Journal of Actuarial, Finance and Risk Management (JAFRM)
Vol. 1(1), June, 2022.

Forecasting of Retirement Insurance Filled via Internet by ARIMA Models

Lovena Louisa[1], Rifky Fauzi[2], Edwin Setiawan Nugraha[3*]

[1,3]Study Program of Actuarial Science, School of Business, President University, 17550, Indonesia
[2]Department of Mathematics, Institut Teknologi Sumatera, 35365, Indonesia
**Corresponding author: edwin.nugraha@president.ac.id*

Abstract— Pension fund insurance is critical for everyone because it can guarantee a good life during retirement because retirement is a period when someone no longer gets a steady income. Technological advances make it easier for retirement insurance applications. By using ARIMA Models, we can predict the number of internet users who apply for retirement insurance via the internet, using the monthly data of the Social Security Administration from January 2008 to October 2020. The data used has a steady increasing trend with some seasonal components, so it needs to be removed first. ARIMA models use the assumption that the data is stationary, so the data must be tested using the ADF test command in R. After seeing the plotting of ACF and PACF, 9 ARIMA models are formed. ARIMA model is selected based on the smallest AIC. By using 95% confidence it can be concluded that ARIMA (9,1,9) is the best model for forecasting.

Keywords— ACF; ADF Test; ARIMA Models; Forecasting; PACF; Pension fund; Seasonal.

I. INTRODUCTION

As we know, pension fund insurance is critical for employees of a company. It is essential because the insurance can protect the employees from unexpected expenses when entering retirement. The other reasons why pension fund insurance is essential:

1. When entering the retirement period, an employee is no longer working. In other words, the employee has no income to fulfill their needs.
2. Having no income does not mean that an employee does not have an expense. In this situation, they must have enough savings or have some investment.
3. Their savings are very limited in amount and will run out if they use it as the only source to cover your expenses.
4. As they get older, some needs, such as health needs, will also increase. That means the cost of their needs will also increase.

When employees do not have pension fund insurance, they have to properly manage their expenses so that the savings that have been collected for retirement do not run out. If their savings are not sufficient for your daily needs, then they are forced to find work at their retirement age. As mentioned before, everyone certainly needs pension fund insurance.

Applying for offline pension fund insurance may be a long process and requires a higher cost than filling online. Therefore, day by day, people are starting to turn to submit via the internet because it is easier and more practical. Consider that people get online application facilities as good as offline applications; we must first predict the number of prospective policyholders who will submit their policies online.

To predict the future number of policyholders who will request their policies via internet applications, we will use ARIMA models as our analysis method. ARIMA (Autoregression Integrated Moving Average) models are commonly used in forecasting and seasonal adjustment of stationary data. With some adjustment, the ARIMA models can be used to predict non-stationary time-series data, which is mostly found in financial time series data. The application of this method can be seen in [1]–[3].

The result of this research will be shown in a table and graphic form that will be validated and verified manually based on the data. Forecasting with ARIMA models was chosen because it has the advantage such as having a great analysis in a random situation, trend, seasonal, and observed in a period of data that have been analyzed. The processes are done by using R Studio software.

II. Literature Review

A. Time Series Analysis

Time Series data is a set of data over a certain period of time. Time series forecasting is forecasting based on the behavior of past data to be projected into the future by utilizing mathematical equations and statistics. Time series data types are divided into several types, among others [4]:

1) *Cycle*

The cycle pattern is a series of changes up or down, so that this cycle pattern changes and varies from one cycle to the next. Cyclical patterns and irregular patterns are obtained by eliminating trend patterns and seasonal patterns if the data used are weekly, monthly, or quarterly. If the data used is annual data, what should be eliminated is the trend pattern only.

2) *Random*

An irregular random pattern, so it cannot be described. These random patterns are caused by unforeseen events such as wars, natural disasters, riots, etc. Because the shape is irregular or does not always occur and cannot be predicted, the random variation pattern in the analysis is represented by an index of 100% or equal to 1.

3) *Trend*

Trend or trend is a long-term component that has a certain trend in the data pattern, either in an increasing or decreasing direction from time to time, so that the long-term trend pattern rarely shows a constant pattern. Techniques that are often used to obtain the trend of a time series data are linear moving averages, exponential smoothing, and the Gompertz model, where these techniques only use past data to get the trend pattern and do not take into account other factors that influence product demand.

4) *Seasonal*

The seasonal pattern shows a movement that repeats itself from one period to the next on a regular basis. This seasonal pattern can be indicated by data that are grouped on a weekly, monthly, or quarterly basis, but for data in the form of annual data, there is no seasonal pattern. This seasonal pattern should be calculated on a weekly, monthly, or quarterly basis depending on the data used for each year, and this seasonal pattern is expressed in numerical form. The technique used to determine the value of the seasonal pattern is the moving average method, winter's exponential smoothing, and classical decomposition.

B. ARIMA Forecasting

ARIMA (Autoregressive Integrated Moving Average) was first developed by George Box and Gwilym Jenkins for modeling time series analysis. ARIMA is often called Box-Jenkins models. ARIMA represents three models, namely the autoregressive model (AR), moving average (MA), and autoregressive and moving average model (ARMA) [5]. The stage of implementation in the search for the model are:

1) Identification of temporary model using past data to obtain models from ARIMA. The identification stage is carried out by observing the ACF (Autocorrelation Function) and PACF (Partial Autocorrelation Function) estimation patterns obtained from the data which are then used to obtain model predictions that match the data pattern.
2) Interpretation or the parameter estimation of the ARIMA model using past data.
3) Diagnostic testing to test the feasibility of the model. If the model is not feasible, then carry out the identification, estimation, diagnostic testing steps to get a proper model.
4) Application, namely forecasting the future value of the series data using the tested method.

C. Autocorrelation Function (ACF)

ACF is the correlation between data in time t with the previous time period $t-1$. The mean and variance of a periodic series data may not be useful if the series is not stationary, however, the maximum and minimum values can be used for plotting purposes. How the key statistic in the periodic series analysis is the autocorrelation coefficient.

$$\bar{Y} = \frac{\sum_{t=1}^{n-k} (Y - \bar{Y})(Y_{t+k} - \bar{Y})}{\sum_{t=1}^{n} (Y_t - \bar{Y})^2} \tag{1}$$

for *k* = 1, 2, …, *n*.

$$\bar{Y} = \frac{\sum_{t=1}^{n} Y_t}{n} \tag{2}$$

r_k is the estimator of ρ_k

D. Partial Autocorrelation Function (PACF)

Partial autocorrelation is used to measure the level of intelligence between X_t and X_{t-k}, if the effect of the lag time is considered separate. The only purpose of the periodic series analysis is to help determine the correct ARIMA model.

$$r_{kk} = \begin{cases} r_1 \\ \dfrac{r_k - \sum_{j=1}^{k-1} r_{k-1,j} r_{k-j} \qquad , \text{If } \ k = 1}{1 - \sum_{j=1}^{k-1} r_{k-1.j} r_{-j.} \qquad , \ \text{if } \ k = 2, 3} \end{cases} \tag{3}$$

$$PACF = \phi_{k_k} = Corr(Y_t, Y_{t-k} \mid Y_{t-1}, Y_{t-2}, ..., Y_{t-k+1}) \tag{4}$$

Based on the Yule-Walker equation:

$$\rho_j = \phi_{k_1} \rho_{j-1} + \phi_{k_2} \rho_{j-2} + ... + \phi_{k_k} \rho_{j-k} \tag{5}$$

where ρ_k denoting ACF, ϕ_{k_k} denoting PCAF, $\hat{\phi}_{k_k}$ denoting estimator of ϕ_{k_k} for $j = 1, 2, ..., k$. Note that $\rho_j = \rho_{-j}$ and $\rho_0 = 1$

E. Autoregressive Model (AR)

Autoregressive (AR) is an observation at time t expressed as a linear function with respect to the previous time p plus a random residual at that white noise, which is an independent and normal distribution with $mean = 0$ and $variance = \sigma^2$. Determination of the partial autocorrelation coefficient is used to measure the closeness level between Y_t and Y_{t-k} if the effect of time lag $1,2,3, \dots, k$. The purpose of using partial autocorrelation coefficients in periodic series data analysis is to help determine the right ARIMA model for forecasting, especially to determine the p order of the AR model (p) [6].

$$X_t = e_t + \phi_1 X_{t-1} + \phi_2 X_{t-2} + ... + \phi_p X_{t-p} \tag{6}$$

F. Moving Average (MA)

The autocorrelation coefficient is the same as a correlation coefficient. The difference lies in the coefficient of this autocorrelation which describes the relationship (association) between the values of the same variable but different periods. Autocorrelation provides important information about the arrangement or structure and pattern of data. The autocorrelation function is useful for finding correlations between data & is useful for determining the order *q* on MA (*q*).

$$X_t = e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - ... - \theta_q e_{t-k} \tag{7}$$

Moving Average Order 1:

$$X_t = e_t - \theta e_{t-1} \tag{8}$$

G. Autoregressive and Moving Average (ARMA)

The ARMA model of the order *p* and *q* (AR (*p*) and MA (*q*)) is a combination of the Autoregressive Model (AR) and the Moving Average (MA) shown in the formula below:

$$Y_t = e_t \phi_1 X_{t-1} + \phi_2 X_{t-2} + ... + \phi_p X_{t-p} - \theta_1 e_{t-1} - \theta_2 e_{t-2} - ... - \theta_q e_{t-k} \tag{9}$$

H. Stationary and Non-Stationary Data in Time Series

According to [7] stationary means that there is no growth or decline in data. That is, the data fluctuations are around a constant average value, independent of the timing and variance of these fluctuations. The visualization form of a time series data plot is often sufficient to determine whether the data is stationary or non-stationary. ACF and PACF plots also support for use in checking the instability of data. ACF plots that tend to be slow or linearly decrease indicate that the data is not stationary in the mean. According to [7], the values of stationary data autocorrelation will drop to zero after the second or third lag, while for non-stationary data, autocorrelation values differ significantly from zero for several time periods. Data stationary testing is indispensable for the time series data method. Forecasting can be done when the data conditions are stationary. Data that is not stationary has 3 conditions, namely not stationary in variance, not stationary in the mean, or not both stationary. Data that is not stationary in variance can be transformed to make the data stationary. Data that is not stationary in the mean needs to be differentiated (differencing) to stationary it. Both of the data are not stationary, transformed, and differentiated (differencing) to make it stationary.

I. Forecasting

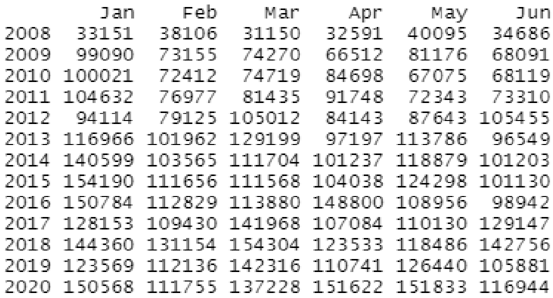
Forecasting is one of the methods for doing the planning and control to face uncertain moments in the future. Forecasting is the use of past data from a variable or collection of variables to estimate its value in the future. There are 2 methods of forecasting, qualitative and quantitative methods [5]. The qualitative method is based on opinion and descriptive analysis, while the quantitative method is based on mathematics calculation.

III. Result and Discussion

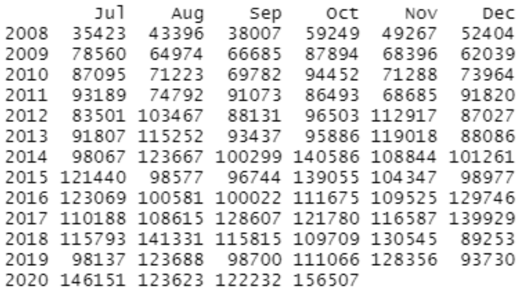
A. Data Preparation

Data used for this research is monthly data of Social Security Administration (SSA) for Retirement Insurances Applications Filled via Internet from January, 2008 to October, 2020 [8]. This data is divided into two sets of data, data training and data testing. Data training uses the data from January, 2008 to December, 2018 while data testing uses the data from January, 2019 to December, 2019.

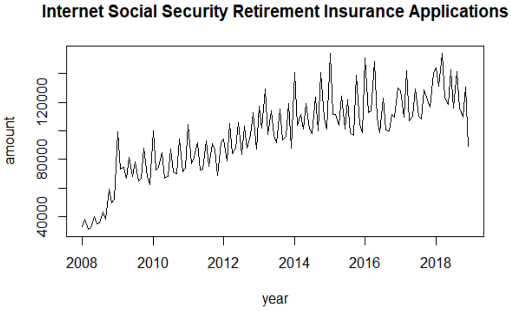
To understand more about this data, the time series plot can be seen in Figure 1. The training and testing data set will be obtained from this data (Figure 1 and Figure 2). In order to evaluate the obtained model, we separate the data into training and testing data. The testing data is from January, 2008 to December, 2018 as seen in Figure 3. This data will be compared to the forecast result. A good model can predict data that was not previously used in training data.



```
        Jan    Feb    Mar    Apr    May    Jun
2008  33151  38106  31150  32591  40095  34686
2009  99090  73155  74270  66512  81176  68091
2010 100021  72412  74719  84698  67075  68119
2011 104632  76977  81435  91748  72343  73310
2012  94114  79125 105012  84143  87643 105455
2013 116966 101962 129199  97197 113786  96549
2014 140599 103565 111704 101237 118879 101203
2015 154190 111656 111568 104038 124298 101130
2016 150784 112829 113880 148800 108956  98942
2017 128153 109430 141968 107084 110130 129147
2018 144360 131154 154304 123533 118486 142756
2019 123569 112136 142316 110741 126440 105881
2020 150568 111755 137228 151622 151833 116944
```



```
        Jul    Aug    Sep    Oct    Nov    Dec
2008  35423  43396  38007  59249  49267  52404
2009  78560  64974  66685  87894  68396  62039
2010  87095  71223  69782  94452  71288  73964
2011  93189  74792  91073  86493  68685  91820
2012  83501 103467  88131  96503 112917  87027
2013  91807 115252  93437  95886 119018  88086
2014  98067 123667 100299 140586 108844 101261
2015 121440  98577  96744 139055 104347  98977
2016 123069 100581 100022 111675 109525 129746
2017 110188 108615 128607 121780 116587 139929
2018 115793 141331 115815 109709 130545  89253
2019  98137 123688  98700 111066 128356  93730
2020 146151 123623 122232 156507
```



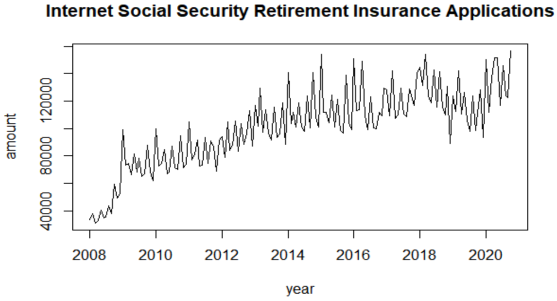


Figure 1. Time Series Plot for Retirement Insurance Application Figure 2. Time Series Plot for
Training Data

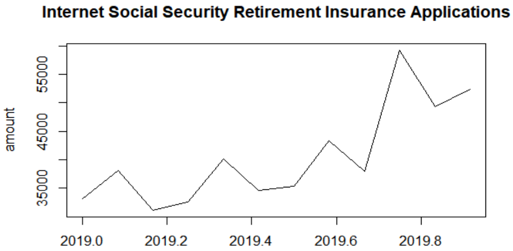


Figure 3. Time Series Plot for Testing Data

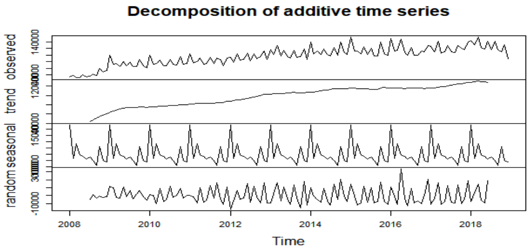


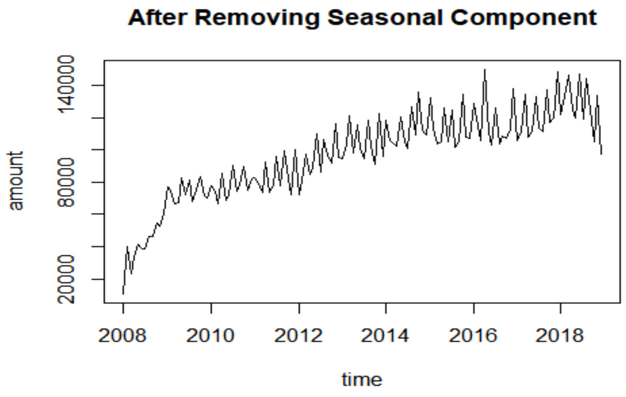
Figure 4. Decomposition plot

B. Decomposition

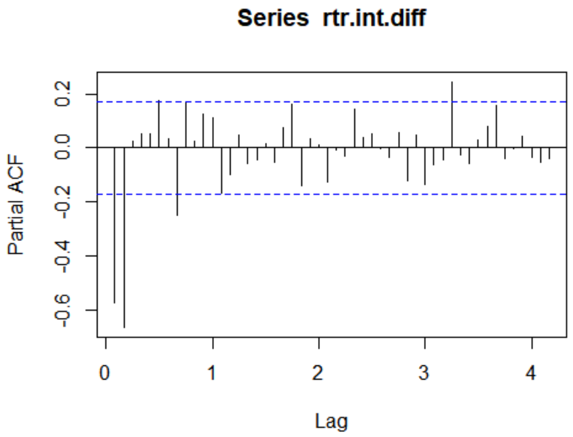
This decomposition shows that our data have a steady increasing trend with some seasonal components. The seasonal component has caused a huge difference in data, so we need to remove the seasonal character first. After we remove the seasonal component, the plot can be seen in Figure 4.

C. Stationarity Check

ARIMA Models used an assumption that the data is stationary. Thus, we need to check whether our data is stationer or not using the ADF test command in R Studio. If the result of p-value is lower than 0.05, that is 0.01, then the data is stationer. The result of the ADF test will be shown in the picture below:



D. Model Specification



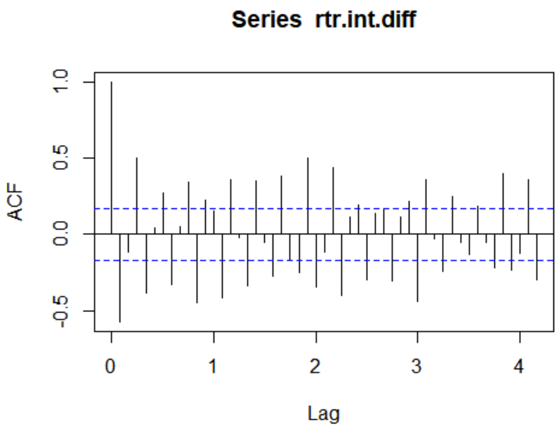


Figure 5. Autocorrelation function function

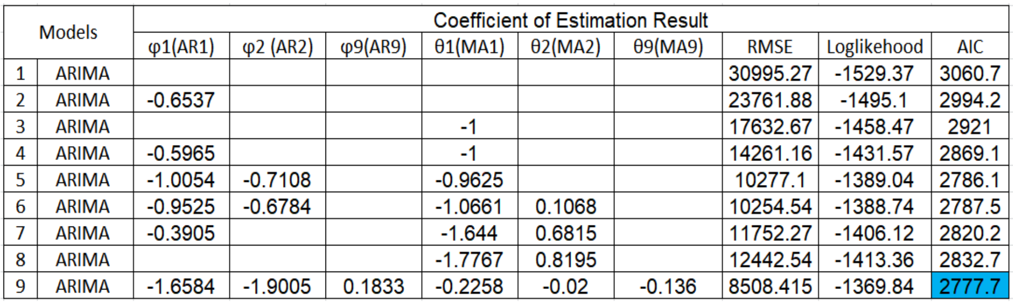
Figure 6. Partial Autocorrelation

In order to obtain a model from the training data, we first identify the ACF and PACF of the data. The ACF and PACF are calculated by using R Studio. The results are depicted in Figure 5 and 6.

TABLE 1.
MODEL SPECIFICATION

MODELS	**P**	**D**	**Q**
ARIMA (0,1,0)	0	1	0
ARIMA (1,1,0)	1	1	0
ARIMA (0,1,1)	0	1	1
ARIMA (1,1,1)	1	1	1
ARIMA (2,1,1)	2	1	1
ARIMA (2,1,2)	2	1	2
ARIMA (1,1,2)	1	1	2
ARIMA (0,1,2)	0	1	2
ARIMA (9,1,9)	9	1	9

TABLE 2
COEFFICIENT OF ESTIMATION



Models		Coefficient of Estimation Result								
		φ1(AR1)	φ2 (AR2)	φ9(AR9)	θ1(MA1)	θ2(MA2)	θ9(MA9)	RMSE	Loglikehood	AIC
1	ARIMA							30995.27	-1529.37	3060.7
2	ARIMA	-0.6537						23761.88	-1495.1	2994.2
3	ARIMA				-1			17632.67	-1458.47	2921
4	ARIMA	-0.5965			-1			14261.16	-1431.57	2869.1
5	ARIMA	-1.0054	-0.7108		-0.9625			10277.1	-1389.04	2786.1
6	ARIMA	-0.9525	-0.6784		-1.0661	0.1068		10254.54	-1388.74	2787.5
7	ARIMA	-0.3905			-1.644	0.6815		11752.27	-1406.12	2820.2
8	ARIMA				-1.7767	0.8195		12442.54	-1413.36	2832.7
9	ARIMA	-1.6584	-1.9005	0.1833	-0.2258	-0.02	-0.136	8508.415	-1369.84	2777.7

To choose which model that we will use, we need to see which model that have the smallest value AIC among of all models that have been identified. Thus, ARIMA (9,1,9) is the best model that we have.

E. Diagnostic Test

The residual analysis can be proven by looking at Q-Q plot, the histogram, the Shapiro test, and the Ljung-Box test.

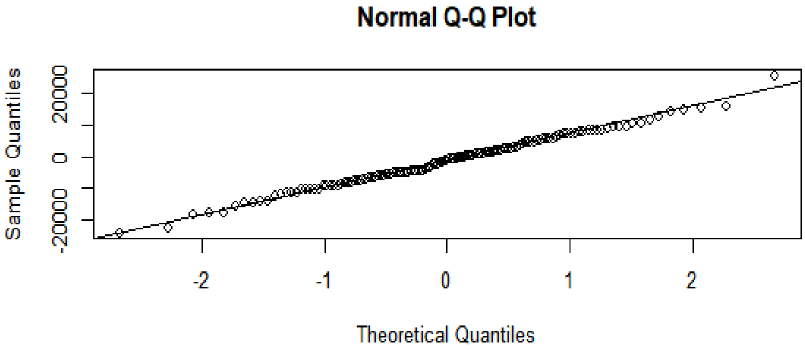


Figure 7. Normal Q-Q Plot

From the plotting Figure 7, we see that the line is almost directly above the dot that appears in the figure. This indicates that our model is correct, and we can next to the forecasting.

Figure 8 shows the curve from the histogram that establishes a normal curve, so we visually can be sure that our residuals are normally distributed.

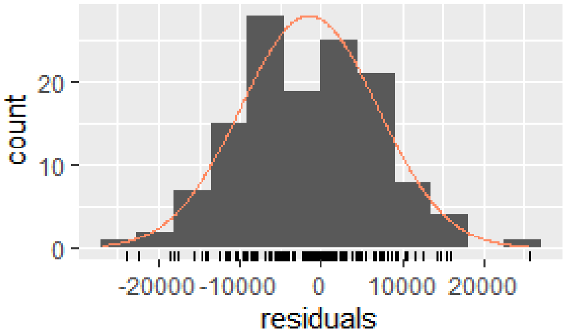


Figure 8. Normal Curve

After checking the residual of data, we need to check by testing with Shapiro test and Ljung-Box test. We need to test the data step by step, starting from the Shapiro test. If the *p*-value is bigger than 0.05, then we can continue to the next step, the Ljung-Box test. There are 4 steps in the Ljung-Box test, lag is equal to 12, 24, 36, and 48. If all the test results with p-value bigger than 0.05, then we can move to the forecasting section.

F. Forecasting

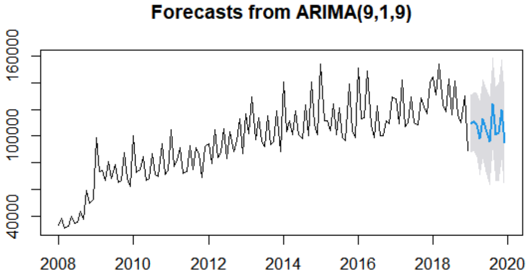


Figure 9. Time Series Plot of Forecasting Result

The forecast result for model ARIMA(9,1,9) with a confidence of 95% can be seen in Figure 9. The solid line is denoting the actual data. The forecast result is denoted thick blue line. The lower limit and upper limit for forecast results are denoted as grey area. The detailed result for the forecast data and the actual data can be seen in Table 3. As seen that, though the model is generated from training data the model can also be used to model the testing data. From the table, despite the forecast results are good. At least, the real data lie between lower and upper limits.

TABLE 3.
FORECAST DATA AND ACTUAL DATA.

Month	Real Data	Upper Limit	Lower Limit	Forecast Result
January	123,569	131,165	87,849	109,507
February	112,136	132,747	89,385	111,066
March	142,316	132,547	84,993	108,770
April	110,741	126,730	70,238	98,484
May	126,440	142,708	82,851	112,780
June	105,881	137,050	73,753	105,402
July	98,137	129,216	63,058	96,137
August	123,688	159,292	89,898	124,595
September	98,700	137,260	65,687	101,474
October	111,066	139,600	66,338	102,969
November	128,356	157,939	82,342	120,141
December	93,730	133,331	56,588	94,960

IV. CONCLUSION

The forecasting using ARIMA models to forecast future retirement insurance applications via internet is quite accurate even there are some points that are light different from the actual data. And by this result, we can conclude that ARIMA (9,1,9) is the best model to forecast.

REFERENCES

[1] U. A. Hafiz, F. Salleh, M. Garba, and N. Rashid, “Projecting Insurance Penetration Rate in Nigeria: An ARIMA Approach,” *Rev. GEINTEC-GESTAO INOVACAO E Tecnol.*, vol. 11, no. 3, pp. 63–75, 2021.

[2] M. D. Kartikasari and N. Imani, “Time Series Analysis of Claims Reserve in General Insurance Industry,” *Proc. Book*, vol. 60007, pp. 1–060007, 2018.

[3] V. S. Kumar, D. K. Satpathi, P. P. Kumar, and V. Haragopal, “Forecasting motor insurance claim amount using ARIMA model,” in *AIP Conference Proceedings*, 2020, vol. 2246, no. 1, p. 020005.

[4] L. Arsyad and L. Arsyad, “Peramalan Bisnis.,” *BPFE Yogyak.*, 2001.

[5] J. L. Whitten, L. D. Bentley, K. C. Dittman, and others, *Systems analysis and design methods*. Irwin Homewood, IL, 1989.

[6] J. D. Cryer and K.-S. Chan, *Time series analysis: with applications in R*, vol. 2. Springer, 2008.

[7] W. Makridakis, C. Steven, and V. E. Wheelwright, “McGee. 1999,” *Metode Dan Apl. Peramalan*.

[8] “Social Security,” *SSA Open Data | Count of Monthly Retirement Insurance Applications Filed via the Internet | 2008-2011*. [Online]. Available: https://www.ssa.gov/open/data/retirement-insurance-online-apps.html