Forecasting Retail Sales Trends Using Autoregressive Integrated Moving Average

Luky Fabrianto

Master of Computer Science-Postgraduate Program, STMIK Nusa Mandiri, Jakarta, Indonesia

Abstract. One year since its inception, in the first nine months sales fluctuated but were in a profitable transaction value and could cover operational costs, but in the last three months the sales value tends to be flat and declining, this is thought to be caused by stagnant supply of goods from distributors so that many consumer needs are not available in stores. It is time for the store management to turn to other distributors and provide opportunities for members to invest in investing for the continued existence of the store.

This study examines the right method for predicting store sales trends, the data used is sales data one year since the store was established. The results obtained are about 5% outliers in transactions and large transaction fluctuations inter-day so that smoothing is needed to minimize shocks in transaction data, the last stage of this study is forecasting sales trends using the ARIMA method, the model obtained is ARIMA (3, 1, 2) with a percentage of errors or MAPE of 2.5% and an RMSE value of 112883,14236. For the next 15 days the sales value is predicted to be in the range of Rp. 3,187,116 to Rp. 3,612,883.

1. Introduction

The method of forecasting which is mostly developed by researchers in economics scope is a quantitative forecasting method where the method is divided into two types, the first is the regression method and the second is the time series method[1].

There are several studies related to forecasting in various fields that have been done before, such as: Comparison of several forecasting methods from the fields of statistics and computer science in sales forecasting, the results of ARIMA from the field of statistical science can outperform several forecasting methods from the field of computer science[2], research that compares methods Forecasting between ARIMA and Support Vector Machine (SVM) results MAPE ARIMA 7.07% and SVM 9.59% [3], the use of ARIMA is used in forecasting the import value of Indonesian iron ore from 2008 - 2017 to explain the advantages and disadvantages of ARIMA, one of the points is the accurate ARIMA method For short-term forecasting[4], and the comparison between the ARIMA forecasting method with the Radial Basis Function (RBF), the ARIMA error is smaller than RBF and RBF-ARIMA[5].

In this study, transactions recorded in database storage will be used as research to make a sales forecast for next 15 days or two weeks using the ARIMA forecasting method. The current condition, minimarket is requires investment to improve from the supply side, the management opens up investment opportunities for members (individuals or groups) for 15 days because wholesaler stocks for supplies is carried out twice a month.

2. Methodology

2.1. Outlier Removal

Outlier is a condition of data that deviates from other data sets[6], also can be interpreted as observations that do not follow the majority of patterns and far from the center data[7], and may have a major effect on regression coefficient[8]. If there are outliers in the dataset, diagnostics are needed to identify the outliers, one of which is by removing outliers from the data group then analyzing the data without outliers.

A common method of eliminating outliers is to use quartile values and constraints. Quartile-1 (Q1), quartile-2 (Q2), and quartile-3 (Q3) divide a data sequence into four parts. The limit or IQR (Interquartile

Range) is defined as the difference between Q1 and Q3 or IQR = Q3 - Q1. Outliers can be determined as a value (x) <1.5 * IQR against Q1 and a value (x)> 1.5 * IQR against Q3 [9].

2.2. Smoothing

The basic principle of smoothing is to identify data patterns by smoothing local variations or short fluctuations in a time series, in this case we are using moving average (MA). Moving average (MA) simply is the value at one time influenced by several values from the previous period, moving average is also suitable for data with constant or stationary pattern.

Formula of moving average is[10] :

$$MA = \frac{\sum(n\ new\ value)}{n} \tag{1}$$

2.3. Forecasting

Forecasting data for the future is carried out following systematic steps and following a model that is in accordance with the properties and patterns of the original data, the design of a good systematic step gives the results obtained to be relevant and nearly accurate.

2.4. ARIMA Modeling

ARIMA is actually an attempt to find the most suitable data pattern from a group of data, the ARIMA method fully requires historical data and current data to produce short-term forecasts [11].

The reality in everyday life is that more non-stationary data is compared to stationary data so that the *autoregressive integrated moving average* (ARIMA) level (*p*, *d*, *q*) time series model is more popular than the previously mentioned time series models. The value of *d* in ARIMA *is a differencing* value to make non-stationary data stationary [12].

The ARIMA model consists of four basic steps, namely:

First - Model Identification

In this study, *the ADF test* was used with the null hypothesis the *time series* is not stationary, where if the *p-value* is <0.05 then the null hypothesis is rejected and *the time series* data are considered stationary.

Second - Model Parameter Estimation

Purpose Parameter estimation in *time series* analysis is to form a good model provided that the model parameters must be significant or *p-value* <0.05.

Third - Check Diagnosis

This examination is carried out by testing whether the data is *whitenose* and normally distributed in order to get good forecasting results.

Fourth - Forecasting

At this stage the data is divided into two periods, namely the training period (*train*) and the test period (*test*) and the forecasting period, model building is done using training data.

Best Model Selection Criteria

After the model is identified, more than one model deemed suitable will be formed, for that purpose the smallest *Akaike Info Criterion* (AIC) & MAPE and significant P-Value are used.

3. Results

3.1. Sales Data Characteristics

This study analyzes the sales data for 1 year minus one day because the shops are closed on Eid holidays (364 days), Figure 1 is a sales data plot.

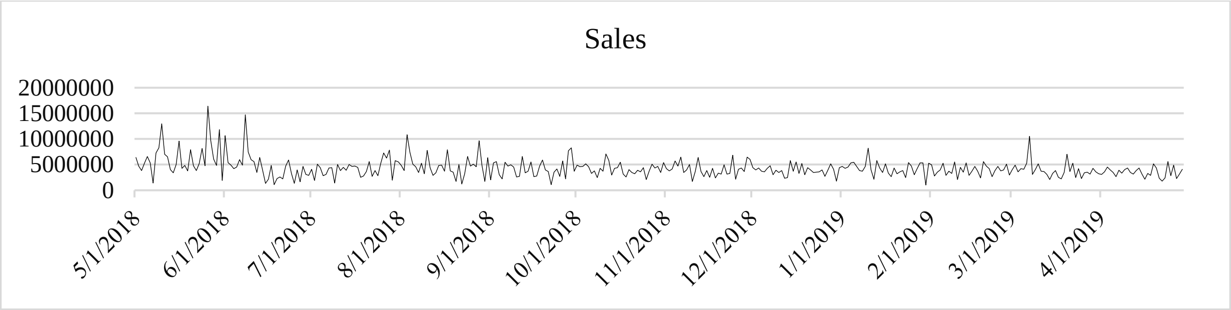

Figure 1. Plot of sales data

3.2. Outliers Removal

To perform outlier removal, the first quintile (Q1), third quintile (Q3) and inter quintile range (IQR) values must be found, as in Table 1. After obtaining the values of Q1, Q3 and IQR, the upper and lower limits can be calculated as below, the lower limit = Q1 - 1.5 * IQR = 625,147.75 and the upper limit = Q3 + 1.5 * IQR = 7,676,661.75. Sales values that are outside the lower and upper limits are considered as outliers, in this study there are 20 sales days or 5% of data that are considered outliers, Figure 2 is sales data without outliers.

Table 1.Quantiles

		Score
Quintiles	Q1	3,269,466
	Q3	5,032,344
	IQR	1,762,878.50

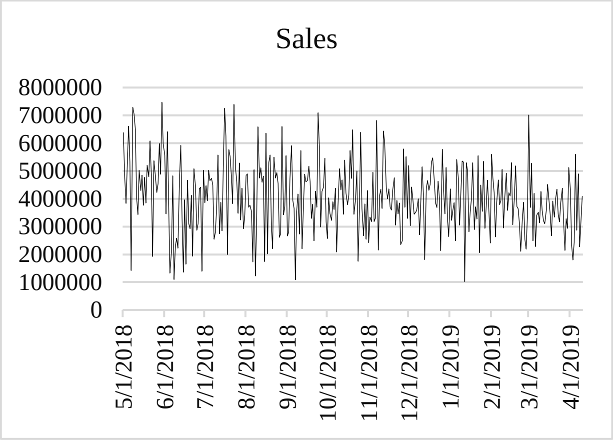

Figure 2. Plot of sales data without outliers

3.3. Smoothing

Based on Figure 4.3 above, it can be seen that the time series shows that sales data is very fluctuating, necessary to have refinement or *smoothing*, refinement can be used in two ways, firstly as forecasting and secondly it is used to reduce or eliminate short fluctuations in a time series. The refinement in this study uses a moving average with a window or window every 15 days, Figure 3 below is the result of time series refinement in this study.

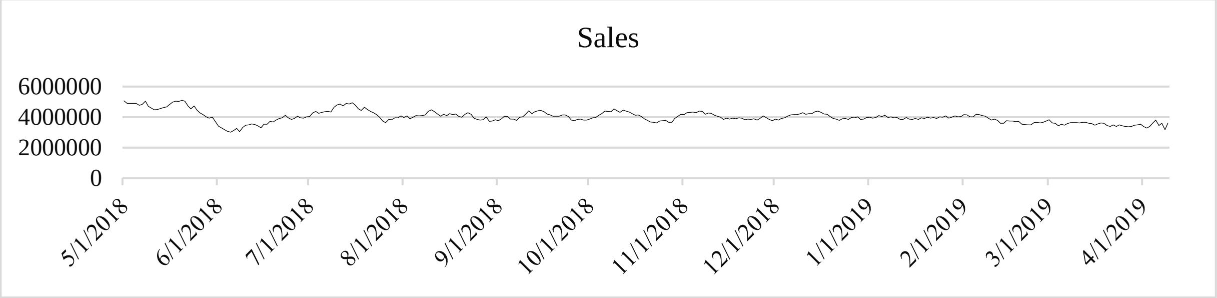

Figure 3. Plot of data after smoothing

3.4. ARIMA Forecasting

3.4.1. Stationary Identification. To determine the stationary of a time series, the calculation of the ADF Statistic and the p-value can be calculated and obtained -2.56628 and 0.100169 so that differencing is necessary, Figure 4 is a visualization of the original sales data plot and the plot of the results of the first and second differencing (seen over difference).

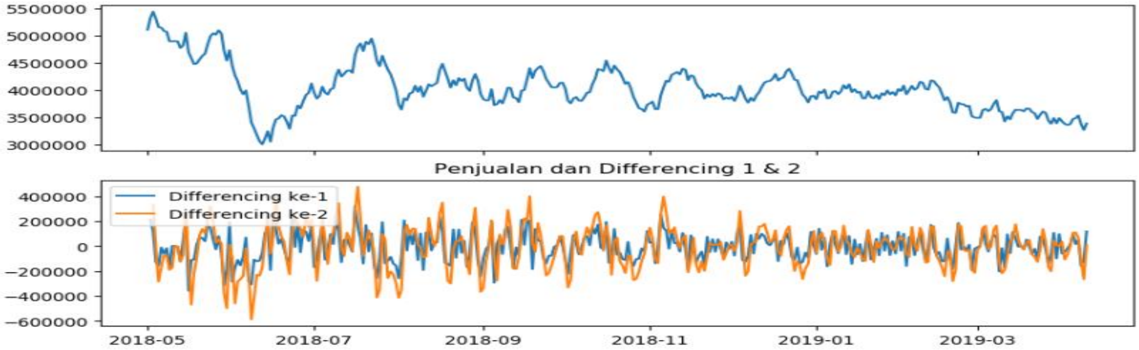

Figure 4. Smoothed data plots and differencing plots 1 & 2

3.4.2. ARIMA Modeling. From the ACF plot in Figure 5 and PACF in Figure 6 below, a provisional estimate of order 1 is used for AR and MA (*p, q*) (*1,1*) values because the first lag has entered the area of significance. The value of differencing (I) is temporary (1), so the ARIMA model p, d, q (1,1,1) is considered the best model so far.

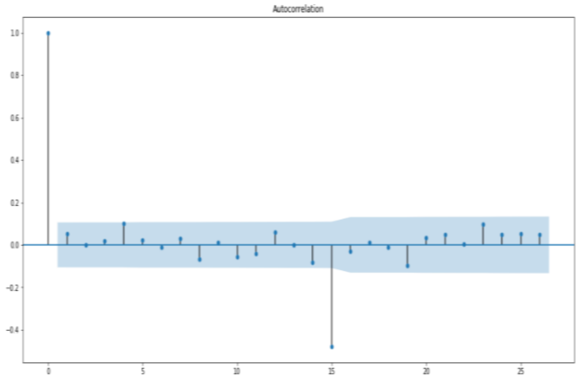

Figure 5. ACF plot

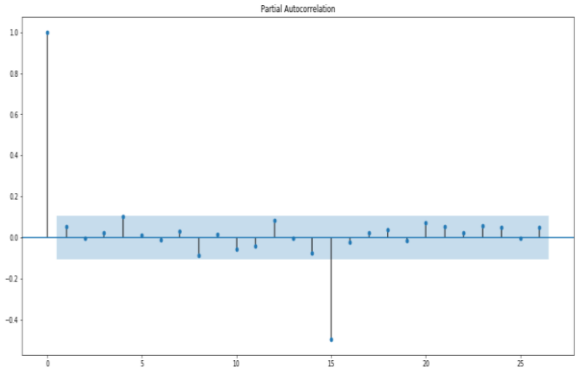

Figure 6. PACF plot

3.4.3. Estimation Test & Cross Validation. Cross validation is needed to determine the model's ability to forecast, validation in time series is not taken randomly but must be sequential, the sales data series is divided in half, the training data is 85% or 293 days and the test data is 15% or 52 days. The plot of ARIMA (1,1,1) with cross validation is shown in Figure 7.

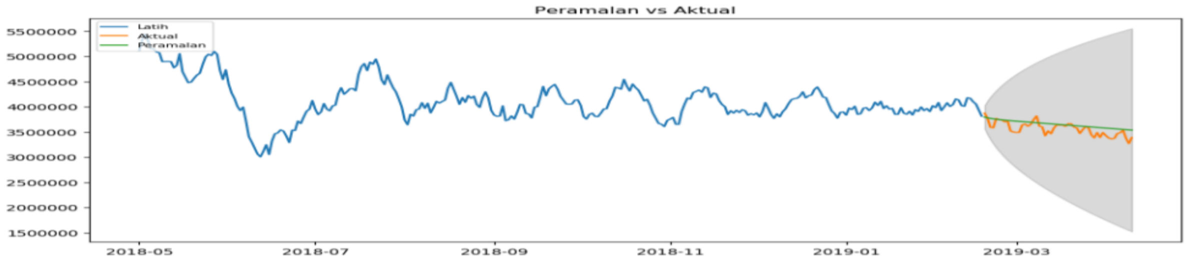

Figure 7. ARIMA forecast plot (1,1,1)

Table 2. Recapitulation of research results

Model	Value AIC	*P-Value*		Significance of	MAPE
ARIMA (1,1,1)	7662.06	AR (1)	0.001	Significant	0.02919
		MA (1)	0.004		
ARIMA (3,1,2)	7639.772	AR (1)	0.000	Significant	0.02561
		AR (2)	0.000		
		AR (3)	0.010		
		MA (1)	0.000		
		MA (2)	0.000		

The ARIMA model (3,1,2) has the smallest AIC value, a significant *P-Value* value and a fairly low MAPE value, so the ARIMA model (3,1,2) is considered the best model, Figure 8 is a plot of the ARIMA model (3,1,2) with cross validation 85% training data and 15% test data.

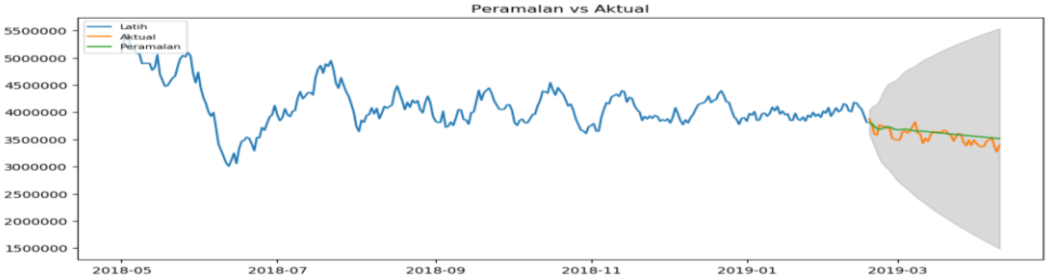

Figure 8.The plot of ARIMA forecasting (3,1,2)

3.4.4. Diagnostic Test

Standardized residual: Describes the error of a reflective residual around the zero mean and fairly uniform variance, as shown in Figure 9.

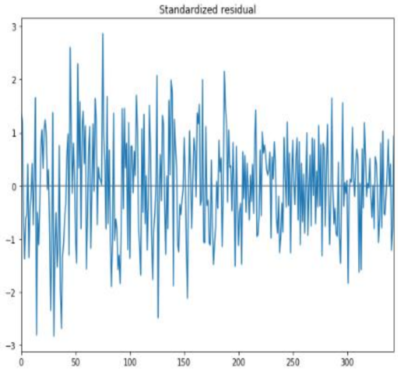

Figure 9. Standardized residual plot

Histogram plus estimated density: The residual density plot shows the normal distribution with zero mean, shown in Figure 10.

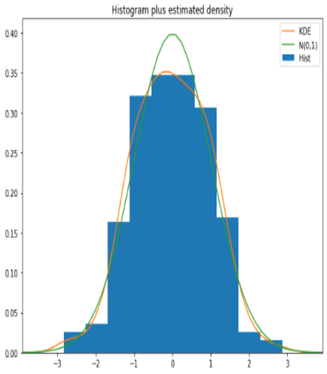

Figure 10. Histogram plus density plot

Normal Q – Q: The location of the residual points tends to follow the red line, as shown in Figure 11.

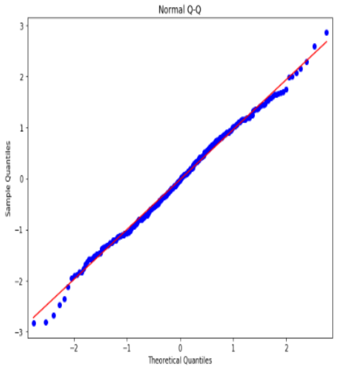

Figure 11. Plot Normal Q-Q

3.5. Forecasting

After the best model has been obtained, namely the ARIMA model (3,1,2), then forecasting for the next 15 days by looking at the trend based on the plot in Figure 12, it is estimated that the sales value is in the range of Rp. 3,300,000 to Rp. 3,500,000. With an RMSE value of 112883.14236, it is estimated that the sales value for the next 15 days will be in the range of Rp. 3,187,116 to Rp. 3,612,883.

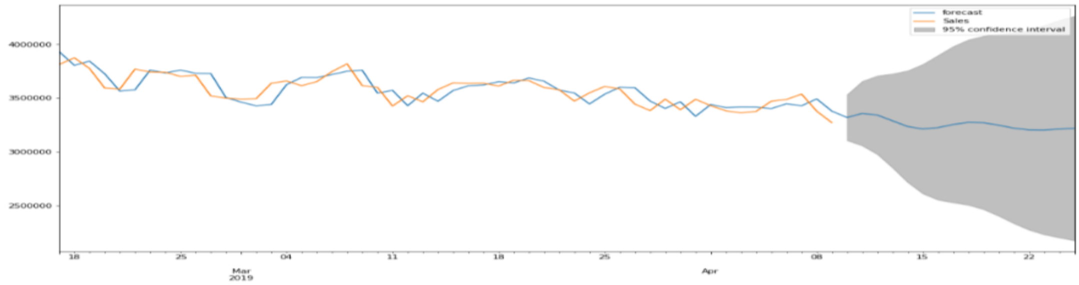

Figure 12. Plot of forecasting the next 15 days

4. Conclusions

There was a sales spike which caused sales data to fluctuate, especially at the beginning of the month of establishment. The first eight months of sales had a seasonal pattern of lag 15 but in the last four months sales tended to be flat and come down. Suggestions from this study include using other data mining methods that are useful for making decisions and future steps, such as association rules, shopping analysis and others, analyzing the alleged relationship between *supply* of goods and sales, and increasing member spending participation.

DAFTAR PUSTAKA

[1] S. H. Wahid, “Ketentuan Pengembalian Setoran Pokok Dalam Undang-Undang No.17 Tahun 2012 Tentang Perkoperasian (Perspektif Undang-Undang Dasar 1945 Dan Hukum Islam),” JURISDICTIE, 2017, doi: 10.18860/j.v5i2.4018.

[2] Abdullah, A. G., & Mulyadi, Y. (2016). Peramalan Beban Listrik Jangka Pendek Melalui Pendekatan Statistik dan Soft Computing. KNTIA, 2.

[3] Yusendra, M. A. E. (2015). Kajian Penerapan Metode Peramalan pada Ilmu Ekonomi dan Ilmu Komputer (Studi Kasus: Penerimaan Mahasiswa Baru Ibi Darmajaya). Prosiding Sembistek 2014, 1(01), 267-279. TIA, 2.

[4] Pavlyshenko, B. M. (2019). Machine-learning models for sales time series forecasting. Data, 4(1), 15. Doi: 10.3390/data4010015.

[5] Molydah, S. (2018). Analisis Perbandingan Implementasi Sarima Dan Support Vector Machine (Svm) Dalam Prediksi Jumlah Wisatawan Mancanegara.(Universitas Islam Indonesia, Yogyakarta)

[6] Widiyanto, S. (2019). Peramalan nilai impor besi baja Indonesia 2008-2017 dengan model arima. JURNAL MANAJEMEN, 11(2), 217-225.

[7] Wiyanti, D. T., & Pulungan, R. (2012). Peramalan Deret Waktu Menggunakan Model Fungsi Basis Radial (RBF) dan Auto Regressive Integrated Moving Average (ARIMA). Jurnal Mipa, 35(2).

[8] Ferguson, T. S. (1961, July). On the rejection of outliers. In Proceedings of the fourth Berkeley symposium on mathematical statistics and probability (Vol. 1, No. 1, pp. 253-287). Berkeley: University of California Press.

[9] Barnett, V., & Lewis, T. (1984). Outliers in statistical data-Osd.

[10] Sembiring, R. K. (1995). Analisis regresi. Bandung: ITB.

[11] Paludi, S. (2009). Identifikasi dan Pengaruh Keberadaan Data Pencilan (Outlier). Majalah Panorama Nasional, 56-62.

[12] Septiadi, M. N. K., Handayani, S., & Mulyadi, C. (2018). Penilaian Persediaan Barang Dagang Dengan Metode Rata-Rata Bergerak. EKA CIDA, 1(2).

[13] Yuanti, A. D. (2016). Perbandingan Model Time Series Seasonal Arima (Sarima) Dan Seasonal Arfima (Sarfima) Pada Data Beban Konsumsi

Listrik Jangka Pendek Di Jawa Timur (Studi Kasus di APD PT. PLN Distribusi Jawa Timur) (Doctoral dissertation, Airlangga University).

[14] Cryer, J. D., & Chan, K. S. (2008). Time series analysis: with applications in R. Springer Science & Business Media.

[15] Wei, W. W. (2006). Time series analysis. In The Oxford Handbook of Quantitative Methods in Psychology: Vol. 2.

[16] Azriati, K. F., Hoyyi, A., & Mukid, M. A. (2014). Verifikasi Model Arima Musiman Menggunakan Peta Kendali Moving Range (Studi Kasus: Kecepatan Rata-rata Angin di Badan Meteorologi Klimatologi dan Geofisika Stasiun Meteorologi Maritim Semarang). Jurnal Gaussian, 3(4), 701-710. ISSN: 2339-2541.

[17] Aziz, A. (2011). Analisis Critical Root Value pada Data Nonstasioner. CAUCHY, 2(1), 1-6. Doi: doi.org/10.18860/ca.v2i1.1794.

[18] Widodo, W. (2005). Metode Autoregresi dan Autokorelasi untuk meramalkan jumlah penjualan pakaian di toko Yuanita Purwodadi. Tugas Akhir Mahasiswa UNNES, Semarang.

[19] Krismiasari, S. (2012). Peramalan Produksi Padi Di Kabupaten Kampar Dengan Metode Box-Jenkins (Doctoral dissertation, Universitas Islam Negeri Sultan Syarif Kasim Riau).

[20] Sugiarto, S. (2017). Penduga Model Arima Untuk Peramalan Harga Tbs Kelapa Sawit Di Propinsi Riau. Jurnal Sains dan Teknologi Industri, 15(1), 35-40.

[21] Hanke, J. E., Reitsch, A. G., & Wichern, D. W. (2001). Business forecasting (Vol. 9). New Jersey: Prentice Hall.

[22] Makridakis, S. (1993). Dkk. Metode dan Aplikasi Peramalan, Jakarta: Airlangga.

[23] A. H. Hutasuhut, “Pembuatan Aplikasi Pendukung Keputusan untuk Peramalan Persediaan Bahan Baku Produksi Plastik Blowing dna Inject Menggunakan Metode ARIMA (Autoregressive Integrated Moving Average) di CV. Asia,” J. Tek. Pomits, vol. 3, no. 2, pp. 70–171, 2014, [Online]. Available: http://ejurnal.its.ac.id/index.php/teknik/article/viewFile/8114/1846.

[24] Hatidja, D. (2011). Penerapan Model Arima Untuk Memprediksi Harga Saham PT. Telkom Tbk. Jurnal Ilmiah Sains, 11(1), 116-123. Doi: 10.35799/jis.11.1.2011.53.

[25] Pamungkas, M. B., & Wibowo, A. (2019). Aplikasi Metode ARIMA Box-Jenkins Untuk Meramalkan Kasus Dbd Di Provinsi Jawa Timur. The Indonesian Journal of Public Health, 13(2), 183.doi: 10.20473/ijph.v13i2.2018.183-196.

[26] Sutanto, P., Setiawan, A., & Setiabudi, D. H. (2017). Perancangan Sistem Forecasting di Perusahaan Kayu UD. 3G dengan Metode ARIMA. Jurnal Infra, 5(1), 325-330.

[27] Wibowo, A. (2018). Model peramalan indeks harga konsumen kota Palangka Raya menggunakan Seasonal ARIMA (SARIMA). Matematika, 17(2).

[28] Rahmadayanti, R., Susilo, B., & Puspitaningrum, D. (2015). Perbandingan Keakuratan Metode Autoregressive Integrated Moving Average (ARIMA) dan Exponential SMOOTHING Pada Peramalan Penjualan Semen di PT. Sinar Abadi. Rekursif: Jurnal Informatika, 3(1).

[29] Yuniarti, A. (2010). Perbandingan metode peramalan eksponensial smoothing dan Box-Jenkins (ARIMA) musiman (Doctoral dissertation, Universitas Islam Negeri Maulana Malik Ibrahim).

[30] Ul Ukhra, A. (2014). Pemodelan dan peramalan data deret waktu dengan metode SEASONAL ARIMA. Jurnal Matematika UNAND, 3(3).

[31] Linda, P., Situmorang, M., & Tarigan, G. (2014). Peramalan Penjualan Produksi Teh Botol Sosro Pada PT. Sinar Sosro Sumatera Bagian Utara Tahun 2014 Dengan Metode Arima Box-Jenkins. Saintia Matematika, 2(3), 253-266.

[32] As’ad, M., Wibowo, S. S., & Sophia, E. (2017). Peramalan Jumlah Mahasiswa Baru Dengan Model Autoregressive Integrated Moving Average (Arima). JIMP-Jurnal Informatika Merdeka Pasuruan, 2(3)., doi: 10.37438/jimp.v2i3.77.

[33] Raihan, R., Effendi, M. S., & Hendrawan, A. (2016). Forcasting Model Exsponensial Smoothing Time Series Rata Rata Mechanical Availability

Unit Off Highway Truck CAT 777D Caterpillar. POROS TEKNIK, 8(1), 1-9.

[34] Aprilia, D. (2016). Perbandingan Metode Peramalan Exponential Smoothing Dan Moving Average (Studi Prediksi Jumlah Penderita TB Paru di Provinsi Jawa Timur) (Doctoral dissertation, Universitas Airlangga).

[35] Setiawan, E., Murfi, H., & Satria, Y. (2016). Analisis Penggunaan Metode Kernel Density Estimation pada Loss Distribution Approach untuk Risiko Operasional. Jurnal Matematika Integratif ISSN, 1412, 6184.

[36] Putri, R. M., & Widodo, E. (2018). Application of Support Vector Machine Method For Rupiah Exchange Rate To Us Dollar Forecasting. In PROSIDING SEMINAR NASIONAL & INTERNASIONAL (Vol. 1, No. 1).

[37] Naufal, M. F. (2017). Peramalan Jumlah Wisatawan Mancanegara Yang Datang Ke Indonesia Berdasarkan Pintu Masuk Menggunakan Metode Support Vector Machine (SVM) (Doctoral dissertation, Institut Teknologi Sepuluh Nopember).

[38] Samsiah, D. N. (2008). Analisis Data Runtun Waktu Menggunakan Model Arima (p, d, q). Program studi Matematika Fakultas Sains dan Teknologi. UIN Sunan Kalijaga. Yogyakarta.