawadi, K.L. and S.A. Neslin (1998). "The effect of promotion on consumption: Buying more and consuming it faster." *Journal of Marketing Research* 35, 390-398.
3. Battese, G.E., R.M. Harter and W.A. Fuller (1988). "An error components model for prediction of county crop areas using survey and satellite data." *Journal of the American Statistical Association* 83, 28-36.
4. Baltagi, B.H. and Q. Li (1992). "Prediction in the one-way error component model with serial correlation." *Journal of Forecasting* 11, 561-567.
5. Erdem, T. (1996). "A dynamic analysis of market structure based on panel data." *Marketing Science* 15, 359-378.
6. Kackar, R. N. and D. Harville (1984). "Approximations for standard errors of estimators of fixed and random effects in mixed linear models." *Journal of the American Statistical Association* 79, 853-862.
7. Verbeke, G. and G. Molenbergs (2000). *Linear mixed models for longitudinal data*. Springer-Verlag,New York.
8. Yaffee, R., 2000, *Time Series Analysis and Forecasting*, Academic Press