Predictions of new students interest in entering the XXX University with statistic modeling

E D Priyana

Department of Industrial Engineering, Faculty of Engineering, Universitas Muhammadiyah Gresik, Gresik, Indonesia

Email: eftadhartikasari@umg.ac.id

Abstract. XXX is the establishment of one of Muhammadiyah's charitable businesses in the field of education. From the beginning of its formation on May 25, 1980 until now XXX has been able to open 18 study programs plus other study programs at the Faculty of Health and the Faculty of Engineering which are planned to be opened. The addition of the study program is expected to filter the number of new student interests to become XXX residents. But unfortunately, the many universities in Gresik were opened, making XXX have to work extra in attracting new students. This research will only be focused on single forecasting on the number of new student interests in several study programs and one university. In the descriptive statistical results, it was obtained the results that Management Study Program contributed the most interest in entering students and as Study Programs which were not new, Fisheries Cultivation Study Program contributed the least interest of students entering. The results of cluster analysis were formed by 8 groups. Group 5 consisting of Management Study Program, Industrial Engineering and PGSD contributed the most interest in terms of the average count per study program. Forecast with Moving Average also proves that this group has a tendency to rise. ARIMA results (1 0 0) for one University produces a value that falls smoothly with a less significant value, but if left unchecked it can also harm the campus.

1. Introduction
XXX is one of Muhammadiyah's charitable endeavors in the field of education. XXX began to be inaugurated by the regent of KDH Level II in Gresik Regency, Mr. Colonel Wasiadji with a location on Jalan Kh.Kholil No.90 Gresik, on May 25, 1980 under Number: E.1 / 017-V / 1980 [1, 2]. From the beginning only opened 4 Faculties (Faculty of Agriculture, Faculty of Fisheries, Faculty of Economics, Faculty of Teacher Training and Education) at the beginning of its formation, XXX has now been able to add to its Study Program by grouping several Faculties (Faculty of Agriculture, Faculty of Economics, Faculty of Teacher Training and Education (FKIP), Faculty of Islamic Religion, Faculty of Engineering, Faculty of Psychology, Diploma Program) [2, 3].

Predictions of new students 'interest in entering the XXX university with statistic modeling is a title carried by the author in the hope of getting a forecast with minimal error value, which means the predicted value obtained will get closer to perfect. Statistics itself is a tool that has long been known by the general
public. Prediction with the smallest error is the best model that can be used for many things. The first is to analyze whether the number of new students who enter XXX is in line with the targets management wishes to achieve. The second is for further research, is opening a new study program felt necessary for student worship. The results of this study are also expected to be an alternative reference for other researchers or writers as further research. [4].

The statistical method used to predict interest in entering new students in XXX University is the Autoregressive Integrated Moving Average (ARIMA) forecasting model which is known to have high accuracy in predicting univariate data. This study will use data on the number of new students who entered from the beginning of the XXX University standing until the last available data. Not only predicting the whole student entering XXX university, the researcher also presents descriptive domain analysis of several study programs, with grouping using cluster analysis (the least group and the most group) so that XXX university management can consider appropriate sustainability management to strengthen the university [5, 6].

The ARIMA method studied was developed by George Box and Gwilyn Jenkins in 1976 (Makridakis et al, 1999) [7]. This method has the concept of ignoring the predictor variables in preparing the forecast model. The ARIMA approach is believed to be able to improve the accuracy of the predictions produced. This trust in research on ARIMA has made ARIMA’s concepts contribute a lot in various fields. One of the studies was conducted by Bagoes Rahmat Widiarso in 2012, which used ARIMA as the basic concept of forecasting research in rainfall in Ngawi District with the result that the Post Mantingan showed poor performance, so special handling was needed [8].

2. ARIMAZ

Autoregressive Integrated Moving Average (ARIMA) is one of the models used for forecasting time series data. ARIMA is a non-stationary ARMA model that has been differentiated so that it becomes a stationary model. The assumptions that must be met are the stationarity of the data and residuals which are white noise. White noise which means residuals do not autocorrelate and have a normal distribution of $N(0, \sigma^2)$.

The models in the non-seasonal ARIMA are as follows [9, 10].

a. Autoregressive Model (AR (p))
   
The Autoregressive (AR) model with order p denoted by AR (p) with the general form of the AR (p) model:
   
   $Z_t = \theta_1 Z_{t-1} + \theta_2 Z_{t-2} + \cdots + \theta_p Z_{t-p} + a_t$  \hspace{1cm} (1)
   
   b. Model Moving Average (MA(q))
   
The Autoregressive Model (AR (p)) with the order q is denoted by MA (q) with the general form of the MA (q) model:
   
   $Z_t = \alpha_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \cdots - \theta_q a_{t-q}$  \hspace{1cm} (2)
   
   c. Model Autoregressive Moving Average (ARMA(p,q))
   
   ARMA model with order p and q is a mixed model between autoregressive order p and moving average order q. The general models:
   
   $Z_t = \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + \cdots + \phi_p Z_{t-p} + \alpha_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \cdots - \theta_q a_{t-q}$  \hspace{1cm} (3)

   Model Autoregressive Integrated Moving Average (ARIMA(p,d,q))
   
   The general model autoregressive order p, Integrate order d, and moving average order q (ARIMA (p, d, q)) is the result of a merger between a stationary process with a stationary non-stationary product. In this case, d is the order of differntiation. The forms of the ARIMA model (p, d, q) are:
   
   $\phi(B)(1-B)^d Z_t = \theta(B) a_t$  \hspace{1cm} (4)
   
   $\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p$  \hspace{1cm} (5)
   
   $\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q$  \hspace{1cm} (6)
3. Descriptive statistics

In this study, the beginning of things done to find out the data description is to make descriptive statistics. Descriptive statistics used are by making bar charts, to determine the prospect of study programs that have the most amount of interest to the smallest. Figure 1 shows the results that Management Program study is a study program that is able to absorb the most students. While the study program that stood for a long time but did not absorb many students was the Fisheries Cultivation Study Program (Prodi Budidaya Perikanan).

Figure 2 shows an overview of the number of incoming interests per period. In 2007/2008 it was seen that Informatics Engineering Study Program was the most popular study program. But in the 2012/2013 period the amount of entry interest tended to drop significantly and began to stabilize in 2015. In contrast to the Management Study Program, which used to be less attractive in early 2007/2008, its presence in 2018/2019 actually contributed to many interested people.

Figure 3 shows prodi groups are less desirable. It can be seen that the Study Program for Aquaculture is less desirable, although it continues to move up smoothly, the numbers are very small or even can be said to be not very meaningful. This is not in line with the environment of Gresik which is famous for the cultivation of ponds that this study program should be able to become a doctrine for the Gresik region.

Figure 4 show the output of cluster analysis. 8 groups will be formed. The number of groups is prioritized to find out the tendency of homogeneity of more significant characteristics [11-14]. Table 1 is

![Figure 1. The average number of students entering interest per study program](image1)

![Figure 2. Time guide of the most interested study programs](image2)

![Figure 3. Time-consuming study program not interested](image3)
the result of cluster analysis. It was found that Islamic Religious Education program study and Agrobiotechnology program study formed a single group, which means that the two study programs did not have homogeneous characteristics in filtering the amount of student interest to enter with any study program.

![Dendrogram cluster analysis](image)

**Figure 4.** Dendrogram cluster analysis

| Program Study                  | Group | Program Study                  | Group |
|-------------------------------|-------|-------------------------------|-------|
| S2 Management                 | I     | Management                    | V     |
| S2 English Language Education |       | Industrial Engineering        |       |
| Shipping Technique            |       | PGSD                          |       |
| Entrepreneurship              | II    | Electrical Engineering        | VI    |
| PIAUD                         |       | Psychology                    |       |
| Law                           |       | Agrobiotechnology             | VII   |
| Islamic Education             | III   | Mathematics Education         | VIII  |
| Accounting                    | IV    | Technical Information         |       |
|                               |       | English Language Education    |       |

Figure 5 is an actual data plot without groups 1 and 2, because both groups are new study programs that do not have enough data to be analyzed using time series analysis. It can be seen that group V is the most absorbing group of new students. It is inversely related to group VII which is a less desirable study group, and group VIII which has an interest rate that tends to go down in a certain period [15].
Figure 6. shows the results of forecasting with the Moving Average (MA) method without including groups I and II because of the limitations of existing data. The figure shows the plot of group V which tends to increase significantly compared to other groups, while group VI moved slightly up but not significantly [16, 17].

Figure 7 is a time series of data plots of actual student interest as a whole. The graph shows fluctuating but less irregular values. In the 2007/2008 period, the interest in entering new students tended to increase significantly and in 2012 the plot tended to be irregular and even decreased [18-20].

4. ARIMA

The assumptions that must be fulfilled in ARIMA are stationary in mean and stationary in variance. This is not found in traditional forecasting models such as Moving Average. It is this consideration of stationarity that makes ARIMA more accurate.

Figure 8 the transformation results show that the rounded value is 1 which means the data is stationary in the variance. Next is the analysis of stationarity in the mean by looking at the ACF value (Figure 9). The way to determine stationarity in the mean is to look at the lag that is out of line. If there are more than 3 lags out of line then the data is considered not stationary in the mean. Since there is no data coming out, it can be ensured that the data is stationary in the mean.
Table 2 is the result of the ACF test with a 95% confidence interval. The comparison of the decision area used by $T_{count}$ adalah $T_{table}$ ($\alpha = 0.05; \text{db} = 11$) with value of 1.796. T test statistics are used to detect lags to $z_t$ with $z_{t-k}$ (k is lag).

**Hypothesis:**

$H_0: \rho_k = 0$

$H_1: \rho_k \neq 0$

Table 2 is output T value in lag 1 with value 2.25 $> 1.796$ which means reject $H_0$, so it can be concluded that there is a significant correlation (between $z_t$ and $z_{t-1}$) in this lag. Besides lag 1, the lag value is less than 1.796, which can be concluded that there is no significant correlation in the other two lags.

**Table 2.** Autocorrelation function on the amount of interest in entering new students of XXX University

| Lag | ACF  | T   | LBQ |
|-----|------|-----|-----|
| 1   | 0.649539 | 2.25 | 6.44 |
| 2   | 0.370223 | 0.94 | 8.75 |
| 3   | 0.172266 | 0.41 | 9.30 |

Figure 10 visually, the PACF plot shows that the data on the amount of interest in entering new University XXX students is stationary in the mean. Seen from no lag coming out of the line. Table 4 is the value that shows the T test for hypothesis testing with a 95% confidence interval. Similar to the ACF test, the decision area used as a comparison for $T_{value}$ is $T_{table}$ ($\alpha = 0.05; \text{db} = 11$) with a value of 1.796. T test statistics are used to detect lags to $z_t$ with $z_{t-k}$ (k is lag).

**Figure 10.** ACF plot data transformation of incoming student interest
Table 3 is the Output T value in lag 1 with a value of 2.25 > 1.796 which means rejecting H<sub>0</sub>, so it can be concluded that there is a significant correlation (between z<sub>t</sub> and z<sub>t-1</sub>) in this lag. Besides lag 1, the lag value is less than 1.796, which can be concluded that there is no significant correlation in the other two lags.

Table 4 shows the output results from the ARIMA model (1 0 0). The model is the best result that has the smallest P-value value from several experiments that have been carried out. There is a limitation of the data obtained so that in ARIMA forecasting (1 0 0) the data is doubled twice, so that it has the number of n as many as 24, which was previously only 12.

### Table 3. Partial autocorrelation function on the interest of new students entering the XXX University

| Lag | PACF   | T    |
|-----|--------|------|
| 1   | 0.649539 | 2.25 |
| 2   | -0.089393 | -0.31 |
| 3   | -0.055173 | -0.19 |

### Table 4. Output ARIMA (1 0 0)

| Type  | Coef  | SE Coef | T    | P    |
|-------|-------|---------|------|------|
| AR 1  | 0.6166 | 0.1698  | 3.63 | 0.001|
| Constant | 396.50 | 47.41   | 8.36 | 0.000|
| Mean   | 1034.1 | 123.6   |      |      |

**Hypothesis:**

H<sub>0</sub>: φ<sub>1</sub> = 0  
H<sub>1</sub>: φ<sub>1</sub> ≠ 0

ARIMA (1 0 0) will produce AR (1) values that involve constant values, so we get 2 parameters in the model. In general, the significance of a constant does not need to be tested so only the autoregressive parameter (φ).

**Hypothesis:**

H<sub>0</sub>: φ = 0  
H<sub>1</sub>: φ ≠ 0

The same is true in the case of testing the ACF and PACF hypotheses. The comparison used in the T-hitung decision area is T-table. Normal comparison tables cannot be used because the analyzed data falls into the small sample category. Table 4 is the result of interpretation of the T<sub>value</sub> for the AR parameter (1) or φ is 3.63 greater than the value of T<sub>table</sub> of 1.717 and P<sub>value</sub> < α. So it can be concluded that the AR (1) model, parameter φ is quite significant. The resulting model equation is:

Z<sub>t</sub> = 0.6166Z<sub>t-1</sub> + 396.5

The data processing also mentions the results of the Mean Square Error (MSE) model of 53929 with a free degree of 22. Table 5 is the Ljung-Box value for detecting white noise processes. Several tests will be conducted, one of which is the correlation test used to detect residual independence and next is the residual normality test model. Residual independence test is performed to detect residual independence between lags. Time series uses this test in Ljung-Box-Pierce.
Hypothesis

There is a correlation between lags

Ljung-Box-Pierce statistical test uses a comparison between $X^2_{\text{value}}$ and $X^2_{\text{table}}$. Reject $H_0$ if the value $X^2_{\text{value}} > X^2_{\text{table}}$. Table 5 shows the Ljung-Box value only at lag 12, this is due to the limited data available so it is not possible to get more data. The value of $X^2_{\text{value}}$ shows the calculation result of 30.5 and the value of $X^2_{\text{table}}$ of 18.3 which means reject $H_0$. If the decision is taken from the side of the $P_{\text{value}}$, the result is less than the error value (5%) so it can be decided that the decline in $H_0$ can be interpreted that between the residuals in lag t with the next lag has a relationship. This is in line with the situation that the output produced only shows lag 12.

| Table 5. ARIMA white noise identification (1 0 0) |
|-------------------------------------------------|
| Lag 12 14 36 48                               |
| Chi-Square statistic                          |
| 30.5  *  *  *                                |
| DF 10  *  *  *                               |
| P-value 0.001  *  *  *                       |

Figure 11 shows the results that no 1 lag was out of bounds, so it can be concluded that the data has been independent. The lack of data variability in the Ljung Box makes the test not yet consistent. The analysis that has been done shows that the best model for the ARIMA test is ARIMA (1 0 0) selected based on the $P_{\text{value}}$. Having succeeded in the variant also meant that it was agreed to analyze ARIMA. A forecast of 10 years is produced in Figure 4 which has been compared with the overall MA results.

Analysis that has been done shows that the best model for the ARIMA test is ARIMA (1 0 0) which is chosen based on the smallest $P_{\text{value}}$. After stationary success in the variance or mean, the ARIMA analysis test will be performed. The forecast for the next 10 years in Figure 4 has been compared with the overall MA results.

Figure 12 shows the results of MA forecasts that tend to be irregular fluctuations. This can be identified because MA has many shortcomings in terms of correlation between series due to the absence of a stationary assumption in variance or mean. While the ARIMA results show a smooth decline in value. ARIMA testing is more reliable because of the close correlation between time series.

Figure 11. The ACF plot of the AR (1) model

Figure 12. Comparison MA VS ARIMA
5. Conclusion
The conclusion obtained from this analysis is that in the case of ARIMA forecasting (1 0 0) the resulting value that in the next 10 periods the number of students entering interest in XXX University will decrease smoothly or may be interpreted to have a less significant value but tends to continue to fall. While in group forecasting with MA, it is produced the value that group 5 which consists of Management Study Program, Industrial Engineering and PGSD has the best forecast among the other groups and it can be assumed that the three study programs can contribute the most new students among other study programs.

References
[1]. Muhammadiyah Central Management 2000 Guidelines for the Islamic Life of Muhammadiyah Citizens Jakarta.
[2]. Team UMG 2017 Academic Guidelines Gresik.
[3]. Asmaria I and Ponirin 2013 Development of Charitable efforts of the Muhammadiyah Organization Medan Journal of Government Science and Social Politics UMA.
[4]. Montgomery, Douglas, Jennings and Kalahci. 2008 Indtroduction To Time Series Analysis and Forecasting America: A John Wiley & Sons Inc.
[5]. Namin, Sima, Namin and Akbar 2018 Forecasting Economic And Financial Time Series: ARIMA VS LSTM. Lubbock USA Texas university.
[6]. Iwok I A and Udoh G M 2018 A Comparative study between the ARIMA – Foutier model and the wavelet model American Of Scientific And Industrial Research. 7-6 pp 137-44
[7]. Wei W W 2006 Time Series Analysis Univariateand Multivariate Methods AmerikaSerikat: Pearson Education Inc.
[8]. Wahyudi 2011 Identification of Extreme Rainfall in Ngawi District Using Generalized Extreme Value and Generalized Pareto Distribution. Surabaya: ITS Press. Walpole R E 1996 Pengantar Statistika Jakarta: PT. Gramedia Pustaka Utama.
[9]. Rachman Rizal 2018 Application of Moving Average and Exponential Smoothing Methods in Garment Industry Production Forecasting Jakarta: STMIK Nusa Mandiri.
[10]. Satya Febi 2015 Use of the ARIMA Method for Short-term Electric Consumption Load Assembling. Semarang: Universitas Negeri Semarang.
[11]. Sukmawati. 2017 Cluster Analysis with Hierarchy Method for Regency / City Grouping in South Sulawesi Province Based on Macroeconomic Indicators.
[12]. Hamidi, Seyed. 2018 Consensus clustering algorithm based on the automatic partitioning similary graph Science Direct Journal. Data & Knowledge Engineering.
[13]. Khadir and Khedaira. 2019 A multiple clustering combination approach based on interative voting process Journal of King Sand University-Computer and Information Sciences.
[14]. Socoro, Joan, Alias and Francesc 2019 Parallel hierarchial architectures for efficient consensus clustering on big multimedia cluster ensembles Journal Information Sciences vol 511 pp 212-28.
[15]. Walder, Adam, & Hanks, Ephraim 2019 Bayesian analysis of spatial gneralized linier mixed models with laplace moving average random fields Journal Pre-Proof.
[16]. Heinrich C, Mikko S P and Almut E D V 2019 Hybrid simulation scheme for volatility modulated moving average fields Journal Mathematics and Computers In Simulation vol 166 pp 224-44.
[17]. Huang, Jing, Huang, Zhijian 2019 Testing moving average trading strategies on ETFs Journal of Empirical Finance.
[18]. Zifwen 1999 ENSO forecasting and modeling the relationship of ENSO with Monson rainfall Skripsi Jurusan Statistika Bogor: FMIPA IPB.
[19]. Fanoodi B, Behnam M and Farzad F J 2019 Reducing demand uncertainty in the platelet supply chain though artificial neural networks and ARIMA models *Computers In Biology and Medicine* **113**(2019) 103415.

[20]. Wang Q, Shuyu L and Rongrong L 2018 China’s dependency on foreign oil will exceed 80% by 2030: Developing a novel NMGM-ARIMA to forecast china’s foreign oil dependence from two dimensions *Science Direct Journal. Energy* **163**(2018) pp 151-67.