Journal of Computer and Communications, 2020, 8, 37-49
https://www.scirp.org/journal/jcc
ISSN Online: 2327-5227
ISSN Print: 2327-5219

Statistical and Machine Learning Methods for Vaccine Demand Forecasting: A Comparative Analysis

Rachel T. Alegado[1,2], Gilbert M. Tumibay[3]

[1]Nueva Ecija University of Science and Technology, Cabanatuan, Philippines
[2]Graduate School, Angeles University Foundation, Angeles, Philippines
[3]Angeles University Foundation, Angeles, Philippines
Email: rachelalegado@neust.edu.ph, tumibay.gibo@auf.edu.ph

How to cite this paper: Alegado, R.T. and Tumibay, G.M. (2020) Statistical and Machine Learning Methods for Vaccine Demand Forecasting: A Comparative Analysis. *Journal of Computer and Communications*, **8**, 37-49.
https://doi.org/10.4236/jcc.2020.810005

Received: September 24, 2020
Accepted: October 25, 2020
Published: October 28, 2020

Abstract

This study aimed to find a suitable model for forecasting the appropriate stock of vaccines to avoid shortage and over-supply. The Auto-Regressive Integrated Moving Average (ARIMA) and Multilayer Perceptron Neural Network (MLPNN) models were used for forecasting time series data. The monthly vaccination coverage was used to develop the models from January 2014 until December 2019. The dataset consists of 72 months of observation, the 60 months of data are used for model fitting from January 2014 to December 2019, and the remaining 12 months of data from January 2019 to December 2019 are used to test the accuracy of the forecast. The most suitable forecast model was selected based on the lowest Root Mean Square Error (RMSE) value and the Mean Absolute Error (MAE). The analytical result shows that the MLPNN model outperformed the ARIMA model in forecasting monthly demand for vaccines. The results will help policymakers improve the proper use of vaccination resources.

Keywords

Vaccine Demand, Forecasting, ARIMA, Machine Learning

1. Introduction

The demands for vaccines have significantly increased due to the occurrence of outbreaks and birth rates in the Philippines. This increase storage, transport capacities, and handling of vaccines, thus resulting in complex and challenging to manage the immunization supply chain. The vaccine demand forecasting tool will help address a host of problems, such as small storage space, low stock, overstocking, and maintenance costs. Forecasting is an essential part of management's decision-making activities and plays a vital role in many areas of the company [1]. It has been applied in various fields such as car fuel consumption forecasting, gold price forecasting, electricity load consumption, wind power, price spike prediction, and so on [2]. The use of demand forecasting models is an excellent part of the vaccine supply chain decision-making process.

This study compares the Multilayer Perceptron Neural Network (MLPNN) and Auto-Regressive Integrated Moving Average (ARIMA) models for vaccine demand forecasting. The suitable forecasting models were selected based on the lowest Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) values. The forecasting results generated using the best model will serve as the basis to provide decision-making solutions to improved immunization planning and program.

The Autoregressive Integrated Moving Average (ARIMA) model is considered the most advanced and robust approach among the traditional forecasting models [3]. It is a type of statistical model for forecasting time-series data based on the Autoregressive (AR) and Moving Average (MA) processes. Many researchers use ARIMA models for forecasting [4] [5] [6]. Authors in [7] utilized the Auto-Regressive Integrated Moving Average (ARIMA) model for forecasting the price of agriculture commodities. The accuracy of the selected model was measure using two parameters: Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE). Similarly, researchers in [8] use ARIMA analysis in forecasting measles immunization coverage, while [9] use ARIMA and ARMA models to propose a forecasting model for household electric consumption and [10] use ARIMA method to develop a vaccine prediction system to ensure that immunization coverage is well optimized.

The Multilayer Perceptron Neural Network (MLPNN) is the most commonly used ANN model for time series forecasting. It is a feed-forward neural network consisting of an input layer, a hidden layer, and an output layer [11]. Currently, the applications of the MLP in forecasting time series data have diversified their field of action. Authors in [12] utilized a multilayer perceptron neural network to forecast the next 12 months' bottled water demand for a small business. Similarly, researchers in [13] use ARIMA and MLPNN model in forecasting network traffic to optimize backhaul network capacity and frequency. Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE) was used to select the best model. In [14], Shamshad *et al.* use MLPNN, ARIMA, and the ETS model to forecast the critical model weather parameters. Despite several researches in forecasting time series data, none of them used ARIMA and MLPNN model to forecast the BCG vaccine demand.

2. Methodology

2.1. CRISP-DM

This study adopted the Cross-Industry Standard Process for Data Mining or

(CRISP-DM) methodology. It is a process model that provides a fluid framework for devising, creating, building testing, and deploying machine solutions [15]. The CRISP-DM methodology consists of six phases that served as the road map in planning and conducting the study. Figure 1 shows the phases of the CRISP-DM methodology.

1) Research Understanding. The research work described in this study aims to develop a vaccine forecasting model using Autoregressive Integrated Moving Average (ARIMA) and Multilayer Perceptron Neural Network (MLPNN) models and to compare their results by evaluating the forecast performance of the models used.

2) Data Understanding. The dataset for this study was obtained from Cabanatuan City Health Office. The object of the research used is the number of infants receiving the vaccines for 72 month period from January 2014 to December 2019 in Cabanatuan City. In addition, this study selected the BCG to be the experimental vaccine for the demand forecasting implementation.

3) Data preparation. In this step, the data were divided into training and testing process of the model. This step includes splitting the series into training containing the first 85% values and testing containing the last 15% of the data set. Moreover, this step applied transformations of data such as identifying and treating outlier data and constructing and decomposing time-series format.

4) Modeling and Forecasting. In this step, the training data set was used to train the statistical and machine, learning models. The auto.arima()function of the forecast package in R was used to fits an ARIMA model. This algorithm automates the ARIMA model's tuning process by using a stepwise search to traverse the model space to select the best model with the smallest Akaike's Information Criterion (AIC) [3]. It uses a variation of the Hyndman-Khandakar algorithm to select an arima model [1].

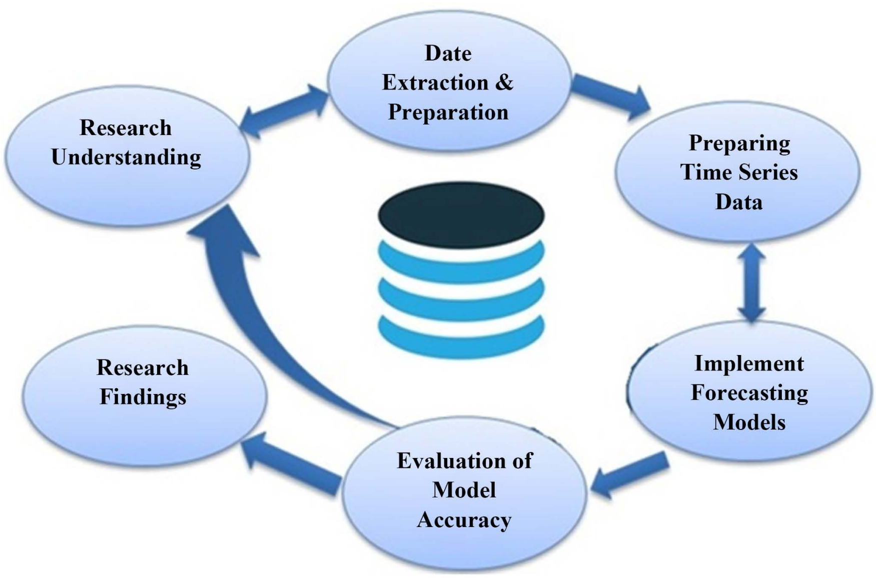

Figure 1. Methodological structure (CRISP-DM).

On the other hand, the mlp() function of nnfor package [16] was used to fits an MLPNN model. This function creates a multilayer perceptron and trains it using a back propagation algorithm. By default, only a hidden layer with five nodes was used, trained 20 times, and different forecasts were combined using the median operator.

5) Evaluation. In this step, the result of ARIMA and MLPNN models is evaluated to determine the model's accuracy. The model will be evaluated using the two performance measures: the Root Mean Square Error (RMSE) and Mean Absolute Error (MAE). The precision of the models is measured based on the lower value of these output measures [17]. The Root Mean Square Error (RMSE) is a commonly used estimate of the difference between the forecasted values of a model and the observed actual values [18]. The equation is shown on Equation (1). The Mean Absolute Error (MAE) is measured as the sum of the expected error values, where all predicted values are required to be positive. The equation is shown on Equation (2).

$$\text{RMSE} = \sqrt{\sum_{i-1}^{n} \frac{(P_i - O_i)^2}{N}} \tag{1}$$

$$\text{MAE} = \frac{1}{n} \sum_{i-1}^{n} |A_i - B_i| \tag{2}$$

2.2. Tools and Specifications

The technologies that helped achieve this study are briefly listed below, along with the methodology applied to achieve the aim of this research:

1) Microsoft Excel 2019 was used to save the raw data in .csv format and pre-process the data and the initial checking of the dataset.

2) The forecasting models were developed using R programming language and using R Studio Desktop IDE. R has several forecasting packages for secure handling of data from time series [19], so it was chosen to increase the research pace efficiency. The R language was used to code the model and visualize the datasets.

3) All the software mentioned above was running on Microsoft Windows 10 Pro machine.

3. Results and Discussion

3.1. Data Assemblage and Preparation with Research Understanding

The monthly vaccination data of Cabanatuan City, Nueva Ecija, from January 2014 to December 2019 was used in this study. The number of observations is 72 months and was divided into two parts to use in training and forecasting. In the first part, 60 monthly data are taken into account from the January 2014 to December 2018. These data are used for model fitting. The remaining 12 months of data from January 2019 to December 2019 were used as an out-of-sample set, using Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) to measure the selected model's forecast accuracy. The R studio Desktop and R Programming language were used to build the ARIMA and MLPNN model. The following steps were used for data preparation:

1) *Load Data Set and Libraries*

The first step in modeling and forecasting is to install and load the necessary packages. Next is to load the data sets from the .csv file, as shown in **Figure 2**.

2) *Detection and Treatment of Outlier data*

Outliers can greatly affect the quality of forecasting. Therefore identification and treatment of these outliers are essential. The first step in determining the outliers appear in the data is by using a box plot diagram. The box plot diagram is a graphical tool usually expressed by quartiles and interquartile, which helps to identify the upper limit and lower limit above which any data lying would be considered outliers. Both the lower and upper limit results are statistically average and should thus be used for forecasting. **Figure 3(a)** and **Figure 3(b)** display the mean and spread of data before and after detection and treat the outlier.

3) *Constructing and Decomposing Time Series Data*

This step includes the creation and decomposition of time series to determine the patterns, cycle, and seasonality of the period.

a) *Time series construction*

In R Library, the ts() function was used to define the frequency to construct a time series. This analysis used a ratio of 12 and 1 to show the monthly and yearly series. **Figure 4** displays the monthly BCG vaccine coverage time series plot beginning from the first month of 2014.

b) *Time series decomposition*

Time series decomposition procedures were carried out to identify the pattern and seasonal factors of vaccination coverage. **Figure 5** shows the decomposition of additive time-series into a random, seasonal, trend, and observe.

4) *Creating Training and Testing Data*

The data set is consisting of 72-month observations from January 2014 to December 2019. The dataset was split into training and testing data. The training

```
#Installing R Packages
install.packages("nnfor")
install.packages("ggplot2")
install.packages("TSstudio")
install.packages("plotly")
install.packages("lubridate")

#Libraries Used
library(nnfor)
library(ggplot2)
library(TSstudio)
library(plotly)
library(lubridate)

#Load the Data
rawData = read.csv("C:/Users/acer/Desktop/BCG.csv")
```

Figure 2. Installing R packages and data pre-processing.

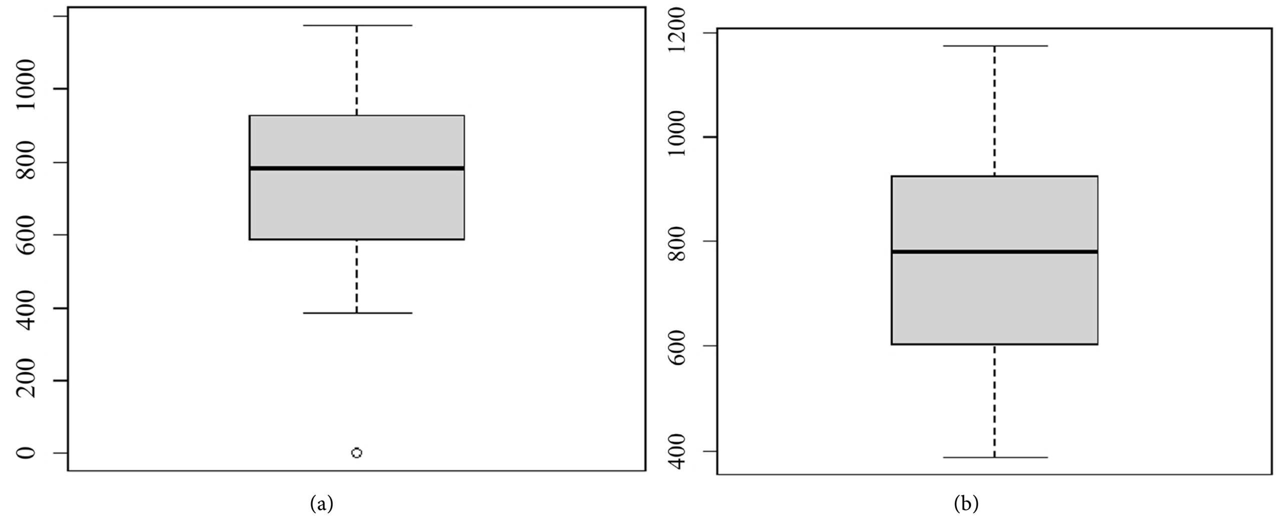

Figure 3. (a) Mean and spread of the data with outlier; (b) Mean and spread of the data after treated the outlier.

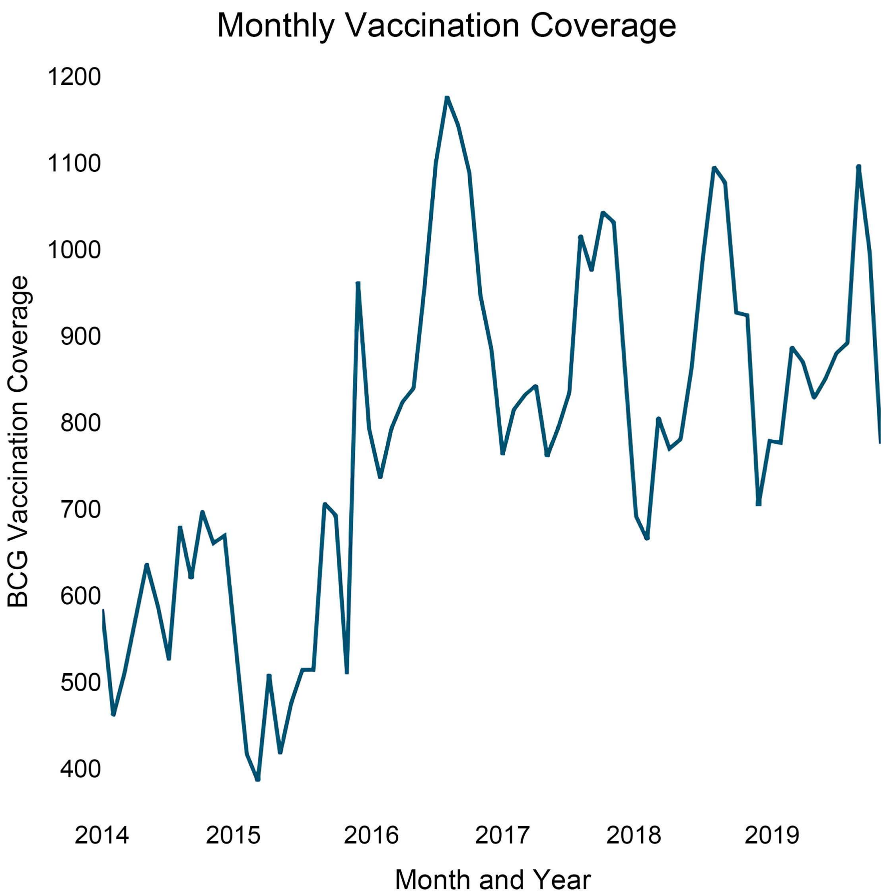

Figure 4. BCG Vaccination coverage data as time series.

dataset consists of 59-month observations from January 2014 and November 2018, while the testing data set consists of 12-month observations from December 2018 and November 2019. The code and result for the splitting dataset are shown in **Figure 6(a)** and **Figure 6(b)**, respectively

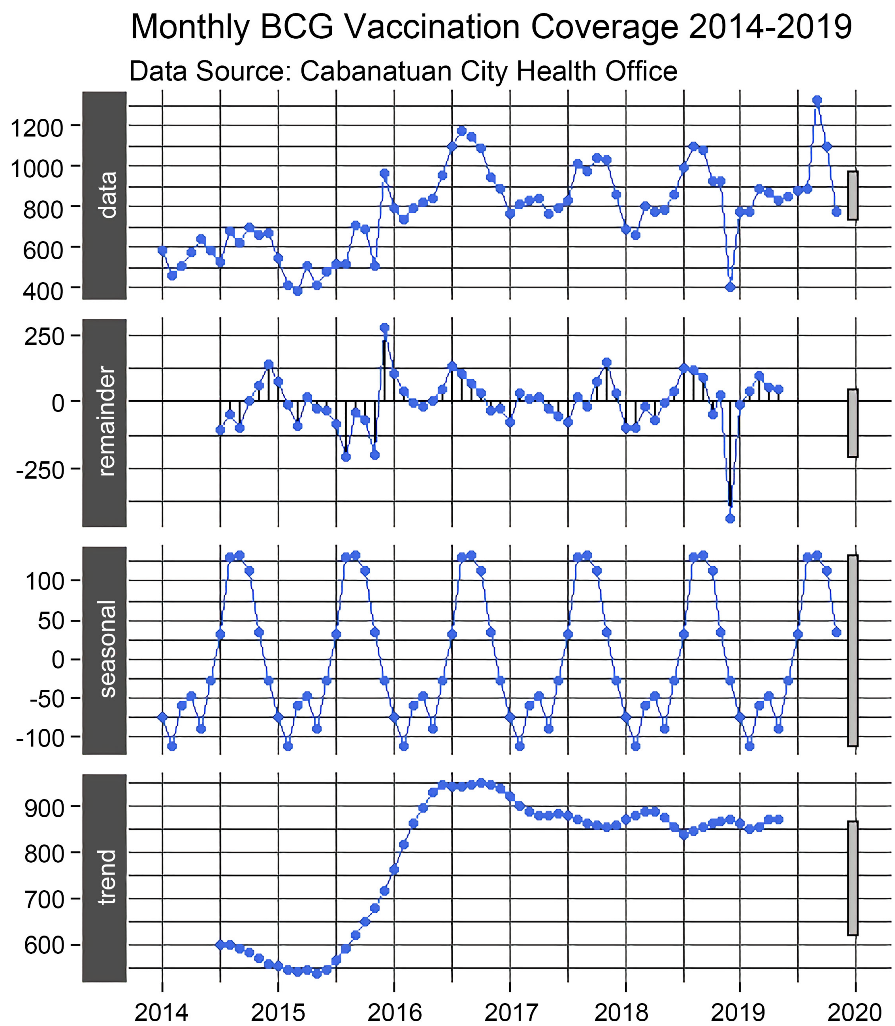

Figure 5. Decomposition of additive time series.

```
# Set the sample out and forecast horizon
h1 <- 12 # the length of the testing partition
h2 <- 60 # forecast horizon

# Splitting the time series object to training and testing partitions
TSData_split <- ts_split(TSData, sample.out = h1)
train <- TSData_split$train
test <- TSData_split$test
ts_info(train)
ts_info(test)
```

(a)

```
> ts_info(train)
 The train series is a ts object with 1 variable and 59 observations
 Frequency: 12
 Start time: 2014 1
 End time: 2018 11
> ts_info(test)
 The test series is a ts object with 1 variable and 12 observations
 Frequency: 12
 Start time: 2018 12
 End time: 2019 11
```

(b)

Figure 6. (a) R code for splitting dataset; (b) Training and testing dataset information.

3.2. Modeling and Implementation

1) ARIMA(p, d, q) Model

The order for ARIMA(p, d, q) model was determined, and the best fit model was identified. The ACF (Autocorrelation function) and PACF (Partial autocorrelation function) are the tools used for the selection of ARIMA(p, d, q). Likewise, the auto.arima function from the forecast package in R will do it automatically to find the order (p, d, q). Figure 7 shows the result of the auto.arima function.

The ARIMA(p, d, q) model that are suitable for monthly series is ARIMA(1, 0, 0)(0, 1, 1) [12]). The obtained ARIMA(1, 0, 0)(0, 1, 1) [12] model result has been shown in Figure 8.

```
> md1
Series: train
ARIMA(1,0,0)(0,1,1)[12]

Coefficients:
         ar1     sma1
      0.8701  -0.7848
s.e.  0.0801   0.4360

sigma^2 estimated as 11754:  log likelihood=-291.61
AIC=589.22   AICc=589.78   BIC=594.77
```

Figure 7. ARIMA Model.

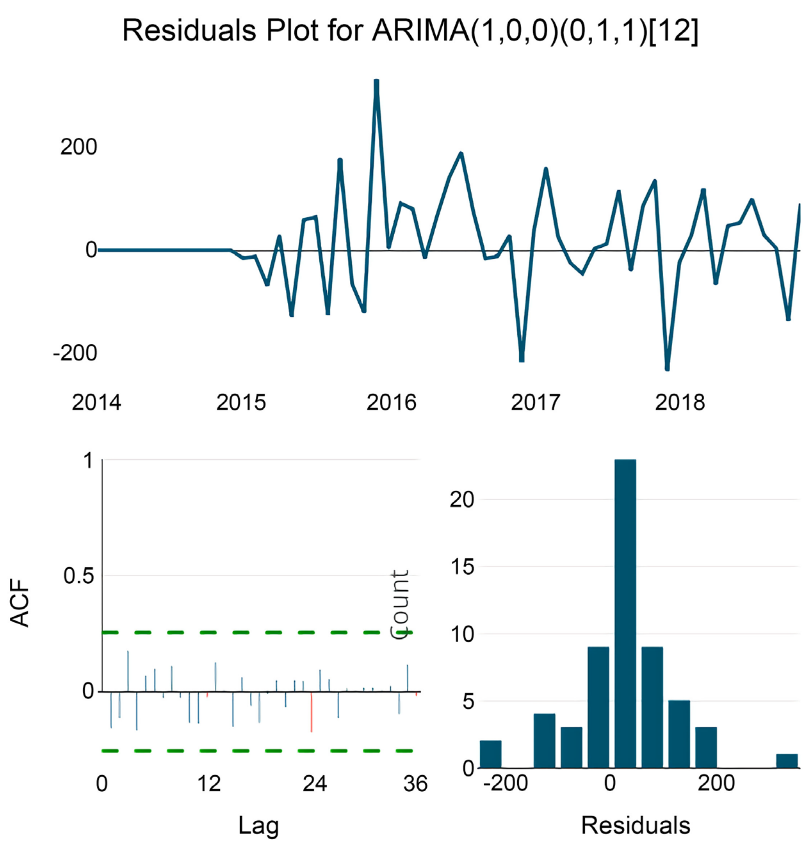

Figure 8. Residual Plot for ARIMA(1, 0, 0)(0, 1, 1) ([12]).

The ARIMA(1, 0, 0)(0, 1, 1) [12]) model was used for forecasting next 12 months of vaccination coverage from December 2018 to November 2019. Figure 9 shows the fitted, forecasted, test and train data of ARIMA(1, 0, 0)(0, 1, 1) [12]) model.

2) MLPNN Model

The analysis of Multilayer Perceptron Neural Network or MLPNN model was implemented using nnfor package of R. The result of the training of the MLPNN using 59 observations was done with five hidden nodes, 20 repetitions, and univariate lags: (1, 2, 10, 11, 12). The MSE of monthly vaccination coverage is 31.7939. The MLPNN model architecture, which consists of one input, hidden, and output layer, as shown in Figure 10. The input layer was consists of 16 nodes, which served as the input for the entire network. The gray input nodes are autoregressions, while the magenta ones are deterministic inputs. The hidden layer is consists of 5 nodes, while the output layer has only one output.

The fitted MLPNN model was used to forecast the next 12 months of BCG vaccination coverage from December 2018 to November 2019. Figure 11 shows the fitted, forecasted, test and train data of MLPNN model.

3.3. Evaluation

This study compares the results obtained from the MLPNN model developed

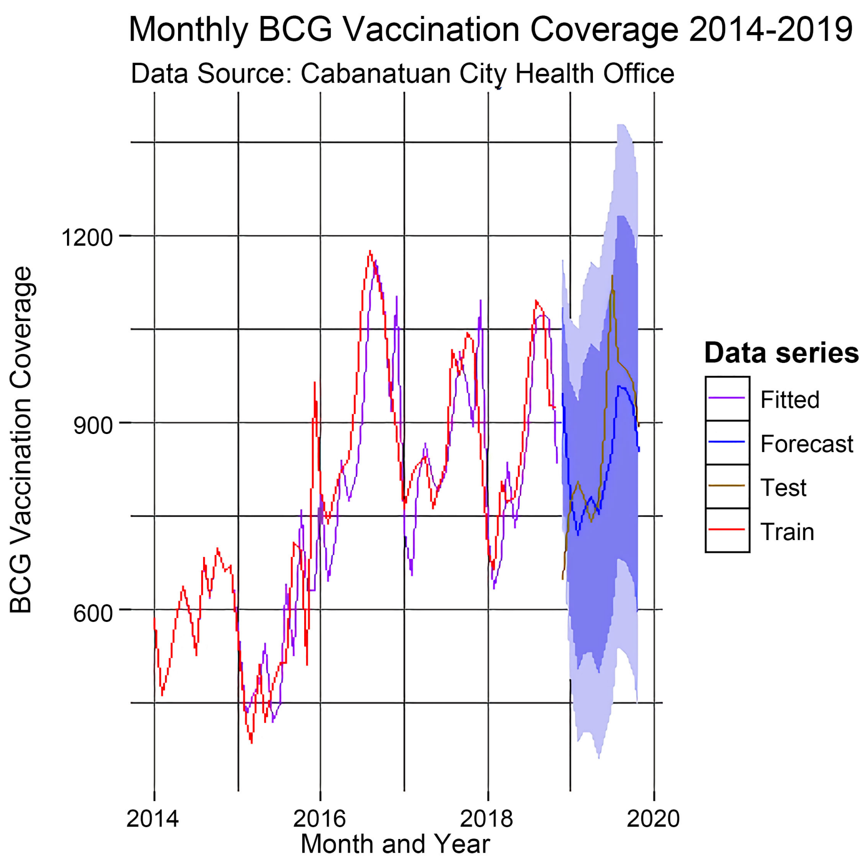

Figure 9. Fitted, Forecasted, Test and Train data of ARIMA(1, 0, 0)(0, 1, 1) [12] model.

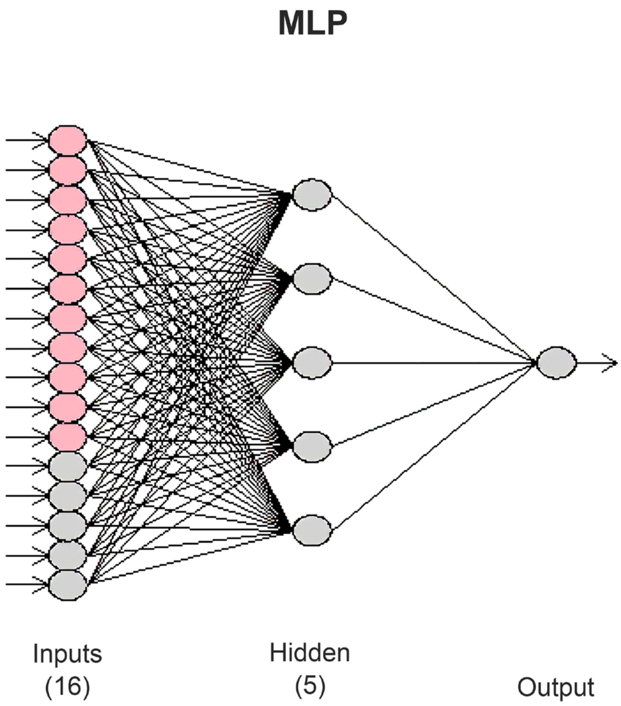

Figure 10. MLP model architecture.

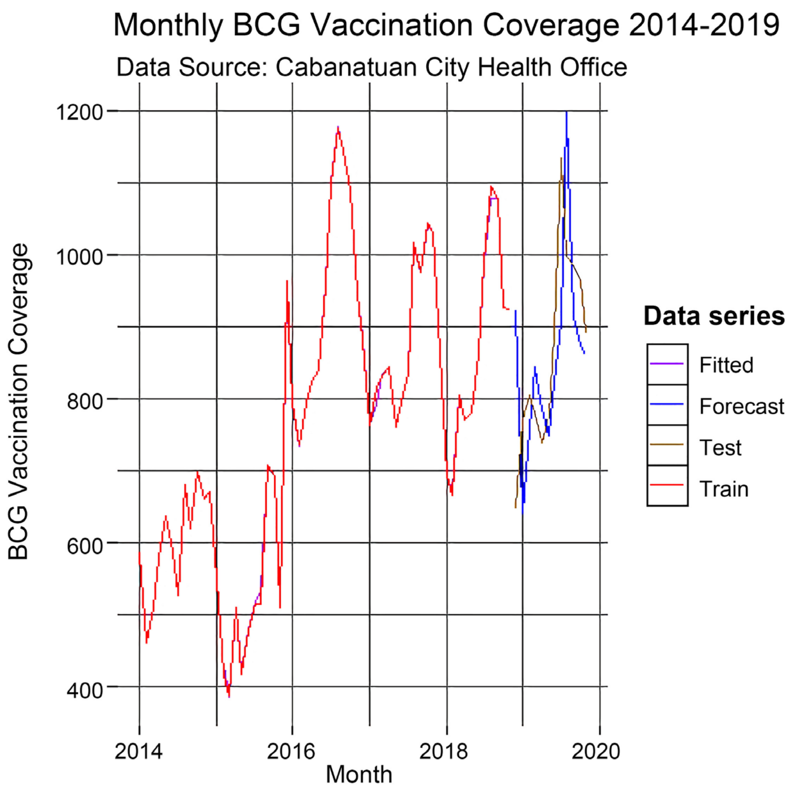

Figure 11. Fitted, Forecasted, Test and Train data of MLPNN model.

with the ARIMA model. To provide a clearer understanding of the performance of the selected methods, the models' accuracy measures are shown in Table 1.

Table 1. Accuracy measure for the models ARIMA(1, 0, 0)(0, 1, 1) [12] and MLPNN.

Models	RMSE	MAE
ARIMA(1, 0, 0)(0, 1, 1) [12]	94.68	64.04
MLPNN	5.63	2.45

The result shows that the performance metrics RMSE and MAE are low for MLPNN model. The smaller the error values, the better the model's performance. Therefore it can be concluded that the MLPNN model performs well than the ARIMA model in forecasting BCG vaccination coverage.

4. Conclusion

The goal of this research was to find a suitable model for forecasting the appropriate stock of vaccines to avoid shortage and over-supply. The MLPNN and ARIMA model was used for forecasting the monthly vaccine demand from January 2014 to December 2019. Then, it chooses the suitable forecasting method using the RMSE and MAE accuracy measures. The results showed that the MLPNN model is superior to the ARIMA model in forecasting the monthly vaccine demand. This result coincided with the previous literature that uses MLPNN and ARIMA in forecasting [20] [21]. The forecasting results in this study can help policymakers to have better decisions in improving vaccination coverage. In future studies, a further experiment may be carried out by applying this approach to a broader scale and using additional forecasting methods, particularly the hybrid model.

Acknowledgements

The authors are grateful to the staff, directors, and administrators of Cabanatuan City Health Office, Philippines for guidance, assistance, help for data collection and support to make this research work realizable.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Hyndman, R.J. and Athanasopoulos, G. (2018) Forecasting: Principles and Practice. OTexts.

[2] Ameer Amsa, M.G., Aibinu, A.M., Salami, M.J.E. and Balogun, W. (2012) A Review of Forecasting Techniques. *Proceedings on the International Conference on Artificial Intelligence* (*ICAI*). *The Steering Committee of the World Congress in Computer Science*, *Computer Engineering and Applied Computing* (*WorldComp*), *1*.

[3] Krispin, R. (2019) Hands-On Time Series Analysis with R: Perform Time Series Analysis and Forecasting Using R. Packt Publishing, Limited, Birmingham.

[4] Aradhye, G., Rao, A.C.S. and Mohammed, M.M. (2019) A Novel Hybrid Approach

for Time Series Data Forecasting Using Moving Average Filter and ARIMA-SVM. In: *Emerging Technologies in Data Mining and Information Security*, Springer, Singapore, 369-381. https://doi.org/10.1007/978-981-13-1498-8_33

[5] Fattah, J., Ezzine, L., Aman, Z., El Moussami, H. and Lachhab, A. (2018) Forecasting of Demand Using ARIMA Model. *International Journal of Engineering Business Management*, **10**. https://doi.org/10.1177/1847979018808673

[6] Aljandali, A. and Tatahi, M. (2018) Economic Forecasting Using ARIMA Modelling. In: *Economic and Financial Modelling with EViews*, Springer, Cham, 111-142. https://doi.org/10.1007/978-3-319-92985-9_7

[7] KumarMahto, A., Biswas, R. and Alam, M.A. (2019) Short Term Forecasting of Agriculture Commodity Price by Using ARIMA: Based on Indian Market. In: *International Conference on Advances in Computing and Data Sciences*, Springer, Singapore, 452-461. https://doi.org/10.1007/978-981-13-9939-8_40

[8] Alegado, R.T. and Tumibay, G.M. (2019) Forecasting Measles Immunization Coverage Using ARIMA Model. *Journal of Computer and Communications*, **7**, 157-168. https://doi.org/10.4236/jcc.2019.710015

[9] Chujai, P., Kerdprasop, N. and Kerdprasop, K. (2013) Time Series Analysis of Household Electric Consumption with ARIMA and ARMA Models. *Proceedings of the International MultiConference of Engineers and Computer Scientists*, **1**, 295-300.

[10] Sahisnu, J.S., Natalia, F., Ferdinand, F.V., Sudirman, S. and Ko, C.S. (2020) Vaccine Prediction System Using ARIMA Method. *ICIC Express Letters*, *Part B*: *Applications*, **11**, 567-575.

[11] Adhikari, R. and Agrawal, R.K. (2012) A Novel Weighted Ensemble Technique for Time Series Forecasting. In: *Pacific-Asia Conference on Knowledge Discovery and Data Mining,* Springer, Berlin, Heidelberg, 38-49. https://doi.org/10.1007/978-3-642-30217-6_4

[12] Herrera-Granda, I.D., Chicaiza-Ipiales, J.A., Herrera-Granda, E.P., Lorente-Leyva, L.L., Caraguay-Procel, J.A., García-Santillán, I.D. and Peluffo-Ordóñez, D.H. (2019) Artificial Neural Networks for Bottled Water Demand Forecasting: A Small Business Case Study. In: *International Work-Conference on Artificial Neural Networks*, Springer, Cham, 362-373. https://doi.org/10.1007/978-3-030-20518-8_31

[13] Mahmood, A., Kiah, M.L.M., Z'aba, M.R., Qureshi, A.N., Kassim, M.S.S., Hasan, Z.H.A., Azzuhri, S.R., *et al*. (2020) Capacity and Frequency Optimization of Wireless Backhaul Network Using Traffic Forecasting. *IEEE Access*, **8**, 23264-23276. https://doi.org/10.1109/ACCESS.2020.2970224

[14] Shamshad, B., Khan, M.Z. and Omar, Z. (2019) Modeling and Forecasting Weather Parameters Using ANN-MLP, ARIMA and ETS Model: A Case Study for Lahore, Pakistan. *International Journal of Scientific and Engineering Research*, **10**, No. 4.

[15] Stirrup, J. and Ramos, R.O. (2017) Advanced Analytics with R and Tableau: Advanced Analytics Using Data Classification, Unsupervised Learning and Data Visualization. Packt Publishing, Birmingham.

[16] Kourentzes, N. (2017) Nnfor: Time Series Forecasting with Neural Networks, 2017. R package version 0.9, 2, 229. https://CRAN.R-project.org/package=nnfor

[17] Rahman, A. and Ahmar, A.S. (2017) Forecasting of Primary Energy Consumption Data in the United States: A Comparison between ARIMA and Holter-Winters Models. In: *AIP Conference Proceedings*, AIP Publishing LLC, Vol. 1885, 020163. https://doi.org/10.1063/1.5002357

[18] Sunil, S., Acharya, S. and Jogi, A.K. (2019) Application of Hybrid Model for Forecasting Prices of Jasmine Flower in Bangalore, India. *International Journal of Scien-*

tific & Technology Research, **8**, No. 11.

[19] Corbyn, J. (2011) Time Series Analysis—With Applications in R. *Journal of the Royal Statistical Society-Series A*, **174**, 507. https://doi.org/10.1111/j.1467-985X.2010.00681_4.x

[20] Waheeb, W., Shah, H., Jabreel, M. and Puig, D. (2020) Bitcoin Price Forecasting: A Comparative Study between Statistical and Machine Learning Methods.

[21] Kassem, A.A., Raheem, A.M. and Khidir, K.M. (2020) Daily Streamflow Prediction for Khazir River Basin Using ARIMA and ANN Models. *Zanco Journal of Pure and Applied Sciences*, **32**, 30-39. https://doi.org/10.21271/zjpas.32.3.4