CAR SALES PREDICTION ANALYSIS WITH COMPUTER BASED THETA METHOD

Jansen Dharma Putra S. Kom [1] dan Djunaidy Santoso, Dipl.Ing.,M.Kom.[2]
[1]Alumni Program Studi Teknik Informatika
[2] Staff Pengajar Program Studi Teknik Informatika
Department of Computer Science Binus University
djunaidyS0533@binus.ac.id

ABSTRACT

The economic development in Indonesia is growing rapidly, competition is increasing. One of the things that are important in supporting the success of a company is to be able to meet the demands of various consumers. Activities that enable us to predict what will happen in the future are known as forecasting. Forecasting Theta methods, by Assimakopoulos Theta and Nikolopoulos. Initially, forecasting only use 2 theta weights parameters, then another theta parameter is added. Currently the Theta method use theta weight parameter. This research is to perform sales forecasting on car sales using Theta Method with five different weights using a system specially designed to perform the calculation. The error is minimized by minimizing the Mean Squared Error (MSE). As a result, using Theta weight 3 produces the smallest error.

***Keywords**: Theta method, forecasting, timeseries, Mean Squared Error*

1. INTRODUCTION

The economic development in Indonesia is growing rapidly, competition is increasing. One of the things that are important in supporting the success of a company is to be able to meet the demands of various consumers. Activities that enable us to predict what will happen in the future are known as forecasting.
In addition to previously known forecasting methods, there is a method introduced by Assimakopoulos Theta and Nikolopoulos (2000). Initially, forecasting only use 2 theta weight parameters, later another theta parameter is added. Currently 5 theta weight parameter is used. By adding these weights, the error generated is smaller. This research is to apply the Theta Method to perform forecasting on car sales, minimizing the error using Mean Squared Error method.

1.1 Scope of the Research.

The scope of this research is as follows:

1. Forecasting method Theta.
2. The analyzed data are cars of type TBR, taking car sales from March 2008 until October 2008.
3. The method of forecasting accuracy is used to determine the theta method. The best theta is the one where the MSE (Mean Squared Error) is minimized.
4. Develop an application to determine the most appropriate method of forecasting. This application compares various methods. Using Theta by the number and weight of different theta and the best method will be obtained by considering the minimum MSE, and then calculates the value of forecasting.
5. The results of forecasting are only used as information for the company to make decisions.

This Research will discuss the following matters:

1. Analysis of factors - factors that affect the sales of cars such as the rate of inflation, the purchasing power of the people and other influences that directly or indirectly affect the sales.

2. Integration of applications created using systems running in the company.

1.2 Objectives

This research aims to find the best method to forecast sales using Theta by the value and theta different weights, and then the accuracy of forecast will be tested by minimizing MSE.

2. LITERATURE REVIEW

2.1 Theta Method

Design an application program that is used to compare Theta method with the value and weight of different theta.

Suppose $\left\{X_1,...,X_n\right\}$ is a univariate timeseries that is to be observed. From the series, a new series $\left\{Y_1\left(\text{¯}\right),...,Y_n\left(\text{¯}\right)\right\}$ is formed such that $Y_t''(\text{¯}) = \text{¯}\, X_t''$ where X_t'' is the second difference of X_t and $Y_t''\left(\text{¯}\right)$ is the second difference of $Y_t\left(\text{¯}\right)$ that has a solution as stated by (Box, 2013)

$$Y_t(\text{¯}) = a_{-} + b_{-}(t-1) + \text{¯}\, X_t$$

where a_{-} and b_{-} are constants. So $Y_t(\text{¯})$ is equivalent to the linear function of X_t by adding a linear trend. Assimakopoulos and Nikolopoulos (2000) states that $Y_t(\text{¯})$ is a “theta line” (Assimakopoulos, 2000). For a fixed value of ¯, given the value of $Y_1\left(\text{¯}\right)$ and $Y_2\left(\text{¯}\right)-Y_1\left(\text{¯}\right)$ which simplifies to the sum of squared differences as follows:

$$\sum_{i=1}^{t}\left[X_t - Y_t\left(\text{¯}\right)\right]^2 = \sum_{i=1}^{t}\left[\left(1-\text{¯}\right)X_t - a_{-} - b_{-}\left(t-1\right)\right]^2 \qquad (2)$$

This is equivalent to the sum of squares difference knowing a_{-} and b_{-}. Therefore this become a simple regression mapping $\left(1-\text{¯}\right)X_t$ against time $t-1$ (Rob J. Hyndman, 2003). Thus the solution is:

$$\hat{b}_{-,n} = \frac{6\left(1-\text{¯}\right)}{n^2-1}\left(\frac{2}{n}\sum_{t=1}^{n} tX_t - \left(n+1\right)\bar{X}\right) \qquad (3)$$

and $\hat{a}_{-,n} = \left(1-\text{¯}\right)\bar{X} - \hat{b}_{-,n}\left(n-1\right)/2$ (4)

The fitted timeseries is as follows:

$$\bar{Y}\left(\text{¯}\right) = \hat{a}_{-,n} + \hat{b}_{-,n}\left(n-1\right)/2 + \text{¯}\,\bar{X} = \bar{X} \qquad (5)$$

It is clearly seen that $\frac{1}{2}\left[Y_t\left(1+p\right)+Y_t\left(1-p\right)\right] = X_t$ because $\hat{a}_{1+p,n} + \hat{a}_{1-p,n} = 0$ and $\hat{b}_{1+p,n} + \hat{b}_{1-p,n} = 0$.

The forecast using Theta method is obtained by calculating the average weight of $Y_t\left(\text{¯}\right)$ using different values of ¯. Assimakopoulos and Nikolopoulos (2000) explained how to get the forecast value for ¯ = 0 and ¯ = 2. They defined the following:

$$X_{n+h} = \frac{1}{2}\left[Y_{n+h}\left(0\right) + Y_{n+h}\left(2\right)\right] \qquad (6)$$

where : X_{n+h} = forecast value (Ft)

$Y_{n+h}\left(0\right)$ = forecast value from theta weighted 0

$Y_{n+h}\left(2\right)$ = forecast value from theta weighted 2

$Y_{n+h}(0)$ is obtained by extrapolating linearly from the equations (2-6) and $Y_{n+h}(2)$ is obtained using simple exponential smoothing (SES) on the timeseries $\{Y_t(2)\}$. Thus,

$$Y_{n+h}(0) = \hat{a}_{0,n} + \hat{b}_{0,n}(n+h-1) \qquad (7)$$

approach by (Makridakis, Wheelwright, & McGee, 1983) stated

$$Y_{n+h}(2) = \sim \sum_{i=0}^{n-1} (1-\sim)^i Y_{n-i}(2) + (1-\sim)^n Y_1(2) \qquad (8)$$

where $\sim$ is the coefficient of smoothing for SES.

From the above result, it can be combined to get a simple forecast for X_{n+h}. From equations (2-8),

$$Y_{n+h}(2) = \sim \sum_{i=0}^{n-1} (1-\sim)^i \left[\hat{a}_{2,n} + \hat{b}_{2,n}(n-i-1) + 2X_{n-i}\right] + (1-\sim)^n \left(\hat{a}_{2,n} + 2X_1\right)$$

$$= \hat{a}_{2,n} + \hat{b}_{2,n}\left[n - \frac{1}{\sim} + \frac{(1-\sim)^n}{\sim}\right] + 2X_{n+h} \qquad (9)$$

where X_{n+h} is the SES forecast for the timeseries $\{X_t\}$. Given that $\hat{a}_{2,n} = -\hat{a}_{0,n}$ and $\hat{b}_{2,n} = -\hat{b}_{0,n}$,

$$X_{n+h} = X_{n+h} + \tfrac{1}{2}\hat{b}_{0,n}\left(h - 1 + \frac{1}{\sim} - \frac{(1-\sim)^n}{\sim}\right) \qquad (10)$$

For large *n*,

$$X_{n+h} = X_{n+h} + \tfrac{1}{2}\hat{b}_{0,n}\left(h - 1 + 1/\sim\right) \qquad (11)$$

Thus this is the SES by adding trend where the slope of the trend is half of the trend line which is adjusted to the original timeseries.

2.2 Durbin-Watson Statistics

The Durbin-Watson test statistics tests the hypothesis that there is no auto correlation on the residue. Durbin-Watson statistics is explained below: (Keller, 2014):

$$D-W = \frac{\sum_{t=2}^{n} (e_t - e_{t-1})^2}{\sum_{t=1}^{n} {e_t}^2} \qquad (12)$$

where : e_t is Xt – Ft

The Durbin-Watson distribution is symmetric around 2, which is the middle value. Therefore the confidence level can formed involving five regions as shown in figure 1.

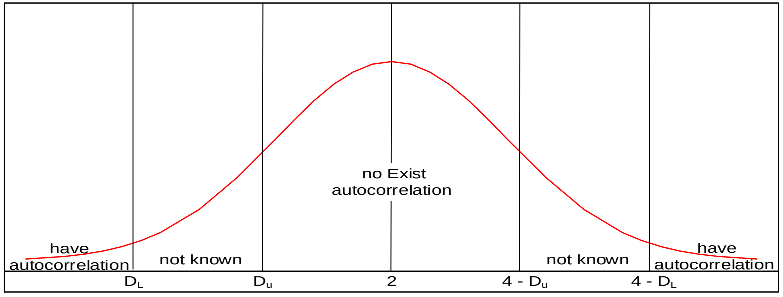

Figure 1. Durbin-Watson distribution
(Makridakis, Wheelwright, & McGee, 1983)

The five intervals are (Keller, 2014):

1. Less than D_L means there is a positive autocorrelation.
2. Bewteen D_L and D_U means that it cannot be determined.
3. Between D_U and 4 – D_U means there is o autocorrelation.
4. Between 4 – D_U and 4 – D_L means that it cannot be determined.
5. Greater than 4 – D_L means that there is a negative autocorrelation.

2.3 *Mean Squared Error* (MSE)

Makridakis, Wheelwright, and McGee uses several statistical standards to measure he accuracy of the forecast (Makridakis, Wheelwright, & McGee, 1983) The measurements shows the match between the model and the historical data. The comparison of the MSE value against forecast may provide indication of the accuracy of the model in the forecast:

$$e_t = X_t - F_t \tag{13}$$

where : e_t = error for period t.

Xt = actual data for period t

Ft = forecast for period t

If there is an observation and forecast for n period of time, then there exists n errors and the standard statistics measurement can be defined as Mean Squared Error,

$$\text{MSE} = \sum_{i=1}^{n} e_t^2 / n \tag{14}$$

2.4 File Input (Database)

In the program application, the input is obtained from *Microsoft Excel* (ekstensi .xls). The first two columns of this file is loaded. The first column contains the period and the second column contains the sales at the corresponding time. The data contain more than 30 rows such that a good forecast can be obtained.

2.4 Unified Modelling Language (UML)

UML is a standard language to create a blue print of software. UML can be used to visualize, determine and create documentation of the software design. There are several diagrams in UML:

1. *Class Diagram* to show the relationship among objects.
2. *Object Diagram* is the object and relation as a representation of the prototype.
3. *Component Diagram* is the components and relations that illustrate system implementation.
4. *Deployment Diagram* contains configuration from node and objects.
5. *Use case Diagram* is used to organize use cases and behaviors.
6. *Sequence Diagram* represents the time sequence of message and the object life line
7. *Collaboration Diagram* represents the time sequence of message and the objects in the interactions.
8. *Flow Diagram* represents the work flow of the activities, focused on the operations between objects.
9. *Activity Diagram* is a diagram the represents the *life cycle* of the object in a transition from one state to the other state.

3.1 RESEARCH DATA

Sales data is obtained from a car dealer, taking sales data for 30 weeks, from March 2008 until October 2008. The variable used is period (t), sales data (Xt) and forecast (Ft).

The hypothesis testing is using *Theta* Method with *theta* weight of 2, 3, 4 and 5 of which theta weight will provide the minimum MSE. The data has 30 periods, outliers is initially checked. When an outlier is detected, it will be corrected using Moving Average smoothing, followed by smoothing using theta weights and calculate the forecast where the weights has value 0 and 2 (for theta weighted 0, forecast is performed using Simple Linear Regression, whereas for other forecast value use the simple exponential smoothing.

The result of the forecasting data, will be calculated using MSE in order to get the accuracy.

3.2 Application Design

The Use case diagram of the system can be seen in figure 2 below. The system performs three main functions, data correction, forecasting using manual theta weight and forecasting using the best theta weight.

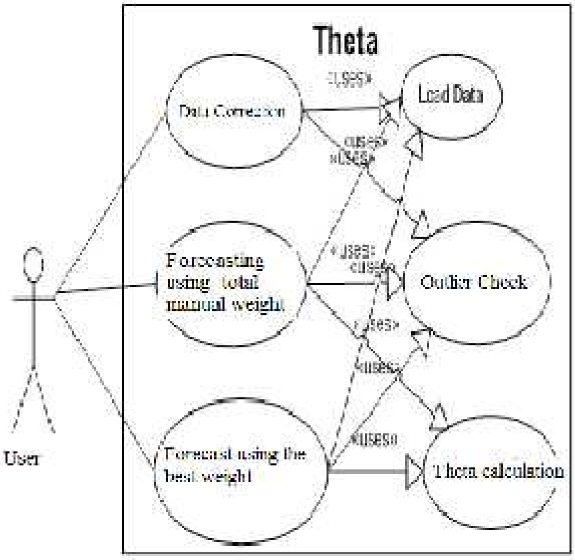

Figure 2. *Use Case Diagram* of the System

4 RESULTS AND DISCUSSION

4.1 Results

The following table shows the sales data for 8 months.

Table 1. Sales data of car type TBR from march 2008 until October 2008

Period	Month	Week Number	Sales	Period	Month	Week Number	Sales
1	March	4	5	16		3	6
2	April	1	5	17		4	13
3		2	5	18		5	13
4		3	8	19	August	1	10
5		4	7	20		2	2
6	May	1	11	21		3	6
7		2	5	22		4	3
8		3	13	23	September	1	14
9		4	6	24		2	6
10	June	1	19	25		3	8
11		2	2	26		4	4
12		3	2	27	October	1	6
13		4	8	28		2	4
14	July	1	13	29		3	7
15		2	8	30		4	7

The sales data has maximum value 19 and minimum value 2. Standard Deviation is 3.9 and the mean is 7.5. Graphically, the trend is shown in figure 3.

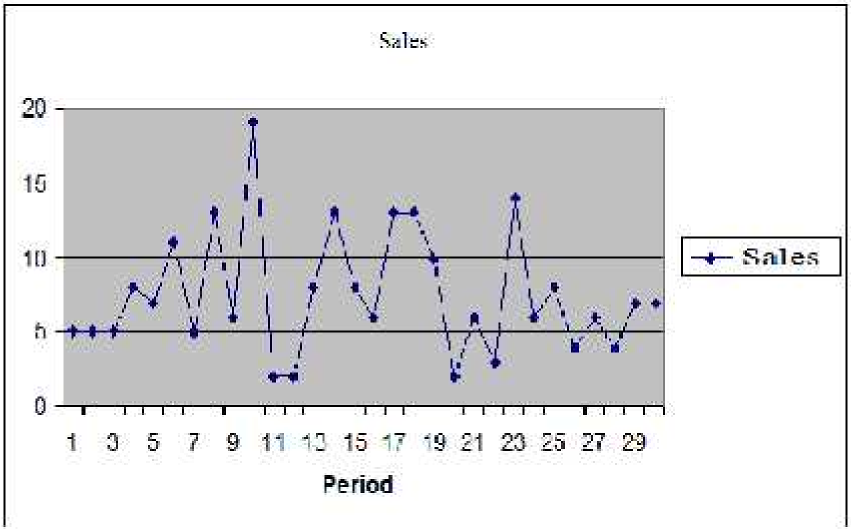

Figure 3. Sales Trend graph for 30 periods.

After performing correction, the trend can be seen in figure 4.

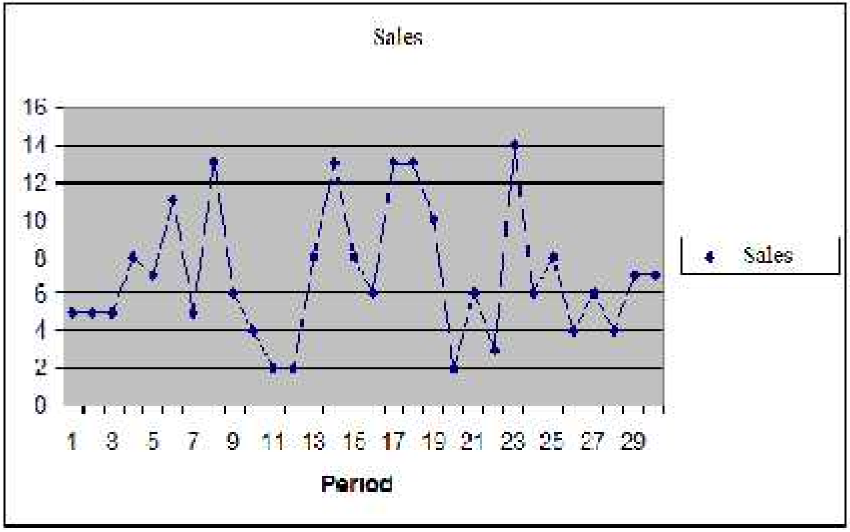

Figure 4. Sales Trend graph after correction

4.2 Discussions

The forecasting method is using theta weighted 2, 3, 4 and 5. Each method will calculate the forecast using the best alpha for each and every methods. The use of alpha is based on experiments, starting fro alpha = 0.1 until alpha = 0.9. The theta weight used is from -3.0 until 2.0, however, the weight 0 and 2 is definitely used because it serves as the standard for theta method and limited such that there are no duplicate weights used. The following shows the results in tabular and graphic forms.

Table 2. The comparison of using Alpha against the value from the Accuracy Method MSE.

	2 theta	3 theta	4 theta	5 theta
Alpha	MSE	MSE	MSE	MSE
0.1	14.04292	12.13138	12.13420	12.13825
0.2	14.44100	12.16304	12.18197	12.18598
0.3	15.02405	12.32256	12.39456	12.39844
0.4	15.63675	12.57458	12.72433	12.72803
0.5	16.28638	12.88094	13.12333	13.12682
0.6	17.01492	13.21252	13.55568	13.55894
0.7	17.87514	13.55093	13.99924	14.00225
0.8	18.93384	13.88413	14.44015	14.44290
0.9	20.28449	14.20283	14.86855	14.87103

From 2 it can be seen that at alpha = 0.1 for different theta gives the smallest MSE value. Therefore the forecasting is set to use alpha = 0.1

Table 3. Durbin-Watson Statistics

Statistics	Forecasting method using alpha = 0.1			
	2 theta	3 theta	4 theta	5 theta
D-W	1.73	1.8	1.79	1.79

With n=30, k=1 and 5% level of significance, the value d_L = 1.34 and d_U = 1.49 and 4- d_U = 2.51, where the value of d_U < D-W < 4-d_U. Therefore it is concluded that the error is random.

4.3 Hypothesis Testing

The hypothesis is tested to determine which forecast theta values namely weight 2, 3, 4 or 5 will result in the minimum MSE. The following table shows the comparison betwrrn accuracy method MSE using theta wight 2, theta weight 3, theta weigh4 and theta weight 5:

Table 3. Comparison of f\Forecast Accuracy

Accuracy Method	Forecasting Method with alpha = 0.1			
	2 theta	3 theta	4 theta	5 theta
MSE	14.04292	12.13138	12.1342	12.13825

It shows that theta weight 3 produces the smallest error. Although theta weight 4 and 5 also shows a small error. Theta weight 2 shows a significantly larger error. Thus, theta weight 3 can be used as input in making decision.

5. CONCLUSION

Based on the result and the comparison of forecasting Theta with weight 2,3, 4 and 5, it concluded that:

[a]. The best method to use for forecasting the car sales of type TBR is Theta Method with Theta weight 3. This method however has a weakness that it can only predict for one period in the future, therefoe need the actual sales to predict the next period.

[b]. The result using Forecasting Theta Method is able to follow the actal pattern.

[c]. Forecasting using Theta Method has a weakness in which this method is not able to explain shirt term patterns.

[d]. From Durbin-Watson test and graphs it can be concluded that the error has a random behavior, whih means that the

error does not have a residue after the forecasting model is applied.

6.RECOMMENDATION

This method can serve as a reference in making decisions on sales prediction. As a recommendation, researches using different weights can be performed and different method for finding the accuracy can be selected.

7.REFERENCES

[1]. Assimakopoulos, V., & Nikolopoulos, K. (2000). Thetheta model: a decompositionapproach to forecasting. *International Journal of Forecasting*, (pp. 521-530).

[2]. Bernard Bercu, F. P. (2013). A SHARP ANALYSIS ON THE ASYMPTOTIC BEHAVIOR OF THE DURBIN–WATSON STATISTIC FOR THE FIRST-ORDER AUTOREGRESSIVE PROCESS. *EDP Sciences*, (pp. 500-530).

[3]. Box, G. (2013). Box and Jenkins: time series analysis, forecasting and control. Springr.

[4]. Keller, G. (2014). *Statistics for Management and Economics*. Cengage.

[5]. Makridakis, Wheelwright, & McGee. (1983). *Forecasting: Methods and Applications*. New York: Wiley.

[6]. Rob J. Hyndman, B. B. (2003). Unmasking Theta Methodan. *International Journal of Forecasting*, (pp. 287-290).

[7]. White, N. E. (1977). The Durbin-Watson Test for Serial Correlation with Extreme Sample Sizes or Many Regressors. Econometrica.

[8]. Booch, G., Rumbaugh, J., Jacobson, I. (2005). *The Unified Modeling Language User Guide*, Massachusetts : Addison Wesley.

[9]. Box, G. E. P., Jenkins. G. M. (1994). *Time Series Analysis: Forecasting and Control*. New Jersey : Prentice Hall.

[10].Hyndman, R J., Billah, B. (2003). Unmasking the Theta method,. *International Journal of Forecasting*, Elsevier, vol. 19(2), pages 287-290.

a. http://www.buseco.monash.edu.au/depts/ebs/pubs/wpapers/2001/wp5-01.pdf
b. Akses : 10 Agustus 2008

[11].Makridakis, S., Wheelwright, S.C., McGee, V.E. (1999). *Forecasting: methods and applications*. New York : John Wiley and Sons.

[12].Nikolopoulos, K., Assimakopoulos V., Bougioukos, N., Petropoulos, F. (2008). "Advances in the Theta model", Working Papers 0023, Tripolis : University of Peloponnese, Department of Economics.http://econ.uop.gr/~econ/RePEc/pdf/AdvancesintheThetamodel.pdf Akses : 10 Agustus 2008

[13].Nikolopoulos, K., Assimakopoulos V. (2000), The Theta Model: a decomposition approach to forecasting, *International Journal of Forecasting* vol. 16, pp. 521-530.

[14].Shneiderman, B. (1998). *Designing the User Interface: Strategies for Effective Human-Computer Interaction, (3rd ed.)*. Menlo Park, CA : Addison Wesley.

[15].Supranto, J. (2001). *Statistic: Theory & Application*. Jakarta : Erlangga.

[16].Webster, C. E. (1986). *The Executive's Guide to Business and Economic Forecasting*. Chicago, Illinois : Probus Publishing Company.