Sales forecasting newspaper with ARIMA: A case study

Cite as: AIP Conference Proceedings **1931**, 030017 (2018); https://doi.org/10.1063/1.5024076
Published Online: 09 February 2018

Carina Intan Permatasari, Wahyudi Sutopo, and Muh. Hisjam

Sales Forecasting Newspaper with ARIMA: A Case Study

Carina Intan Permatasari[a)], Wahyudi Sutopo[b)], and Muh. Hisjam[c)]

Department of Industrial Engineering, Faculty of Engineering, Sebelas Maret University, Surakarta 57126, Indonesia

[a)]Corresponding author: carinaintan1996@gmail.com
[b)]wahyudisutopo@gmail.com
[c)]hisjam@gmail.com

Abstract. People are beginning to switch to using digital media for their daily activities, including changes in newspaper reading patterns to electronic news. In uncertainty trend, the customers of printed newspaper also have switched to electronic news. It has some negative effects on the printed newspaper demand, where there is often an inaccuracy of supply with demand which means that many newspapers are returned. The aim of this paper is to predict printed newspaper demand as accurately as possible to minimize the number of returns, to keep off the missed sales and to restrain the oversupply. The autoregressive integrated moving average (ARIMA) models were adopted to predict the right number of newspapers for a real case study of a newspaper company in Surakarta. The model parameters were found using maximum likelihood method. Then, the software Eviews 9 were utilized to forecasting any particular variables in the newspaper industry. This paper finally presents the appropriate of modeling and sales forecasting newspaper based on the output of the ARIMA models. In particular, it can be recommended to use ARIMA (1, 1, 0) model in predicting the number of newspapers. ARIMA (1, 1, 0) model was chosen from three different models that it provides the smallest value of the mean absolute percentage error (MAPE).

INTRODUCTION

Inaccuracies between supply and demand on the company making it inefficient. The impact arising are overs, less, and the forecast is not referable to the next production [1]. Benefits provided from forecasting is helpful in decision-making for solving problems and developing business strategies [2]. In case study of the newspaper industry, there are several things that can affect supply and demand. People are beginning to switch to using digital media for their daily activities, including changes in newspaper reading patterns to electronic news [3]. In trend uncertainty, the customers of the printed newspaper also have switched to electronic news. It has some negative effects on the printed newspaper demand, where there is often an inaccuracy of supply with demand which means that many newspapers are returned.

The newspaper has a special character *i.e.* daily production and there is no stock because the information cannot be sold in delay. If the information is delayed then the product is not sold commercially. So the industry newspaper there has been previous research which will conduct a discussion of model and predicted the newspaper include the determination of the appropriate forecast by using data mining to meet customer demand in the newspaper industry with the method RRBF [4]. From other research, there is also a discussion of the newspaper production predicted accurately by means of uniting the desires of the customer with the amount of paper that is not sold by the method of fuzzy clustring [2]. There is also research that discusses allows prediction of the demand of newspaper with accurate customer desire to unite with the minimize returns of newspapers so there happen excessive production which led to the loss by the method of ARIMA [5]. From the most research about ARIMA are discusses the modeling and forecasting [6]. This article is discusses the appropriate pattern to try on a real case of a newspaper industry in solo had the characteristics of daily production. Which at the moment are having problems of inaccuracies reaches 5 – 10%.

The 3rd International Conference on Industrial, Mechanical, Electrical, and Chemical Engineering
AIP Conf. Proc. 1931, 030017-1–030017-10; https://doi.org/10.1063/1.5024076
Published by AIP Publishing. 978-0-7354-1623-9/$30.00

Modeling is observation from earlier data by developing a model to describe the structure of the estimates in corresponding with the future [7]. While the demand forecast is predicting future demand action by learning from the past [8]. Various models of demand forecast include Autoreggressive (AR), Moving Average (MA), Autoregressive Moving Average (ARMA), Autoregressive Integrated Moving Average (ARIMA) [9]. Compared with the early AR, MA and ARMA model, ARIMA model is more flexible in the application and more accurate in the quality of the simulative or predictive results [5]. The ARIMA analysis, an identified underlying process is generated based on observations to a time series for generating a good model which shows the process-generating mechanism precisely [9]. Some previous research about demand forecast for car demand is a forecast with long-term characteristics and can be stored [10]. Model for forecasting sales of distributors in plastic industry that have long-term characteristics and can be saved anyway [11]. In the case of electrical load forecasting also have long-term characteristics and can be stored [12]. While in this research is the real problems in the newspaper case study to develop forecasting parameters. The aim of this research is to select an appropriate ARIMA model in forecasting newspaper demand.

The article is organize as follows, in part 1 is present background research and outlined problems in the real system. In part 2, present is the basic theory in feasibility analysis. In part 3, present is a method to solve the problem. Discussion and analysis are present in part 4 and the conclusions are present in part 5.

METHODOLOGY

The Box – Jenkins method or ARIMA is used for forecasting short term. For the long term modeling this result cannot constant. ARIMA can be defined as the combination of two autoregressive (AR) model that is integrated with the Moving Average (MA) model. Writing the notation Autoregressive Integrated Moving Average is an ARIMA (p, d, q) [13]. P is the degree of process of AR, d is the order of differencing and q is the degree of MA process.

Autoregressive model with the ordo of the AR (p) model of ARIMA (p,0,0) is stated as follows [9]:

$$Y_t = \theta_0 + \theta_1 Y_{t-1} + \theta_2 Y_{t-2} + \cdots + \theta_p Y_{t-p} + e_t \quad (1)$$

Where :

Y_t = Stationary time series

θ_0 = Constant

θ_p = Parameter of autoregressive model

e_t = Residual time (t)

Moving Average model with the ordo of the MA (q) or ARIMA (0,0, q) is stated as follows:

$$Y_t = \emptyset_0 - \emptyset_1 Y_{t-1} - \emptyset_2 Y_{t-2} - \cdots - \emptyset_q Y_{t-q} + e_t \quad (2)$$

Where :

Y_t = Stationary time series

$\emptyset_0$ = Constant

$\emptyset_q$ = Coefficient of the model which shows the moving average weights

e_t = Residual tense used

To ensure the results obtainable right from ARIMA has accurate and reduce the level of error can be used with four models-selection criteria include root mean square error (RMSE), mean absolute error (MAE), mean absolute percentage error (MAPE) and Theil Inequality Coefficient.

TABLE 1. Model – accurancy metrics

Criteria	Formula
Root Mean Squared Error (RMSE)	$\sqrt{\sum_{t=T+1}^{T+h} (\hat{y}_t - y_t)^2 / h}$
Mean Absolute Error (MAE)	$\sum_{t=T+1}^{T+h} \lvert \hat{y}_t - y_t \rvert / h$
Mean Absolute Percentage Error (MAPE)	$100 \sum_{t=T+1}^{T+h} \left\lvert \frac{\hat{y}_t - y_t}{y_t} \right\rvert / h$
Theil Inequality Coefficient	$\frac{\sqrt{\sum_{t=T+1}^{T+h} (\hat{y}_t - y_t)^2 / h}}{\sqrt{\sum_{t=T+1}^{T+h} \hat{y}_t^{\,2} / h} - \sqrt{\sum_{t=T+1}^{T+h} y_t^{\,2} / h}}$

The first assessment criterion, RMSE is preserves the units of the estimation variable. This approach is more sensitive and minimizes large errors. Nevertheless, the ability to compare different time series is limited with this criterion. Conversely, MAE, the second criterion, determines the error magnitude for a precise set of forecasts. MAE defines how close forecasts are to the actual outcomes. This metric does not consider the direction of the forecasts. Moreover, these criteria determine the precisions of continuous variables. The third criteria is Theil Inequality Coefficient (U_1 and U_2), respectively. The former enables different predictions to be compared, which implies that actual values are compared with predicted values. U_1 provides a range of values on a zero-to-one scale. The nearer U_1 is to zero, the more accurate the prediction is. When faced with alternative predictions, the forecast with the smallest value of U_1 is regarded as the best and is thus selected. Conversely, U_2 performs relative comparisons based on random walk models and prediction models (naïve model). The naïve model may be described as the actual predetermined forecast model applied based on an indiscriminate-walk process. When U_2 levels off at unity, the naïve method is considered to be equally useful for forecasting. $U_2 < 1$ indicates that the forecasting model would work better than the naïve approach. MAPE, the fourth criterion, enables comparison of distinct time-series data without defining the relation or percent error. This metric is significant in instances in which the measured variables are very large [14]. In this research using MAPE because of data availability.

This research is a case study of estimation, model, and predicted sales of the newspaper. The data used is historical data obtained from the company in the form of the number newspaper sales from January 2016 to March 2017. Aim to get more accurate data and can be accounted from an object.

Analysis of the behavior data consists test and non-test stationary use the ADF test, after that analysis model used Box-Jenkins method and software Eviews 9. Box Jenkins method used for estimation model equations mean. At this stage the data verification and validation problems analysis in order to time-series and estimation parameter from newspaper sales index data so obtained the best model to suit the actual circumstances [5].

DISCUSSION AND ANALYSIS

The material of this research was demand newspaper. Data sample for this study are the data obtained from industries newspaper in solo. Plot of the time series original data demand newspaper is given in the figure bellow.

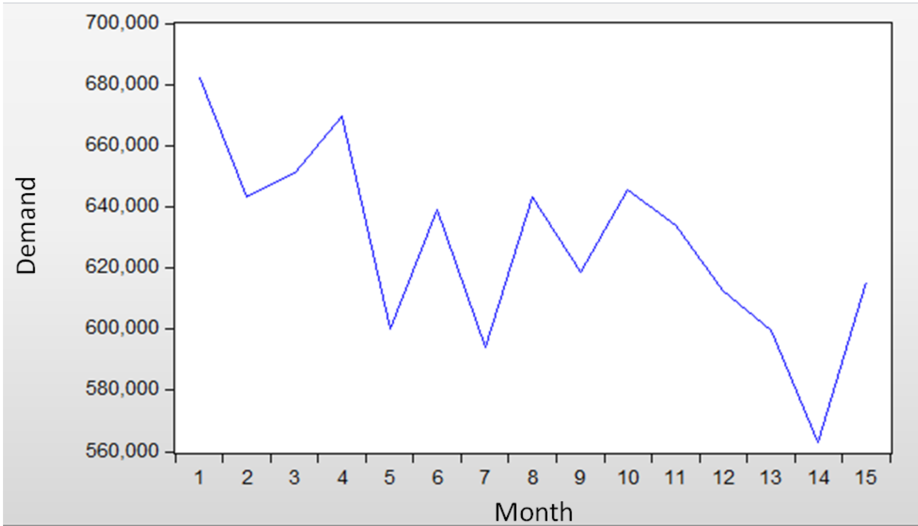


FIGURE 1. The Plot Of The Original Newspaper Sales Data

From the above data plots can be noted that the number of sales is very fluctuating that tends to decline. Based on the plot of the data indicate that the data has not been stationer against mean and variation. In particular needs to be done to test the Augmented Dickey-Fuller (ADF) so that known sales data newspaper has stationary. The result of the ADF test looks like Table 2.

TABLE 2. ADF Test

		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		-0.738086	0.7921
Test critical values:	1% level	-4.297073	
	5% level	-3.212696	
	10% level	-2.747676	

*MacKinnon (1996) one-sided p-values.

The value of the t-statistic in output is – 0.738086, still smaller than the value in table t McKinon at trust level 1%, 5%, or 10%. As well as the value of the Probability of 0.7921 is still greater than the value of the critique of α = 0.05 (0.7921 > 0.05). The results of the output indicates that the data are not stationary. This data indicates need for differentiation and transformation. So that the data becomes stationary. ADF test done first with differentiation results done as in table 3.

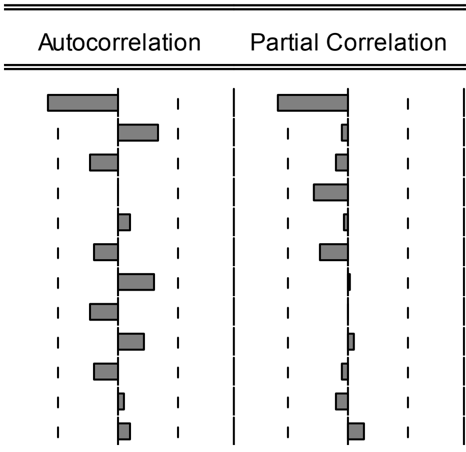
TABLE 3. ADF Test with 1st difference

		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		-3.939905	0.0169
Test critical values:	1% level	-4.297073	
	5% level	-3.212696	
	10% level	-2.747676	

*MacKinnon (1996) one-sided p-values.

The value of the t-statistic in output is -3.939905 is already greater than the value in table t McKinon at trust level 5% and 10%. As well as the value of the probability of 0.0169 is already smaller than the value of the critique of 0.05 (0.0169 < 0.05). Thus the data has been stationary on the differentiation of the first stage (1st difference) and the null hypothesis can be rejected. After that, the next process is to do an analysis of the time series model with ARIMA.

ACF and PACF plot made to identify a suitable data for means of data. The results of the correlogram with the first differentiation will show ACF and PACF graph like Figure 2.



	AC	PAC	Q-Stat	Prob
1	-0.626	-0.626	6.7573	0.009
2	0.354	-0.063	9.0948	0.011
3	-0.258	-0.102	10.448	0.015
4	0.001	-0.316	10.448	0.034
5	0.106	-0.045	10.728	0.057
6	-0.227	-0.257	12.166	0.058
7	0.325	0.016	15.555	0.030
8	-0.256	-0.001	18.010	0.021
9	0.238	0.045	20.555	0.015
10	-0.215	-0.051	23.143	0.010
11	0.065	-0.103	23.454	0.015
12	0.099	0.153	24.548	0.017

FIGURE 2. ACF and PACF

From the above graph model, it can be predicted that the model of ARIMA is used for proper ARIMA (1, 1, 0), ARIMA (0, 1, 1), ARIMA (1, 1.1) without constant. Next do the estimation of the value of C, probability, and AIC on each model.

TABLE 4. Models of ARIMA (1, 1, 0)

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	628259.3	9373.630	67.02412	0.0000
AR(1)	0.226804	0.363252	0.624373	0.5441
SIGMASQ	8.61E+08	4.43E+08	1.941980	0.0760
R-squared	0.046002	Mean dependent var	627452.2	
Adjusted R-squared	-0.112998	S.D. dependent var	31099.19	
S.E. of regression	32809.24	Akaike info criterion	23.81519	
Sum squared resid	1.29E+10	Schwarz criterion	23.95680	
Log likelihood	-175.6139	Hannan-Quinn criter.	23.81368	
F-statistic	0.289322	Durbin-Watson stat	1.992774	
Prob(F-statistic)	0.753850			
Inverted AR Roots	.23			

TABLE 5. Models of ARIMA (0, 1, 1)

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	627862.0	8507.323	73.80253	0.0000
MA(1)	0.127373	0.402339	0.316582	0.7570
SIGMASQ	8.80E+08	4.33E+08	2.032180	0.0649

R-squared	0.025116	Mean dependent var	627452.2
Adjusted R-squared	-0.137365	S.D. dependent var	31099.19
S.E. of regression	33166.45	Akaike info criterion	23.83441
Sum squared resid	1.32E+10	Schwarz criterion	23.97602
Log likelihood	-175.7581	Hannan-Quinn criter.	23.83290
F-statistic	0.154577	Durbin-Watson stat	1.729068
Prob(F-statistic)	0.858456		
Inverted MA Roots	-.13		

TABLE 6. Models of ARIMA (1, 1, 1)

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	628580.8	18621.67	33.75533	0.0000
AR(1)	0.730270	1.290236	0.565998	0.5828
MA(1)	-0.464509	1.736313	-0.267526	0.7940
SIGMASQ	8.18E+08	5.06E+08	1.614447	0.1347

R-squared	0.094349	Mean dependent var	627452.2
Adjusted R-squared	-0.152647	S.D. dependent var	31099.19
S.E. of regression	33388.53	Akaike info criterion	23.90475
Sum squared resid	1.23E+10	Schwarz criterion	24.09357
Log likelihood	-175.2857	Hannan-Quinn criter.	23.90274
F-statistic	0.381985	Durbin-Watson stat	2.089046
Prob(F-statistic)	0.768040		
Inverted AR Roots	.73		
Inverted MA Roots	.46		

To determine the best model is to compare to the four models who are looking for a model with a value of AIC and Schwarz criterion to the smallest. From the results above, it is well known that the best model is the ARIMA (1, 1,0) without constant. Next is doing a diagnostic check to perform a test of normality residue. The results can be seen in Figure 3.

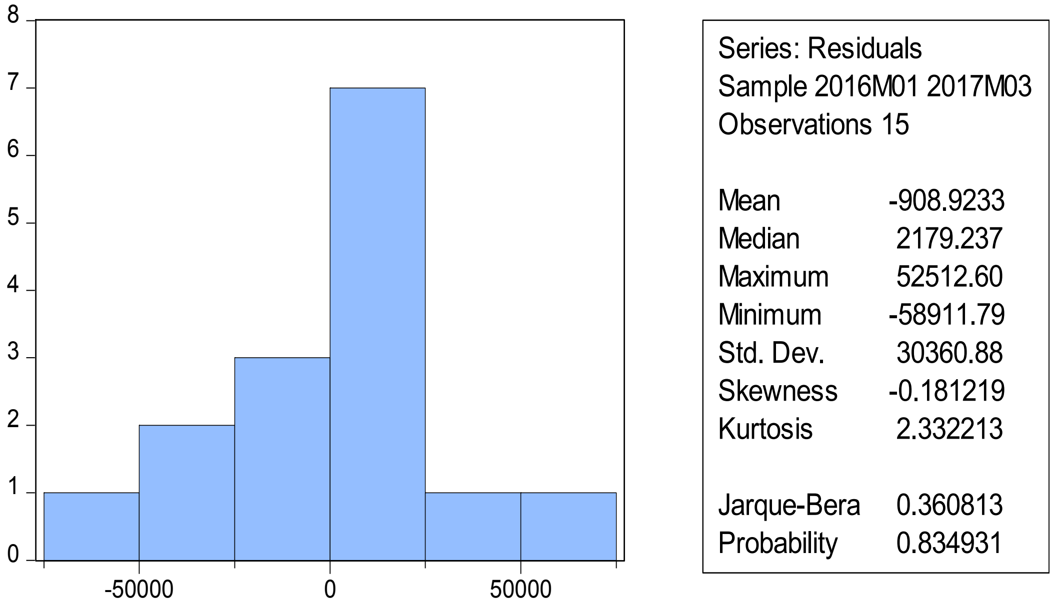


FIGURE 3. The Results Of The Diagnostic Check

Based on the above, it can be seen that the output value of probability > alpha *i.e.* 0.913103 > 0.05 means that the data are normal and has been stationary against the variation. This means that these data have a relatively stable fluctuations from time to time. To prove that data are already normal can use assumptions autocorrelation test and assumptions heteroscedasticity test.

Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	-0.108	-0.108	0.2136	
		2	0.327	0.319	2.3164	0.128
		3	-0.102	-0.049	2.5395	0.281
		4	-0.176	-0.329	3.2609	0.353
		5	0.074	0.115	3.4021	0.493
		6	-0.175	0.011	4.2684	0.511
		7	0.311	0.234	7.3607	0.289
		8	-0.109	-0.075	7.7952	0.351
		9	0.158	-0.041	8.8564	0.355
		10	-0.207	-0.202	11.040	0.273
		11	-0.156	-0.130	12.601	0.247
		12	-0.110	-0.046	13.633	0.254

FIGURE 4. Test Correlation Assumptions

Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	-0.210	-0.210	0.7997	0.371
		2	-0.113	-0.164	1.0515	0.591
		3	-0.006	-0.073	1.0521	0.789
		4	0.029	-0.009	1.0719	0.899
		5	-0.197	-0.217	2.0570	0.841
		6	0.049	-0.056	2.1251	0.908
		7	0.076	0.016	2.3075	0.941
		8	-0.182	-0.201	3.5127	0.898
		9	0.122	0.048	4.1470	0.901
		10	-0.020	-0.082	4.1683	0.939
		11	-0.166	-0.223	5.9347	0.878
		12	-0.151	-0.290	7.8769	0.795

FIGURE 5. Test Assumption Heteroscedasticity

After that it can be determined the sales forecast for short periods of time. The results of the forecast as shown in Figure 6. From the result MSE is 27028, 18; MAE is 21855,49; and MAPE is 3,52.

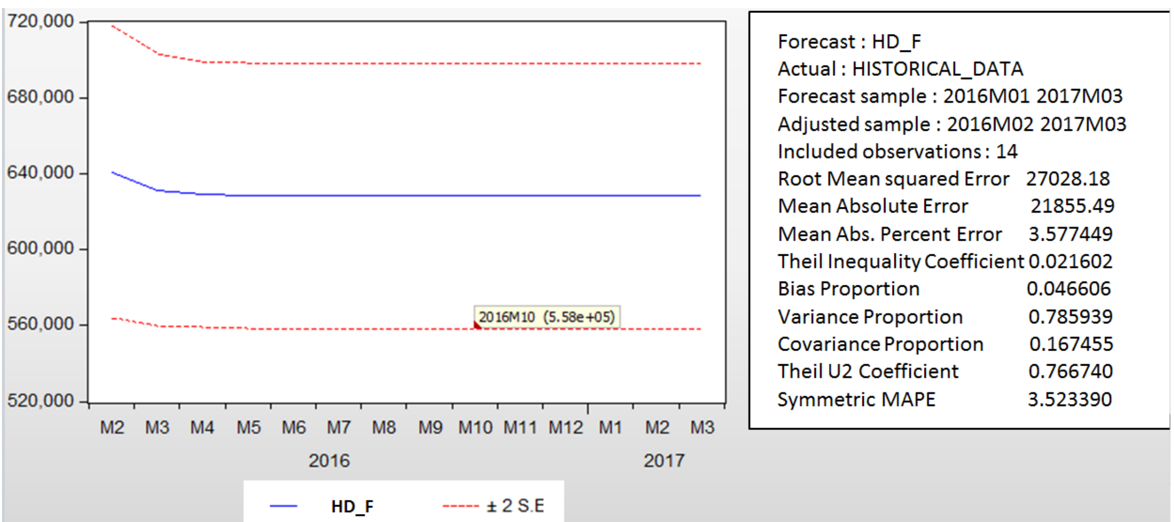


FIGURE 6. The Results Of The Forecast

Forecasting should be reference of the customer order cycle to improve accuracy and lower returned [15]. In order that sustainable improvement it needs to be stated in SOP as done by the large companies, include GMF [16], Pedro Neves Company [2], Hearst Company [17], and newspaper industries of UK & Netherland [18] in improving its performance on an ongoing basis. In general the contributions of forecasting is to improve the relationship between the manufacturers and the consumers. Obtained ARIMA models that enable thefollowing:

$\Delta Y_t = \theta_0 + \theta_1 \Delta Y_{t-1} + e_{t-1}$

$\Delta Y_t = C + 0.226804 \Delta Y_{t-1} + e_{t-1}$

$\Delta Y_t = 628{,}259.3 + 0.226804 + 3.52$

$\Delta Y_t = 628{,}263.046804$

$\Delta Y_t = \pm 628.264$

From the result, the production of newspaper for each month are ± 628.264.

CONCLUSION

The number of newspaper sales data are not stationary and indicates instability the residual variation, then the fluctuating data need ADF test for get the stationary. To perform the demand forecast to be done stationary data with the first differentiation. Based on the selection of the best models of ARIMA model is obtained (1, 1, 0). $\Delta Y_t = C + 0{,}226804\ \Delta Y_{t-1} + e_{t-1}$ and the results for the next 3 months of ± 628,264. ARIMA models can be used to predict the short-term forecast in good, it is supported with a level MAPE 3.52%. Circulation managers can use the results of this forecasting for planning determines the production master plan.

For further research should be able to put a profile of consumers and competition behavior between a newspaper in one agency, then for forecasting needs to be combined the forecasting with competition in every agency.

ACKNOWLEDGMENTS

This study is supported by PT Aksara Solopos. The authors are grateful to PT. Aksara Solopos for helpful to collect the data.

REFERENCES

1. Singh, A. P., Manoj Kumar Gaur, Dinesh Kumar Kasdekar, and Sharad Agrawal, A Study of Time Series Model for Forecasting of Boot in Shoe Industry, *International Journal of Hybrid Information Technology*, vol. 8, no. 8, pp. 143 – 152, 2015.
2. Cardoso, G. and F. Gomide, Newspaper Demand Prediction and Replacement Model based on Fuzzy Clustering and Rules, *International Journal of Information Sciences*, pp. 4799 – 4809, 2007.
3. Spil and Kijl, *A Business Model for the E-Newspaper from a Customer Perspective*, Master thesis: MSc in Business administration, University of Twente, 2014.
4. Lertuthai, M., Manisara Baramichai, and Ungul Laptaned, Development of the Adaptive Forecasting Model for Retail Commodities by Using Leading Indicator: Retailing Rule Based Forecasting Model (RRBF), *Proceedings of the World Congress on Engineering and Computer Science*, vol. 2, 2009.
5. Incesu G., Baris Asikgil, and Mujgan Tez, Sales Forecasting System for Newspaper Distribution Companies in Turkey, *Pak.j.sts.oper.res*. vol. 8, no. 3, pp. 685 – 699, 2012.
6. Rojas, Juan Pedro Sepúlveda, Felipe Rojas, Héctor Valdés-González, dan Mario San Martín, Forecasting Models Selection Mechanism for Supply Chain Demand Estimation, *Procedia Computer Science*, pp. 1060 – 1068, 2015.
7. Adhikari, R. and R.K. Agrawal, *An Introductory Study on Time Series Modeling and Forecasting*, English, Lambert Academic Publishing (LAP), 2013.
8. Raicharoen, T., C. Lursinsap, P. Sanguanbhoki, Application of Critical Support Vector Machine to Time Series Prediction, *Proceedings of the 2003 International Symposium, ISCAS '03*, vol. 5, pp. 741 – 744, 2003.
9. Box, G.E.P. and Jenkins, G.M., *Time Series Analysis: Forecasting and Control*, USA, Revised ed. Holden-Day, 1976.
10. Jansson, Jan Owen, Car Demand Modelling and Forecasting, *Journal of Transport Economics and Policy*, pp. 125 – 140, 1989.
11. Udom, P. and N. Phumchusri, A Comparison Study between Time Series Model and ARIMA Model for Sales Forecasting of Distributor in Plastic Industry, *IOSR Journal of Engineering*, vol. 4, no. 2, pp. 32 – 38, 2014.
12. Dong, Y., Jianzhou Wang, Chen Wang, and Zhenhai Guo, Research and Application of Hybrid Forecasting Model Based on an Optimal Feature Selection System—A Case Study on Electrical Load Forecasting, *Journal of Energies*, pp. 490 – 512, 2017.
13. Makridakis, Spyros., Wheelwright, C, Steven, Mcgree, E, Victor, Metode Mathematical Programming, and System Dynamics Methods, *International Journal of Forecasting*, vol. 12, pp. 329 – 343, 1999.
14. Tularam, Gurudeo Anand and Tareq Saeed, Oil-Price Forecasting Based on Various Univariate Time-Series Models, *American Journal of Operations Research*, vol. 6, pp. 226 – 235, 2016.
15. Hendry, D. F. and Mizon G. E., Forecasting from Structural Econometric Models, *Journal of Economics, Department of Economics Discussion Paper Series*, no. 597, March 2012.
16. Puteri V. E., Sutopo W., Hisjam M., and Asyrofa R., Standard Operating Procedure (SOP) to improve quality

management system (case study: SBU GMF power services), *Lecture Notes in Engineering and Computer Science*, vol. 2, pp. 745 – 750, 2016.
17. Yu, X., Zhiquan Qi, and Yuanmeng Zhao, Support Vector Regression for Newspaper/ Magazine Sales Forecasting, *Procedia Computer Science*, pp. 1055 – 1062, 2013.
18. Deleersnyder, B., Inge Geyskens, Katrijn Gielens, and Marnik G. Dekimpe, How Cannibalistic is the Internet channel? A Study of the Newspaper Industry in the United Kingdom and the Netherlands, *International Journal of Research in Marketing*, pp. 337 – 348, 2002.