Application of ARIMA, hybrid ARIMA and Artificial Neural Network Models in predicting and forecasting tuberculosis incidences among children in Homa Bay and Turkana Counties, Kenya

Siamba Stephen[1*], Otieno Argwings[1], Koech Julius[1]

[1]University of Eldoret, School of Science, Department of Mathematics and Computer Science

*Corresponding author

Email: stephen.siamba@gmail.com

Abstract

Background

Tuberculosis (TB) infections among children (below 15 years) is a growing concern, particularly in resource-limited settings. However, the TB burden among children is relatively unknown in Kenya where two-thirds of estimated TB cases are undiagnosed annually. Very few studies have used Autoregressive Integrated Moving Average (ARIMA), hybrid ARIMA, and Artificial Neural Networks (ANNs) models to model infectious diseases globally. We applied ARIMA, hybrid ARIMA, and Artificial Neural Network models to predict and forecast TB incidences among children in Homa bay and Turkana Counties in Kenya.

Methods

The ARIMA, ANN, and hybrid models were used to predict and forecast monthly TB cases reported in the Treatment Information from Basic Unit (TIBU) system for Homa bay and Turkana Counties between 2012 and 2021. The data were split into training data, for model development, and testing data, for model validation using an 80:20 split ratio respectively.

Results

The hybrid ARIMA model (ARIMA-ANN) produced better predictive and forecast accuracy compared to the ARIMA (0,0,1,1,0,1,12) and NNAR (1,1,2) [12] models. Furthermore, using the Diebold-Mariano (DM) test, the predictive accuracy of NNAR (1,1,2) [12] versus ARIMA-ANN, and ARIMA-ANN versus ARIMA (0,0,1,1,0,1,12) models were significantly different, $p<0.001$, respectively. The 12-month forecasts showed a TB prevalence of 175 to 198 cases per 100,000 children in Homa bay and Turkana Counties in 2022.

Conclusion

The hybrid (ARIMA-ANN) model produces better predictive and forecast accuracy compared to the single ARIMA and ANN models. The findings show evidence that the prevalence of TB among children below 15 years in Homa bay and Turkana Counties is significantly under-reported and is potentially higher than the national average.

Introduction

Background

Tuberculosis (TB) is a highly infectious disease ranked among the top ten most lethal causes of mortality. Approximately 33% of the global population has been plague-ridden with TB, particularly in developing countries [1]. In 2016, over 10 million new TB cases were reported globally with children below 15 years of age accounting for about 7% of those cases [3]. In 2018, about 1 million TB disease cases and over 230,000 TB-related deaths occurred among children below 15 years with about 55% of these reported TB cases going undiagnosed and/or unreported [2].

Pediatric TB is usually overlooked [4] amid diagnosis and treatment challenges. Developing countries account for over 85% of new cases of TB globally with Asian and African countries contributing 61% and 25% of global new TB cases [2] respectively. In 2016, approximately 7 countries, globally, accounted for close to 65% of all new TB cases [2] while in 2019, 30 high TB burdened countries accounted for 87% of all new TB cases while only 8 countries accounted for approximately 67% of the total new TB cases [5].

The TB burden in Sub-Saharan Africa (SSA) is far much greater and is exacerbated by poverty, political strive, and weak health systems which have curtailed the implementation of TB control interventions. Consequently, TB has become an enormous burden to health systems that are already overstretched [6].

Tuberculosis is a disease of major concern in Kenya and is among the top five causes of mortality. Kenya is listed among the top 30 TB high burdened countries [7]. Kenya is also among 14 countries globally that suffer from the TB, TB-HIV co-infection, and Multi-Drug Resistant TB [8] triple burden. The TB prevalence for Kenya in 2015 was 233 per 100,000

population with a mortality of 20 per 100,000. In Kenya, the TB case notification increased from 11,000 to 116,723 between 1990 and 2007 [9] occasioned by the HIV epidemic and improved case detection due to improved diagnostic capacity.

The use of mathematical models in the modeling of epidemic interactions within populations has been detailed extensively. While existing interventions to control TB have been partially successful, within the context of resource constraints, mathematical modeling can increase understanding and result in better policies toward the implementation of effective strategies that would compound better health and economic benefits [10]. In addition, mathematical models are essential in guiding policymakers in resource allocation toward the prevention and control of diseases.

Several studies have utilized ARIMA, Seasonal ARIMA (SARIMA), neural network, and hybrid ARIMA models to model TB incidences in other countries [11, 13] and in these studies, the hybrid models were demonstrated to offer better predictive and forecast accuracy. In Africa, Azeez *et al.* compared the predictive capabilities of the SARIMA and the hybrid SARIMA neural network auto-regression (SARIMA-NNAR) models in modeling TB incidences in South Africa and the SARIMA-NNAR model was found to have better goodness-of-fit [12]. In addition, Li *et al*. compared the the predictive power of the ARIMA and ARIMA-generalized regression neural network (GRNN) hybrid models in forecasting TB incidences in China and concluded that the hybrid model was superior to the single ARIMA model [13].

The ARIMA and Neural Networks models have also been applied in modeling other infectious diseases. Zeming and Yanning predicted HIV-AIDS incidences in China in 2017 using the back propagation (BP) artificial neural network and the ARIMA models and concluded that

the hybrid (BP-ANN) model offered better predictive power compared to the single ARIMA model [14]. Zhou *et al.* modeled the prevalence of schistosomiasis in Qianjiang city in China using a hybrid model of ARIMA-NARNN (Nonlinear Autoregressive Neural Network) and concluded that the hybrid ARIMA-NARNN model produced high-quality prediction accuracy [15]. Yu *et al.* used the hybrid seasonal ARIMA and NARNN model to forecast incidences of the Coxsackie viral infection in Shenzhen China, and concluded the hybrid seasonal ARIMA-NARNN was the best model [16].

While hybrid ARIMA models have been applied in forecasting both the short-term and long-term incidences of infectious diseases in other countries, there has been little to no application of these cutting-edge methods in African countries with the majority of the models limited to only ARIMA models [17]. In Kenya, while ARIMA models have been applied in forecasting disease incidence [18], very little has been done in the application of hybrid ARIMA models in predicting disease incidence except in non-public health settings such as agriculture and economics.

The popularity of ARIMA models stems from their flexibility to represent varieties of time series with simplicity but with a profound limitation stemming from their linear assumptions which in many cases is usually impractical [19] since real-world applications mainly involve data exhibiting non-linear patterns. Consequently, to overcome this disadvantage, non-linear stochastic models such as the ANN models have been proposed [20]. Despite this, a single ANN model is not able to incorporate both linear and non-linear patterns and this has led to the adoption of hybrid models that can address this challenge [21]. To attain a higher degree of

133 predictive and forecasting accuracy, theoretical and empirical findings show that combining
134 different models can be effective [22].

135 To better understand the status of TB infection among children in Kenya, it is important
136 to assess the trend and forecast these incidences using available surveillance data and novel
137 models to elicit a better understanding and innovative interventions to curtail the spread of
138 pediatric TB in Kenya. This study compares linear-based ARIMA, non-linear-based ANN, and
139 hybrid ARIMA in modeling TB incidences among children below years in Homa bay and
140 Turkana counties in Kenya.

141 **Materials and methods**

142 **Study design and setting**

143 This was a retrospective study that utilized aggregated monthly TB cases among children
144 data reported by health facilities located in Homa Bay and Turkana Counties to the National
145 Tuberculosis, Leprosy and Lung Disease Program (NTLLDP) in the Treatment Information from
146 Basic Unit (TIBU) electronic system between January 2012 to December 2021. The data
147 comprised 120 data points. The study utilized data reported by health facilities in Homa bay and
148 Turkana Counties which are among the top 10 TB endemic Counties in Kenya [50].

149 **Data collection and analysis**

150 Tuberculosis case data were abstracted and aggregated for each month between January
151 2012 to December 2021 for health facilities located in Homa bay and Turkana Counties in
152 Kenya. In 2012, the Kenya Ministry of Health (MoH) through the Division of Leprosy,
153 Tuberculosis and Lung Disease transitioned the reporting of TB cases from paper-based to the
154 TIBU system [23]. The TIBU system is a national case-based surveillance system used in the

155 storage of individual cases of TB that are reported to the national TB program monthly with
156 nationwide coverage [24]. This study did not collect or utilize patient-level data.

157 The 120 TB cases data points in the data used in this study were split using the 80:20
158 ratio with 80% (2012 to 2019) of the data assigned to the training set and 20% (2020 to 2021)
159 assigned to the testing set. Splitting the data into training and testing data using the 80:20 ratio,
160 in this case, was based on the available chronological data points as this has been proved to yield
161 test error rate estimates with reasonably low bias and variance [25]. Furthermore, when the
162 training data is large enough, the model learns well and this gives predicted values that are much
163 closer to the actual values [26]. The models were built on the training set and their performance
164 was evaluated on the testing set; this method is the holdout-validation method for model
165 performance evaluation.

166 Data analysis was performed using R statistical software [27] together with applicable
167 packages for analyzing time-series data. The results were summarized using tables and figures.

168 **The Time Series concept**

169 A time series is a sequential set of data points measured over time and is typically
170 composed of the trend, cyclical, seasonal, and irregular (random) components.

171 An autoregressive (AR) model is a type of random process used to describe certain time-
172 varying processes within a time series [28]. The basic idea of AR models is that the present value
173 of a series Y_t can be linearly explained by a function of p past values, that is,
174 $Y_{t-1}, Y_{t-2} +, \ldots, + Y_{t-p}$.

175 Assuming $E(Y_t) = 0$, an AR process of order p can be written as;

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \ldots + \phi_p Y_{t-p} + \varepsilon_t, \quad (1)$$

Where ε_t is white noise (WN), and is uncorrelated with Y_s for all $s < t$

However, if the mean is $E(Y_t) = \mu \neq 0$, then Y_t is replaced by $Y_t - \mu$ to obtain;

$$Y_t - \mu = \phi_1 (Y_{t-1} - \mu) + \phi_2 (Y_{t-2} - \mu) + \ldots + \phi_p (Y_{t-p} - \mu) + \varepsilon_t \quad (2)$$

Equation 2 can also be written as;

$$Y_t = \alpha + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \ldots + \phi_p Y_{t-p} + \varepsilon_t \quad (3)$$

Where;

$$\alpha = \mu(1 - \phi_1 - \ldots - \phi_p)$$

Furthermore, equation (1) can be written in the form;

$$Y_t - \phi_1 Y_{t-1} - \phi_2 Y_{t-2} - \ldots - \phi_p Y_{t-p} = \varepsilon_t \quad (4)$$

However, by applying the backshift operator $BY_t = Y_{t-1}$, we get;

$$(1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p) Y_t = \varepsilon_t$$

Or using notation, we can write;

$$\phi(B) Y_t = \varepsilon_t, \quad (5)$$

Where $\phi(B)$ denotes the autoregressive (AR) operator;

$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p \quad (6)$$

A moving average (MA) model uses the dependency between an observed value and the residual error from a moving average model applied to the lagged observations. This implies that the output variable is linearly dependent on the current and past values of a stochastic term [28]. Consequently, Y_t is a moving average process of order q if

$$Y_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \ldots + \theta_q \varepsilon_{t-q}, \quad (7)$$

Where ε_t is WN and $\theta_1, \ldots, \theta_q$ are constants

On the other hand, a MA(q) process can also be written in the form;

$$Y_t = \theta(B)\varepsilon_t, \quad (8)$$

Where the moving average operator;

$$\theta(B) = 1 + \theta_1 B + \theta_2 B^2 + \ldots + \theta_q B^q = 1 + \sum_{j=1}^{q} \theta_j B^j \quad (9)$$

defines a linear combination of values in the shift operator $B^k \varepsilon_t = \varepsilon_{t-k}$

Autoregressive Integrated Moving Average models (ARIMA) models

An ARMA (p, q) model is a class of stochastic processes whose auto-covariance functions depend on a finite number of unknown parameters. Generally, an ARMA process of orders p and q can be represented mathematically [29] as;

$$Y_t = \sum_{j=1}^{p} \phi_j Y_{t-j} + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} + \varepsilon_t \; \forall \, t \in \mathbb{Z} \quad (10)$$

In the lag operator notation, the ARMA (p, q) process is given by

$$\phi(B) Y_t = \theta(B)\varepsilon_t \; \forall \, t \in \mathbb{Z} \quad (11)$$

Box and Jenkins introduced the ARIMA model in 1960 [30]. The ARIMA model requires only historical time series data on the variable under forecasting. Most importantly, ARIMA models are represented as ARIMA (p, d, q) where p is the number of AR terms, d is the number of non-seasonal differences, and q is the number of lagged forecast errors [31]. The ARIMA model assumes that the residuals are independent and normally distributed with homogeneity of variance and zero mean value.

Seasonal Autoregressive Integrated Moving Average models (SARIMA) models

The SARIMA model is made up of non-seasonal and seasonal components in a multiplicative model. A SARIMA model can be written as ARIMA (p,d,q) $(P, D, Q)S$ where p is the non-seasonal AR order, d is the non-seasonal differencing, q is the non-seasonal MA order, P is the seasonal AR order, D is the seasonal differencing, Q is the seasonal MA order and S is the period of repeating seasonal pattern. Generally, S=12 for monthly data.

Let the backshift operator be presented as $BY_t = Y_{t-1}$

Without differencing, a SARIMA model can be written formally as;

$$\vartheta(B^S)\phi(B)(Y_t - \mu) = \theta(B)\Theta(B^S)\,\varepsilon_t, \forall\, t \in \mathbb{Z} \quad (12)$$

Where on the left of equation 12, the seasonal and non-seasonal AR processes multiply each other, and on the right of equation 12, the seasonal and non-seasonal MA processes multiply each other. Also, in this study, S=12 since monthly TB case data are used.

Artificial Neural Networks (ANNs) models

Artificial Neural Networks have been suggested as alternative and better modeling approaches to time series forecasting [32]. The main goal of ANNs is to construct a model that mimics the human brain intelligence into a machine [33] and is biologically motivated [34]. The most common ANNs are multi-layer perceptrons (MLPs) that utilize a single hidden layer feed-forward network (FNN) [35] made up of the input layer, the hidden layer, and the output layer connected by acyclic links [36]. A neuron is a data processing unit while the nodes in the various layers of ANNs are the processing elements. In this study, the inputs were the lagged TB case observations and the outputs were the predicted or fitted TB cases from the model. According to Zhang [37], the ANN can be written as:

$$Y_t = \gamma_0 + \sum_{j=1}^{q} \gamma_j h\left(\beta_{0j} + \sum_{i=1}^{p} \beta_{ij} Y_{t-i}\right) + \varepsilon_t, \forall\, t \quad (13)$$

Where Y_{t-i} *(i=1, 2, ..., p)* are the *p* inputs, Y_t is the output, *q* are the hidden nodes, γ_j *(j=0, 1, 2, ..., q)* and β_{ij} *(i=0, 1, 2, ..., p; j=0, 1, 2, ..., q)* are the connection weights and ε_t is the random shock; γ_0 and β_{0j} are the bias terms. There is no systematic rule in deciding the choice of *q* while *p*, which is the number of neurons is equal to the number of features in the data. The logistic function *h(.)* is applied as the nonlinear activation function, where:

$$h(.) = \frac{1}{1+e^{-x}} \quad (14)$$

As a matter of fact, the model in equation 13 performs a nonlinear functional mapping from past observations of a time series to the future value. That is:

$$Y_t = f\left(Y_{t-1}, Y_{t-2}, \ldots, Y_{t-p}, v\right) + \varepsilon_t \quad (15)$$

Where *v* is a vector of all parameters and f(.) is a function determined by the structure of the network and the connection weights.

Hybrid (ARIMA-ANN) models

Generally, a time series can be observed as having linear and nonlinear components as shown in equation 16.

$$Y_t = l_t + n_t \quad (16)$$

Where l_t and n_t are the linear (from the ARIMA model) and nonlinear (ANN fitted ARIMA model residuals) components respectively. Residuals from the ARIMA model are fitted with the ANN model.

Proposed Methodology

The proposed methodology for this study was based on the combination of the Box-Jenkins methodology for ARIMA modeling, and the ANN and hybrid ARIMA models as shown in Fig 1.

Fig 1: The Proposed methodology

Model identification and specification

Optimal values of *p, d,* and *q* for the ARIMA model were determined by examining the autocorrelation functions and the best model was determined by testing models with different parameters of *p*, *d,* and *q*. The models were estimated using the maximum likelihood estimation (MLE) method and the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) [38] penalty function statistics were used to determine the best model that minimizes AIC or BIC.

One assumption of the ARIMA model is that the residuals should be white noise. As such, the Ljung-Box Q test [39] was used to test the hypothesis of independence, constant variance and zero mean of the model residuals.

Accuracy measures

Various accuracy measures have been proposed [40] to determine predictive and forecast performance. This study used the Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and the Mean Absolute Percent Error (MAPE) to measure the predictive and forecast accuracy of the three models. The lower the values of these accuracy measures the better the model. Furthermore, MAPE values of 10% or below, 10-20%, and 20-50% should be considered as high accuracy, good accuracy, and reasonable accuracy [41].

The study also compared the predictive accuracy of the forecasts from the three models using the Diebold-Mariano (DM) test [42]. The test was used to test the null hypothesis that two models have similar predictive accuracy.

Ethical approval and considerations

Authorization for use of the data from the TIBU system was obtained from the National Commission for Science, Technology, and Innovation (NACOSTI) through a research permit and letters of approval from the research unit of the department of health services Homabay County and the department of medical services Turkana County.

Results

Exploratory data analysis

There was a total of 120 data points in this dataset. The trend of the TB cases among children below 15 years in Homa bay and Turkana counties in the data is shown in Fig 2 showing a marginal increase in the TB cases reported between 2018 and 2021.

Fig 2: Monthly TB cases among children below 15 years from Homabay and Turkana Counties between 2012 and 2021

The monthly cycle box plot of TB cases is shown in Fig 3 where there is a potential presence of seasonality within the reported TB cases. However, whether or not to account for seasonality in the model depends on whether this would improve model accuracy. This implies that there is need to account for seasonality within the ARIMA model. Furthermore, outliers were detected in some months.

Fig 3: Monthly cycle plot of TB cases

Comparison of model performance in predicting TB cases among children below 15 years in Homabay and Turkana Counties, Kenya

ARIMA Model estimation and accuracy

The Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) were used to pick the best parsimonious model based on the least AIC or BIC estimated values. The best model was ARIMA (0,0,1,1,0,1,12); where p=0, d=0 and q=1 respectively and P=1, D=0 and Q=1 respectively. The Ljung-Box Q test for the best model showed a p-value of 0.971 implying that the ARIMA (0,0,1,1,0,1,12) model residuals were independently distributed.

The best model was made up of non-differenced seasonal AR (1), non-seasonal MA (1) model and seasonal MA (1) polynomials and using equation 12, the ARIMA (0,0,1,1,0,1,12) model equation can be written as:

$$(1 - \vartheta_1 B^{12})(Y_t - \mu) = (1 + \Theta_1 B^{12})(1 + \theta_1 B)\varepsilon_t, \forall\, t \in \mathbb{Z} \quad (17)$$

Where ε_t is WN, and uncorrelated with Ys for each s < t

From the model output, the estimated coefficients were (see Table 1); ma1 = θ_1= 0.291, sar1 = ϑ_1 = 0.997, sma1 = Θ_1 = -0.953 and μ = Intercept = 50.902

Plugging these estimated coefficients into equation 17, yields the model equation:

$$(1 - 0.997B^{12})(Y_t - 50.902) = (1 - 0.953B^{12})(1 + 0.291B)\varepsilon_t, \forall\, t \in \mathbb{Z} \quad (18)$$

Table 1: Estimated model coefficients

	Coefficients	Std. Error	Z-value	Pr(>|z|)
ma1	0.291	0.108	2.701	0.007*
sar1	0.997	0.015	68.167	<0.001**
sma1	-0.953	0.127	-7.512	<0.001**
Intercept	50.902	5.064	10.052	<0.001**

Significance codes: 0 '***' 0.001'**' 0.01 '*' 0.05 '.' 0.1 ' '1

ARIMA Model residual diagnostics

The best ARIMA model was assessed for fit using the standard model residual analysis. In model diagnostic checking, 4 plots were used to test the underlying assumptions as shown in Fig 4. The Q-Q plot was relatively normal except for a few outliers at the tails, with model residuals being normally distributed. Inspection of the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots to test residual randomness to identify patterns or extreme values showed significant auto-correlations at lag 3.

Fig 4: ARIMA Model residual diagnostics

ARIMA model performance

The performance of the ARIMA (0,0,1,1,0,1,12) model was carried out by comparing predicted and forecasted TB cases with the actual TB cases reported and presented in Fig 5 and Fig 6 respectively.

Fig 5: Comparison of ARIMA predicted versus actual (training data) TB cases

Fig 6: Comparison of ARIMA 24-month forecast versus actual (test data) TB cases

ARIMA (0,0,1,1,1,0,1,12) model accuracy

Table 2 compares the accuracy of parameters/measures of the ARIMA (0,0,1,1,0,1,12) model on the training (2012 to 2019) and testing (2020 to 2021) data. The accuracy measures compared were RMSE, MAE, and MAPE correspondingly.

The accuracy measure comparison in table 2 demonstrates that the model performs slightly worse on the testing data an RMSE value of 29.17 compared to 18.74 on the training data.

Large RMSE values generally imply that the fitted model fails to account for important information within the underlying data hence the need to account for such information which might be captured within the non-linearities as shown in the additive hybrid model methodology which provides fitted values that are as closer to the actual/observed data.

Table 2: Model (ARIMA (0,0,1,1,0,1,12)) accuracy comparison on training and testing data

Data	RMSE	MAE	MAPE
Training	18.74	14.39	39.00
Testing	29.17	20.47	33.15

Artificial Neural Network (ANN) model estimation and accuracy

The training dataset was fit using an ANN model using the Neural Network Auto-Regressive (NNAR) function to produce an NNAR (*p*,*P*,*k*) [m] model. The optimal lag parameter, *p*, and the number of nodes in the hidden layer, *k*, were automatically selected while *P*=1 by default. In addition, a decay parameter of 0.001 and a maximum iteration of 200 were pre-set for the model to help restrict the weights from becoming too large and ensure that the model can test different models until the optimal model that has the minimal RMSE produced respectively.

The optimal NNAR model produced was NNAR (1,1,2) [12] with an average of 20 networks each of which was a 2-2-1 network with 9 weights. The plot of the point forecast values from the NNAR (1,1,2) [12] model on the training set against the actual training data were presented in Fig 7 while Fig 8 presents the 24-month forecast based on the NNAR (1,1,2) [12] model. Table 3 shows the accuracy measures from the model. The model seems to perform badly on the testing dataset although the MAPE values below 50% are reasonable.

Fig 7: Comparison of NNAR (1,1,2) [12] predicted TB cases versus actual TB cases (training data)

Fig 8: Comparison of NNAR (1,1,2) [12] 24-month forecast TB cases versus actual TB cases (test data)

Table 3: Model (NNAR (1,1,2) [12]) accuracy comparison between training and testing set

Data	RMSE	MAE	MAPE
Training	18.56	14.58	29.89
Testing	28.65	21.95	38.86

Hybrid (ARIMA-ANN) model estimation and accuracy

Residuals from the optimal ARIMA (0,0,1,1,0,1,12) model were fit using an ANN model and the accuracy measures and comparison of the forecast and prediction were presented in Table 4, Fig 9 and Fig 10 respectively.

Fig 9: Comparison of ARIMA-ANN predicted TB cases versus actual TB cases (training data)

Fig 10: Comparison of ARIMA-ANN 24-month forecasted TB cases versus actual TB cases (testing data)

Table 4: Hybrid ARIMA-ANN model accuracy

Data	RMSE	MAE	MAPE
Training	19.08	15.32	42.42
Testing	27.61	19.69	32.89

Comparison of predictive accuracy of the models

The predictive accuracy of the models was compared using the Diebold-Mariano (DM) test with the null hypothesis that the predictive accuracy of the two models compared are the same. The results in table 5 show that the NNAR (1,1,2) [12] and ARIMA-ANN models present significantly different predictive accuracies, similar to ARIMA (0,0,1,1,0,1,12) versus ARIMA-

ANN models. In general, the ARIMA-ANN model offers better predictive accuracy compared to the NNAR (1,1,2) [12] and ARIMA (0,0,1,1,0,1,12) models.

Table 5: Comparison of predictive accuracy

Model	DM statistic	Loss Function Power	P-value
ARIMA (0,0,1,1,0,1,12) Vs NNAR (1,1,2) [12]	0.732	2	0.466
NNAR (1,1,2) [12] Vs ARIMA-ANN	6.260	2	<0.001
ARIMA (0,0,1,1,0,1,12) Vs ARIMA-ANN	8.732	2	<0.001

Comparison of model performance in forecasting temporal trends of TB incidences

The resulting ARIMA (0,0,1,1,0,1,12), NNAR (1,1,2) [12] and ARIMA-ANN models were used to forecast TB cases for the next 12 months (2022). The forecast results were presented in table 6. The 12-month mean forecasted TB cases was 55, 59 and 52 cases per month based on the ARIMA (0,0,1,1,0,1,12), ANN (1,1,2) [12] and ARIMA-ANN respectively for 2022 giving a total of 657, 706 and 629 TB cases forecasted for the year 2022 from the ARIMA (0,0,1,1,0,1,12), ANN (1,1,2) [12] and ARIMA-ANN models respectively.

Table 6: TB cases 12 month point forecast comparison

Month	ARIMA (0,0,1,1,0,1,12)	NNAR (1,1,2) [12]	ARIMA-ANN
Jan-22	57	53	44
Feb-22	61	53	45
Mar-22	62	52	60
Apr-22	56	58	43
May-22	46	52	44
Jun-22	54	52	71
Jul-22	63	72	54
Aug-22	46	52	52
Sep-22	53	56	62
Oct-22	63	103	49
Nov-22	48	57	46
Dec-22	48	46	59
Mean	**55**	**59**	**52**
Total	**657**	**706**	**629**

Discussion

Although all three models were able to predict TB cases among children below 15 years, the hybrid ARIMA-ANN model was able to offer better predictive performance compared to single ARIMA and ANN models. These findings compare with those from other studies which applied either hybridized ARIMA or SARIMA in the modeling of TB incidences and other infectious diseases [12, 16, 43, 44] with the overall conclusion that hybrid models have better predictive performance. The majority of infectious disease data are neither purely linear nor non-linear and mostly present with both linear and nonlinear properties. As such, single models are not enough in modeling such kinds of data. Hybrid models are found to be most appropriate for the accurate estimation of such data [45]. The use of hybridized ARIMA models has been proposed in recent years and used extensively with improvements proposed over time.

The estimated TB prevalence in Kenya was 259 TB cases per 100,000 population in 2020 [4]. This translates to approximately 134,680 TB cases and with children accounting for about

409 20% (26,936) of these cases [46], the prevalence among children below 15 years was
410 approximately 121 TB cases per 100,000 population of children.

411 From the forecasted mean of 52 to 59 TB cases per month in 2022 for Homabay and
412 Turkana Counties, this study estimates that for the year 2022, the number of TB cases reported
413 would be approximately 629 to 706. However, given that these are estimated reported cases, they
414 most likely represent only about 35% of TB cases since up to 65% of pediatric TB cases are
415 potentially missed each year [2]. Taking this into account, the estimated TB cases for 2022 will
416 be approximately 1797 to 2017 for Homa Bay and Turkana Counties among children. The
417 estimated population of children below 15 years in Homabay and Turkana Counties for 2022 is
418 approximately 1,020,795 [47]. As such, the estimated TB prevalence among children in Homa
419 Bay and Turkana Counties in 2022 would be approximately 176 to 198 TB incidences per
420 100,000 population. These TB prevalence estimates for 2022 are way higher than the estimated
421 national average of 121 per 100,000 population of children below 15 years in 2020.

422 The findings of this study show that the estimated TB prevalence among children below
423 15 years is much higher compared to the estimated national average for 2020. These findings are
424 in line with the WHO newsletter that showed that the number of people developing TB and
425 dying from the disease could be much higher in 2021 and 2022 mainly because of the COVID-19
426 pandemic [48]. These findings also confirm those by Oliwa *et al.* who indicated that notification
427 data may underestimate the TB burden among children [49] while Mbithi *et al*. reported a
428 decrease in TB diagnosis in Kenya by an average of 28% in the year 2020 [50].

Conclusion

The hybrid ARIMA model offers better predictive accuracy and forecast performance compared to single ARIMA and ANN models in modeling TB cases among children below 15 years in Homabay and Turkana Counties.

The findings in this study allude to the under-reporting of TB cases among children below 15 years. As such, there is need to re-look at the TB surveillance framework data more closely to understand existing gaps. There is an urgency to re-align vital resources towards the National TB program to have the TB fight back on track.

Limitations

This study utilized data collected and reported in the TIBU system, as such, the study did not have control over the quality and accuracy of the data.

This study utilized data between 2012 to 2021 which comprised 120 data points. Deep learning algorithms such as ANNs usually demand a large amount of data to allow the algorithm to effectively learn. In this study, there were 96 data points for model development and while this represented 80% of the records, it might not have been sufficient enough to allow proper learning of the algorithm. Furthermore, there were only 24 records used for model validation and this might not have been large enough to allow for better learning by the algorithm.

This study combined data and analysis for Turkana and Homabay County. However, these two counties might present different scenarios when it comes to pediatric TB.

Finally, since the study focused on modeling TB cases among children below 15 years in Homa bay and Turkana counties, the findings might not be generalized to other counties of Kenya.

Acknowledgments

We would like to thank the departments of health in Homa bay and Turkana County for granting permission to access and utilize TB data in the TIBU system, and we hope that this work aids in the realization of the goals of the TB program in these two counties and beyond.

References

1. Liao CM, Lin YJ, Cheng YH. Modeling the impact of control measures on tuberculosis infection in senior care facilities. Building and environment. 2013 Jan 1; 59:66-75.
2. WHO *Global tuberculosis report.* 2018
3. WHO *Global tuberculosis report.* 2017
4. WHO *Global tuberculosis report.* 2020
5. Floyd K, Glaziou P, Zumla A, Raviglione M. The global tuberculosis epidemic and progress in care, prevention, and research: an overview in year 3 of the End TB era. The Lancet Respiratory Medicine. 2018 Apr 1;6(4):299-314.
6. Zumla A, Petersen E, Nyirenda T, Chakaya J. Tackling the tuberculosis epidemic in sub-Saharan Africa–unique opportunities arising from the second European Developing Countries Clinical Trials Partnership (EDCTP) programme 2015-2024. International Journal of Infectious Diseases. 2015 Mar 1;32:46-9.
7. WHO *Global tuberculosis report.* 2016
8. Kimani E, Muhula S, Kiptai T, Orwa J, Odero T, Gachuno O. Factors influencing TB treatment interruption and treatment outcomes among patients in Kiambu County, 2016-2019. PloS one. 2021 Apr 6;16(4):e0248820.

9. Kipruto H, Mung'atu J, Ogila K, Adem A, Mwalili S, Masini E, Kibuchi E. The epidemiology of tuberculosis in Kenya, a high TB/HIV burden country (2000-2013). International Journal of Public Health and Epidemiology Research. 2015;1(1):2-13.
10. Houben RM, Dowdy DW, Vassall A, Cohen T, Nicol MP, Granich RM, Shea JE, Eckhoff P, Dye C, Kimerling ME, White RG. TB MAC TB-HIV meeting participants. How can mathematical models advance tuberculosis control in high HIV prevalence settings? Int J Tuberc Lung Dis. 2014;18:509-14
11. Cao S, Wang F, Tam W, Tse LA, Kim JH, Liu J, Lu Z. A hybrid seasonal prediction model for tuberculosis incidence in China. BMC medical informatics and decision making. 2013 Dec;13(1):1-7.
12. Azeez A, Obaromi D, Odeyemi A, Ndege J, Muntabayi R. Seasonality and trend forecasting of tuberculosis prevalence data in Eastern Cape, South Africa, using a hybrid model. International journal of environmental research and public health. 2016 Aug;13(8):757.
13. Li Z, Wang Z, Song H, Liu Q, He B, Shi P, Ji Y, Xu D, Wang J. Application of a hybrid model in predicting the incidence of tuberculosis in a Chinese population. Infection and Drug Resistance. 2019;12:1011.
14. Zeming Li, Yanning Li A. comparative study on the prediction of the BP artificial neural network model and the ARIMA model in the incidence of AIDS. BMC medical informatics and decision making. 2020 Dec;20(1):1-3.
15. Zhou L, Xia J, Yu L, Wang Y, Shi Y, Cai S, Nie S. Using a hybrid model to forecast the prevalence of schistosomiasis in humans. International journal of environmental research and public health. 2016 Apr;13(4):355.

16. Yu L, Zhou L, Tan L, Jiang H, Wang Y, Wei S, Nie S. Application of a new hybrid model with seasonal auto-regressive integrated moving average (ARIMA) and nonlinear autoregressive neural network (NARNN) in forecasting incidence cases of HFMD in Shenzhen, China. PloS one. 2014 Jun 3;9(6):e98241.

17. Anokye R, Acheampong E, Owusu I, Isaac Obeng E. Time series analysis of malaria in Kumasi: Using ARIMA models to forecast future incidence. Cogent social sciences. 2018 Jan 1;4(1):1461544.

18. Achieng E, Otieno V, Mung'atu J. Modeling the trend of reported malaria cases in Kisumu county, Kenya. F1000Research. 2020 Jun 12;9:600.

19. Ren H, Li J, Yuan ZA, Hu JY, Yu Y, Lu YH. The development of a combined mathematical model to forecast the incidence of hepatitis E in Shanghai, China. BMC infectious diseases. 2013 Dec;13(1):1-6.

20. Zhang GP. A neural network ensemble method with jittered training data for time series forecasting. Information Sciences. 2007 Dec 1;177(23):5329-46.

21. Yolcu U, Egrioglu E, Aladag CH. A new linear & nonlinear artificial neural network model for time series forecasting. Decision support systems. 2013 Feb 1;54(3):1340-7.

22. Khashei M, Bijari M. A new class of hybrid models for time series forecasting. Expert Systems with Applications. 2012 Mar 1;39(4):4344-57.

23. Ministry of Health, Government of Kenya, Division of Leprosy, Tuberculosis, and Lung Disease. *National Monitoring and Evaluation Plan.* 2010

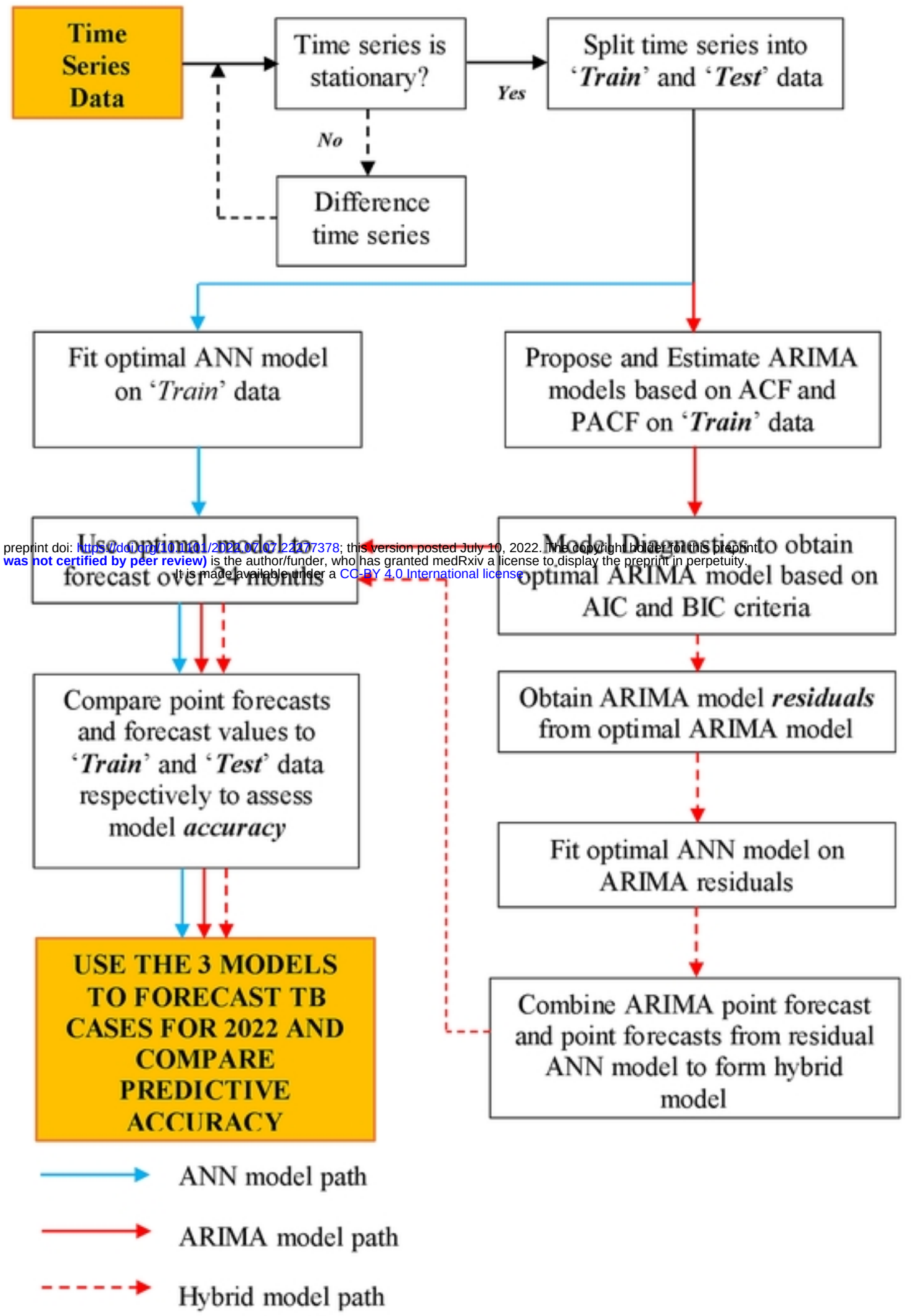
24. Ministry of Health, Government of Kenya, Division of Leprosy, Tuberculosis, and Lung Disease. *Annual Report*. 2012

25. James G, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning. New York: Springer; 2013 Jun.
26. Medar R, Rajpurohit VS, Rashmi B. Impact of training and testing data splits on the accuracy of time series forecasting in machine learning. In 2017 International Conference on Computing, Communication, Control and Automation (ICCUBEA) 2017 Aug 17 (pp. 1-6). IEEE.
27. Team RC. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www. R-project. org/. 2021.
28. Bakar NA, Rosbi S. Data clustering using Autoregressive Integrated Moving Average (ARIMA) model for Islamic country currency: an econometrics method for Islamic financial engineering. The International Journal of Engineering and Science (IJES). 2017;6(6):22-31.
29. Lee J. Univariate time series modeling and forecasting (Box-Jenkins method), Econ 413, Lecture 4. Department of Economics, University of Illinois. 2018.
30. GE P. Box, GM Jenkins, Time series analysis: forecasting and control, revised ed.
31. Shrivastav AK, Ekata D. Applicability of Box Jenkins ARIMA model in crime forecasting: A case study of counterfeiting in Gujarat state. Int J Adv Res Comput Eng Technol. 2012;1(4):494-7.
32. Khashei M, Bijari M. An artificial neural network (p, d, q) model for time-series forecasting. Expert Systems with applications. 2010 Jan 1;37(1):479-89.
33. Kihoro JM, Otieno RO, Wafula C. Seasonal time series forecasting: a comparative study of Arima and ann models. African Journal of Science and Technology. 2006.

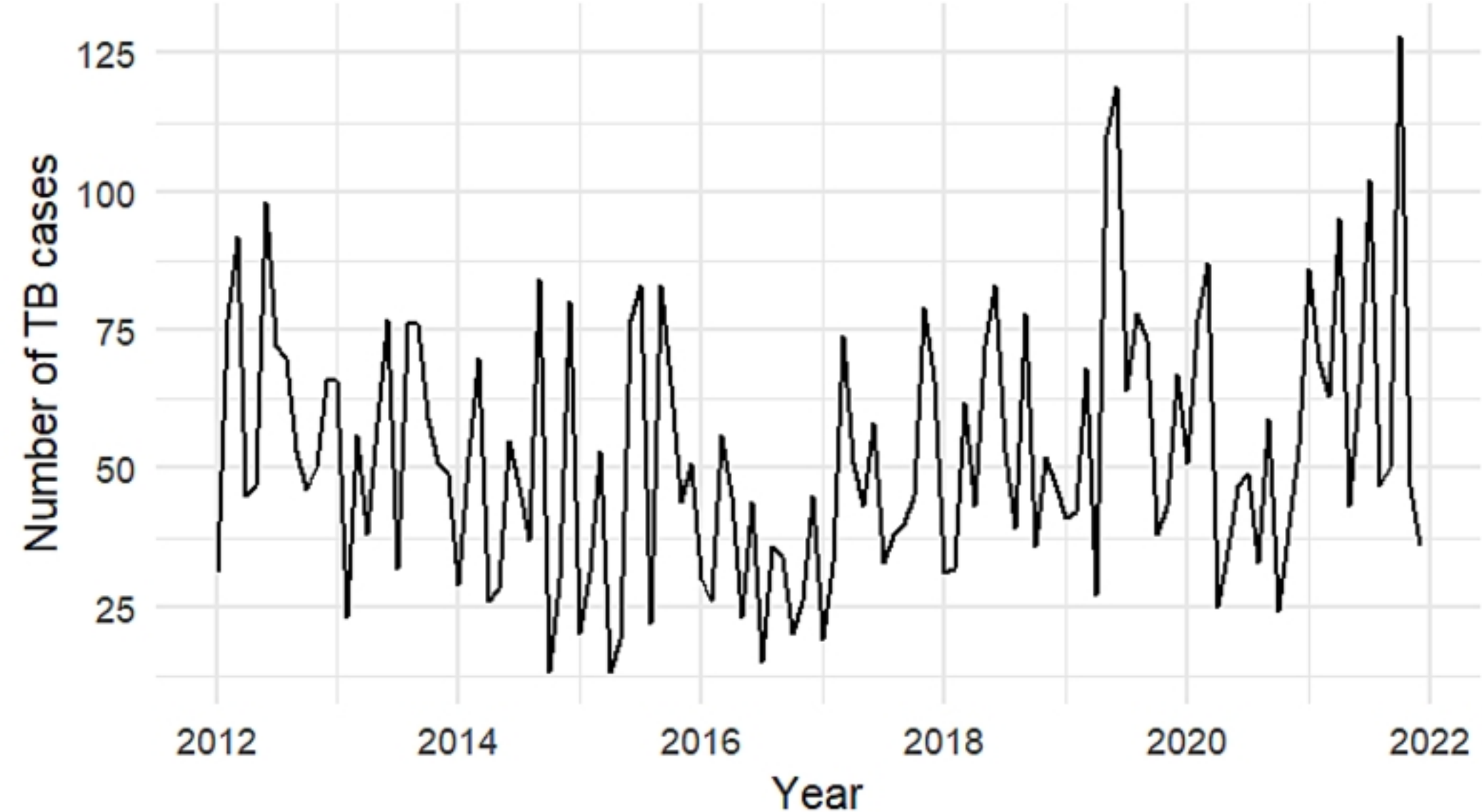
34. Larie D, An G, Cockrell RC. The Use of Artificial Neural Networks to Forecast the Behavior of Agent-Based Models of Pathophysiology: An Example Utilizing an Agent-Based Model of Sepsis. Frontiers in Physiology. 2021;12.
35. Cinar AC. Training feed-forward multi-layer perceptron artificial neural networks with a tree-seed algorithm. Arabian Journal for Science and Engineering. 2020 Dec;45(12):10915-
36. Darji MP, Dabhi VK, Prajapati HB. Rainfall forecasting using neural network: A survey. In 2015 international conference on advances in computer engineering and applications 2015 Mar 19 (pp. 706-713). IEEE.
37. Zhang GP. Time series forecasting using a hybrid ARIMA and neural network model. Neurocomputing. 2003 Jan 1;50:159-75.
38. Chakrabarti A, Ghosh JK. AIC, BIC, and recent advances in model selection. Philosophy of statistics. 2011 Jan 1:583-605.
39. Ljung GM, Box GE. On a measure of lack of fit in time series models. Biometrika. 1978 Aug 1;65(2):297-303.
40. Hamzaçebi C. Improving artificial neural networks' performance in seasonal time series forecasting. Information Sciences. 2008 Dec 1;178(23):4550-9.
41. Lewis C. International and Business Forecasting Methods Butterworths: London. 1982.
42. DIEBOLD F, MARIANO R. Comparing predictive accuracy. journal of business and Economics Statistics, v. 13. 1995.
43. Nyoni SP, Nyoni T. Modeling and Forecasting TB Incidence in Bolivia Using the Multilayer Perceptron Neural Network. International Research Journal of Innovations in Engineering and Technology. 2021 Mar 1;5(3):301.

559 44. Zhou L, Zhao P, Wu D, Cheng C, Huang H. Time series model for forecasting the number of
560 new admission inpatients. BMC medical informatics and decision making. 2018 Dec;18(1):1-
561 1.
562 45. Chakraborty T, Chakraborty AK, Biswas M, Banerjee S, Bhattacharya S. Unemployment rate
563 forecasting: A hybrid approach. Computational Economics. 2021 Jan;57(1):183-201.
564 46. Okwara FN, Oyore JP, Were FN, Gwer S. Correlates of isoniazid preventive therapy failure
565 in child household contacts with infectious tuberculosis in high burden settings in Nairobi,
566 Kenya–a cohort study. BMC infectious diseases. 2017 Dec;17(1):1-1.
567 47. https://www.census.gov/datatools/demo/idb/#/table?COUNTRY_YEAR=2022&COUNTRY
568 _YR_ANIM=2022
569 48. WHO Global Tuberculosis Report. 2021
570 49. Oliwa JN, Maina J, Ayieko P, Gathara D, Kathure IA, Masini E, Van't Hoog AH, van
571 Hensbroek MB, English M. Variability in distribution and use of tuberculosis diagnostic tests
572 in Kenya: a cross-sectional survey. BMC infectious diseases. 2018 Dec;18(1):1-3.
573 50. Mbithi I, Thekkur P, Chakaya JM, Onyango E, Owiti P, Njeri NC, Kumar AM,
574 Satyanarayana S, Shewade HD, Khogali M, Zachariah R. Assessing the real-time impact of
575 COVID-19 on TB and HIV services: the experience and response from selected health
576 facilities in Nairobi, Kenya. Tropical Medicine and Infectious Disease. 2021 May 10;6(2):74.
577 51. Makori L, Gichana H, Oyugi E, Nyale G, Ransom J. Tuberculosis in an urban hospital
578 setting: Descriptive epidemiology among patients at Kenyatta National Hospital TB clinic,
579 Nairobi, Kenya. International Journal of Africa Nursing Sciences. 2021 Jan 1;15:100308.

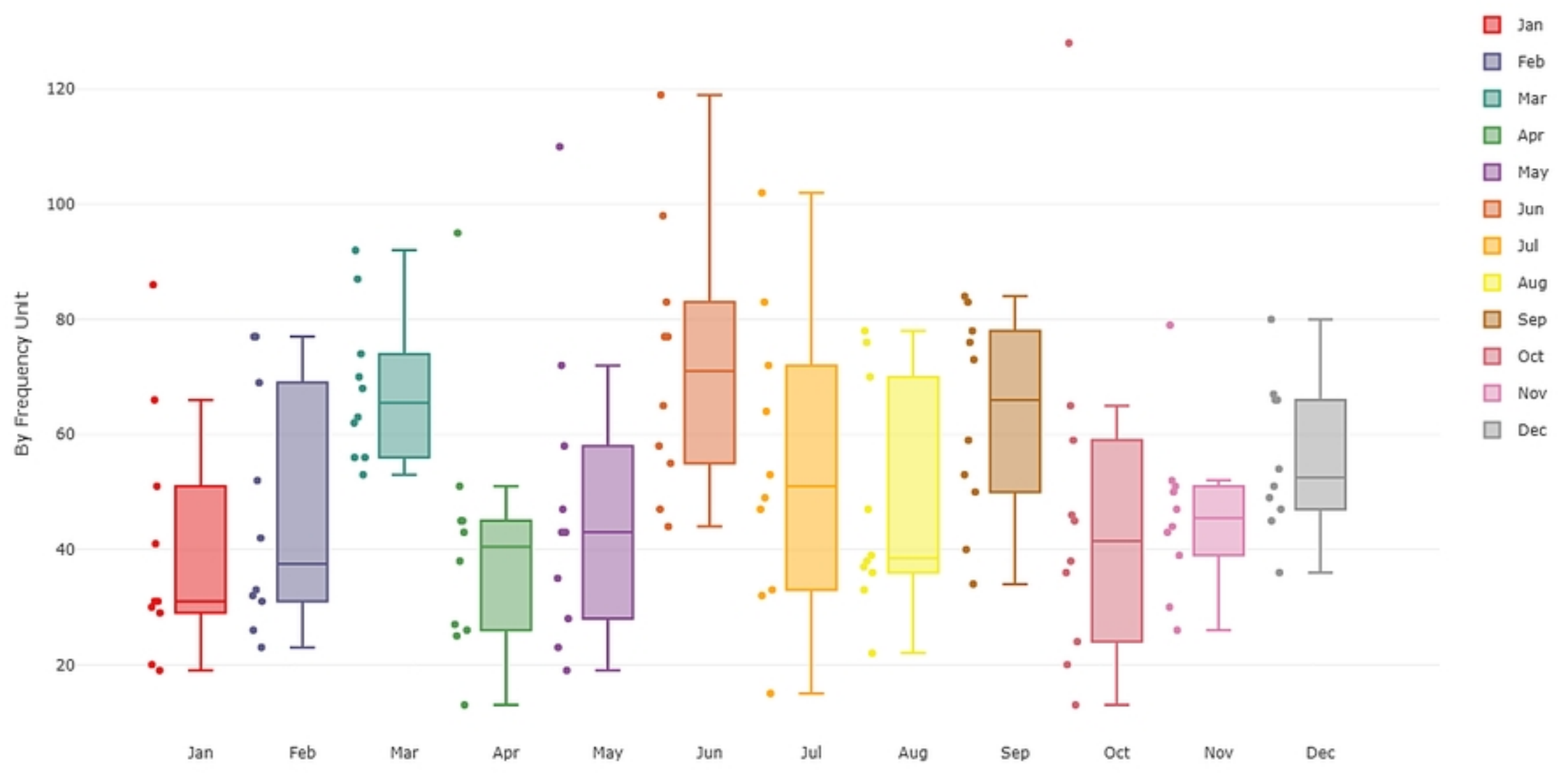
580



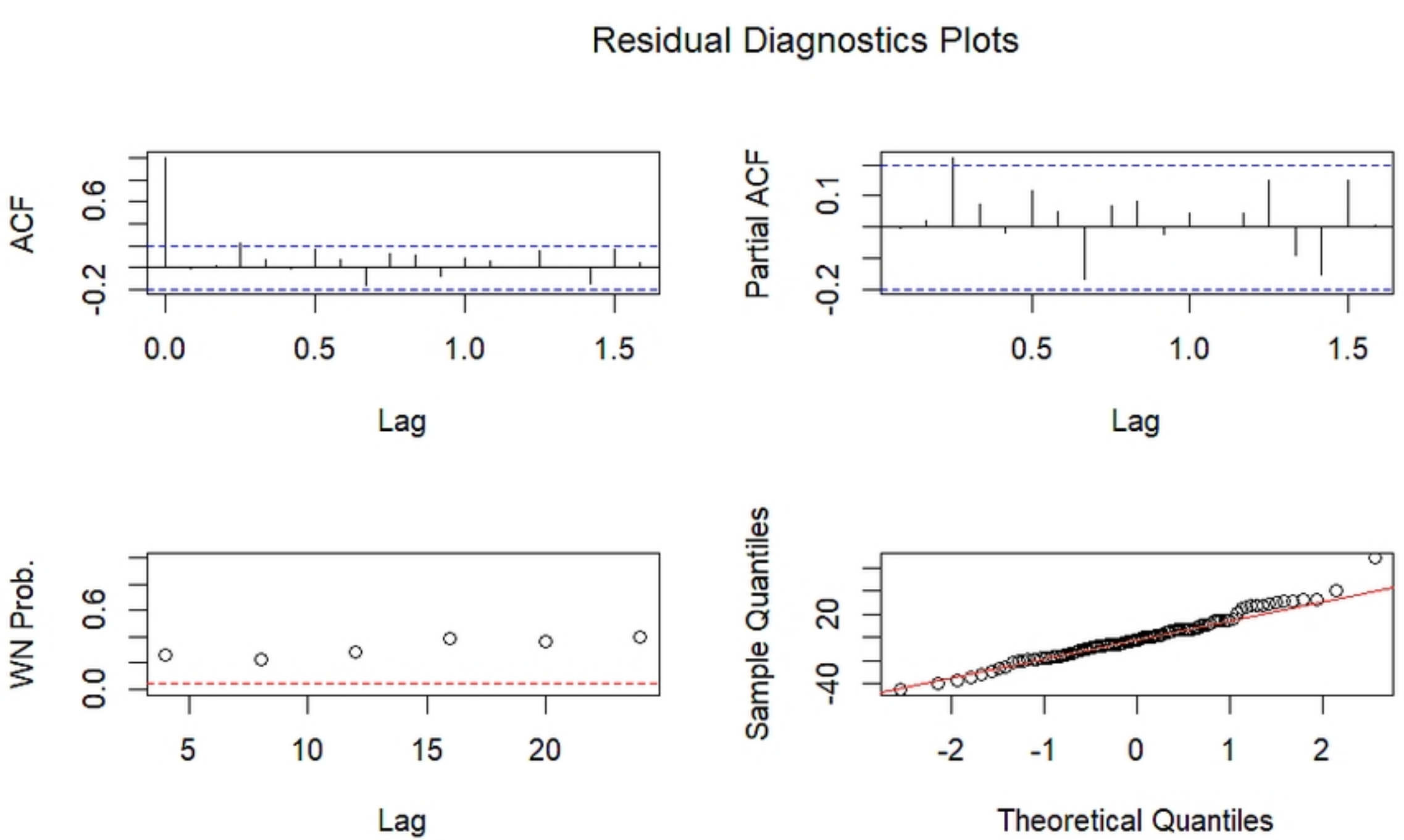
The Proposed methodology



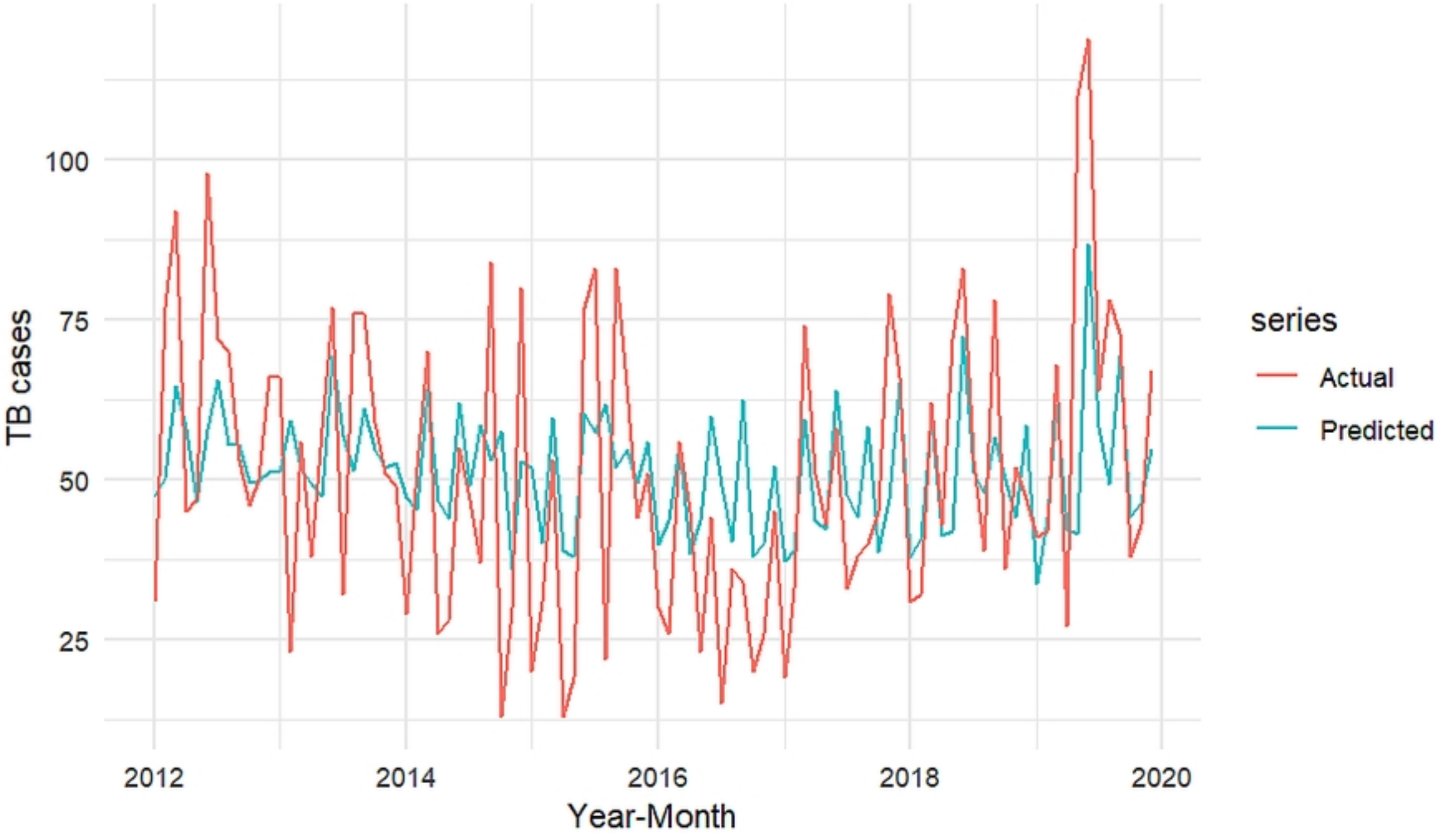
Monthly TB cases among children below 15 years from H



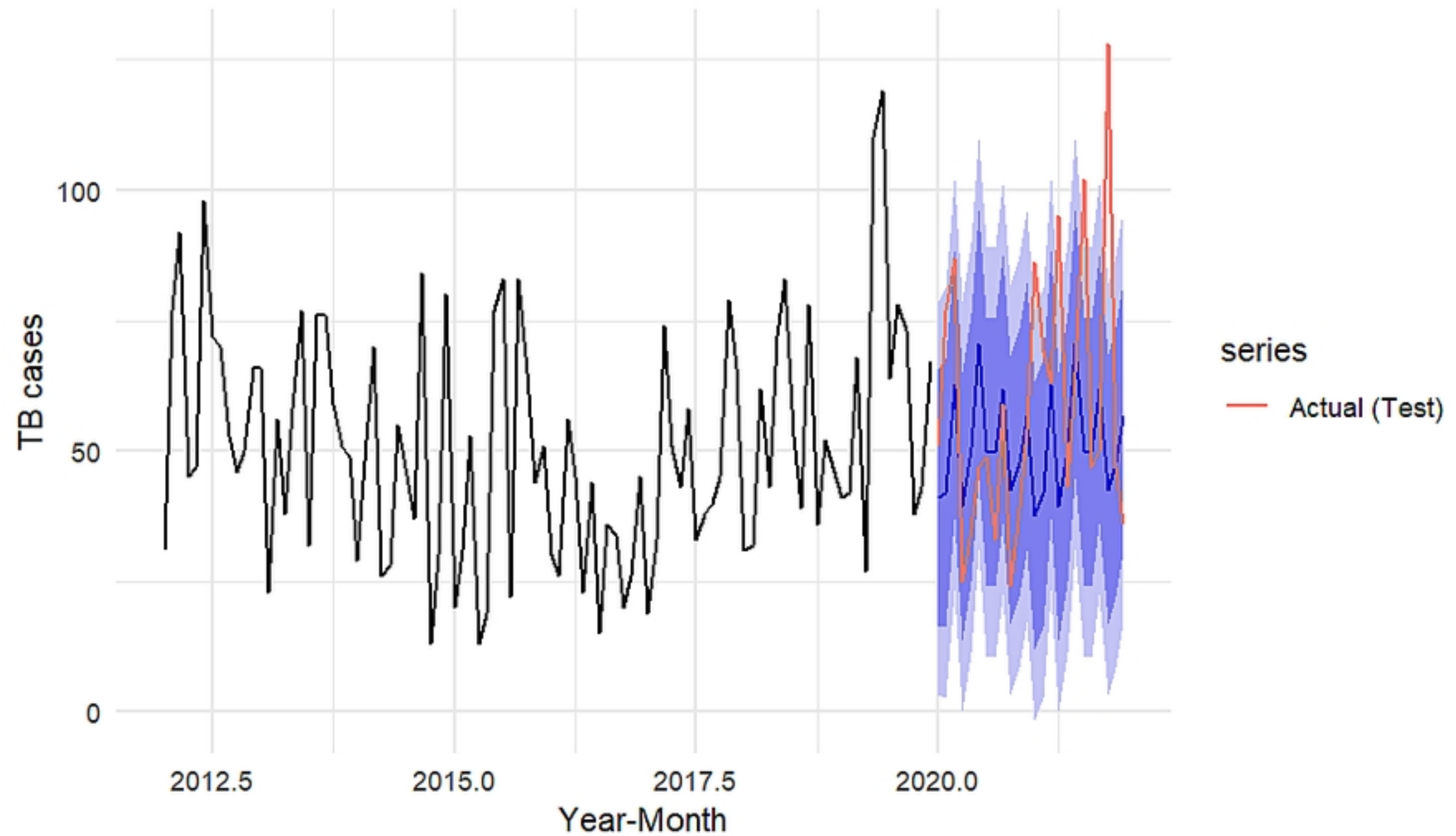
Monthly cycle plot of TB cases



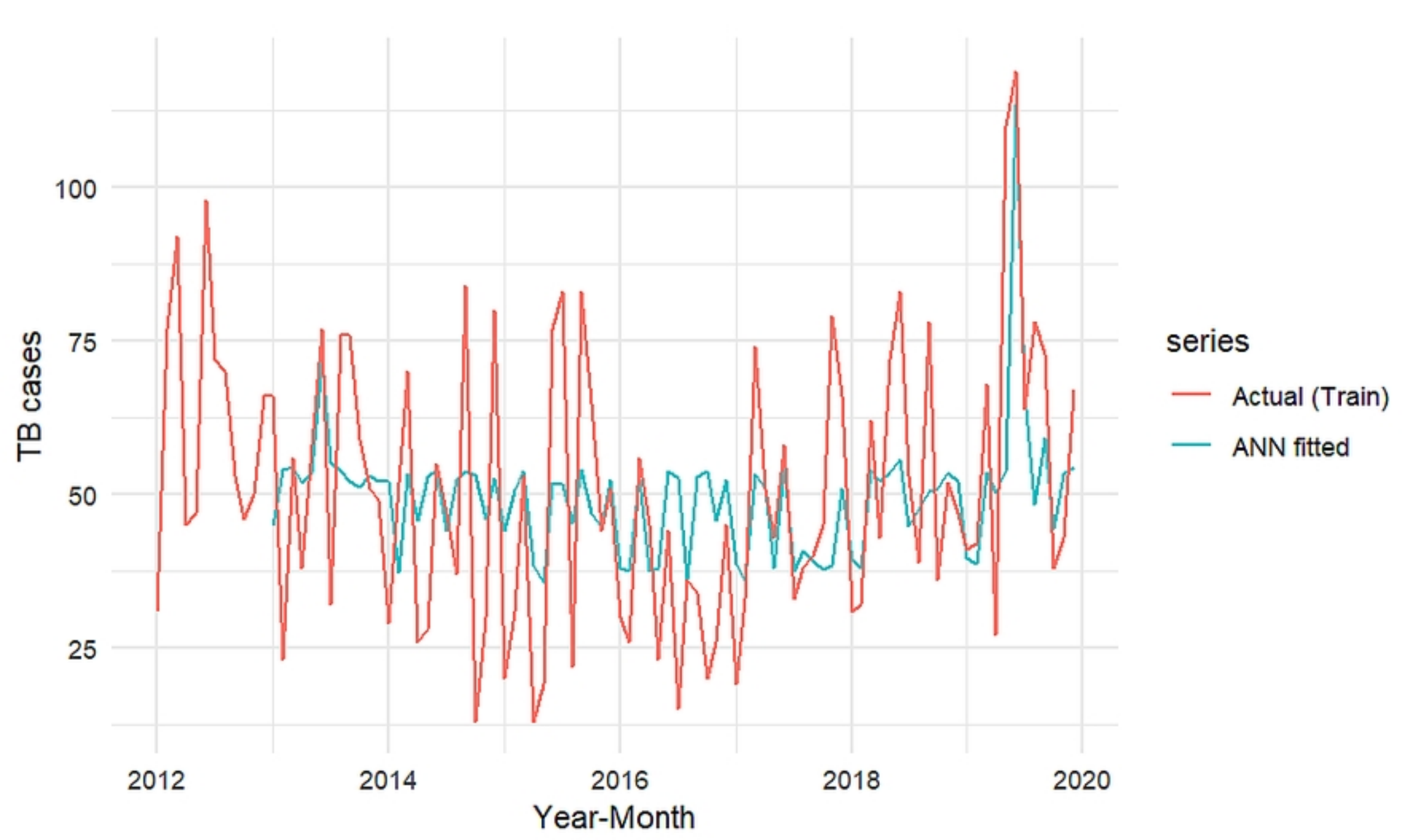
ARIMA Model residual diagnostics



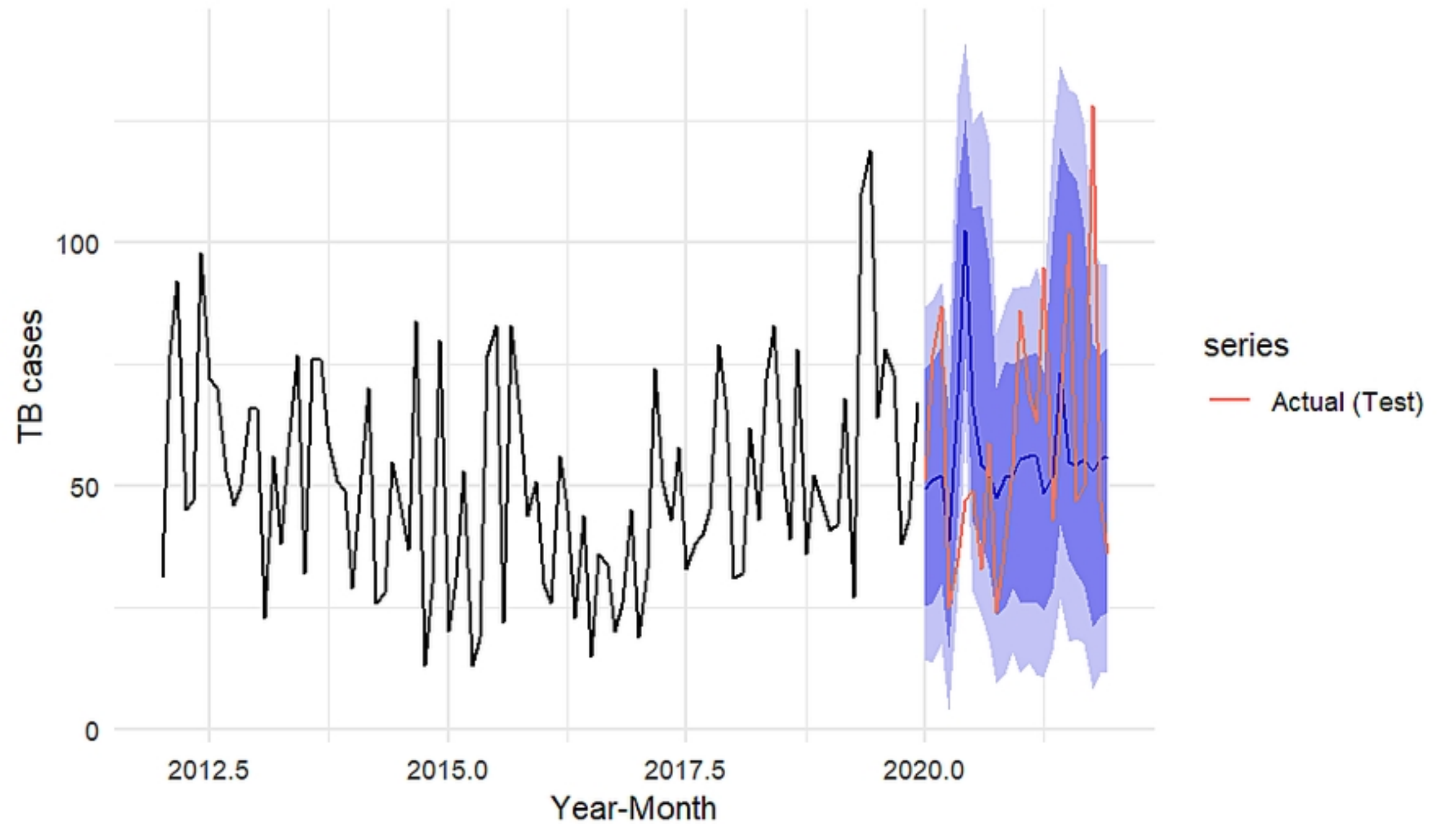
Comparison of ARIMA predicted versus actual (training data) TB



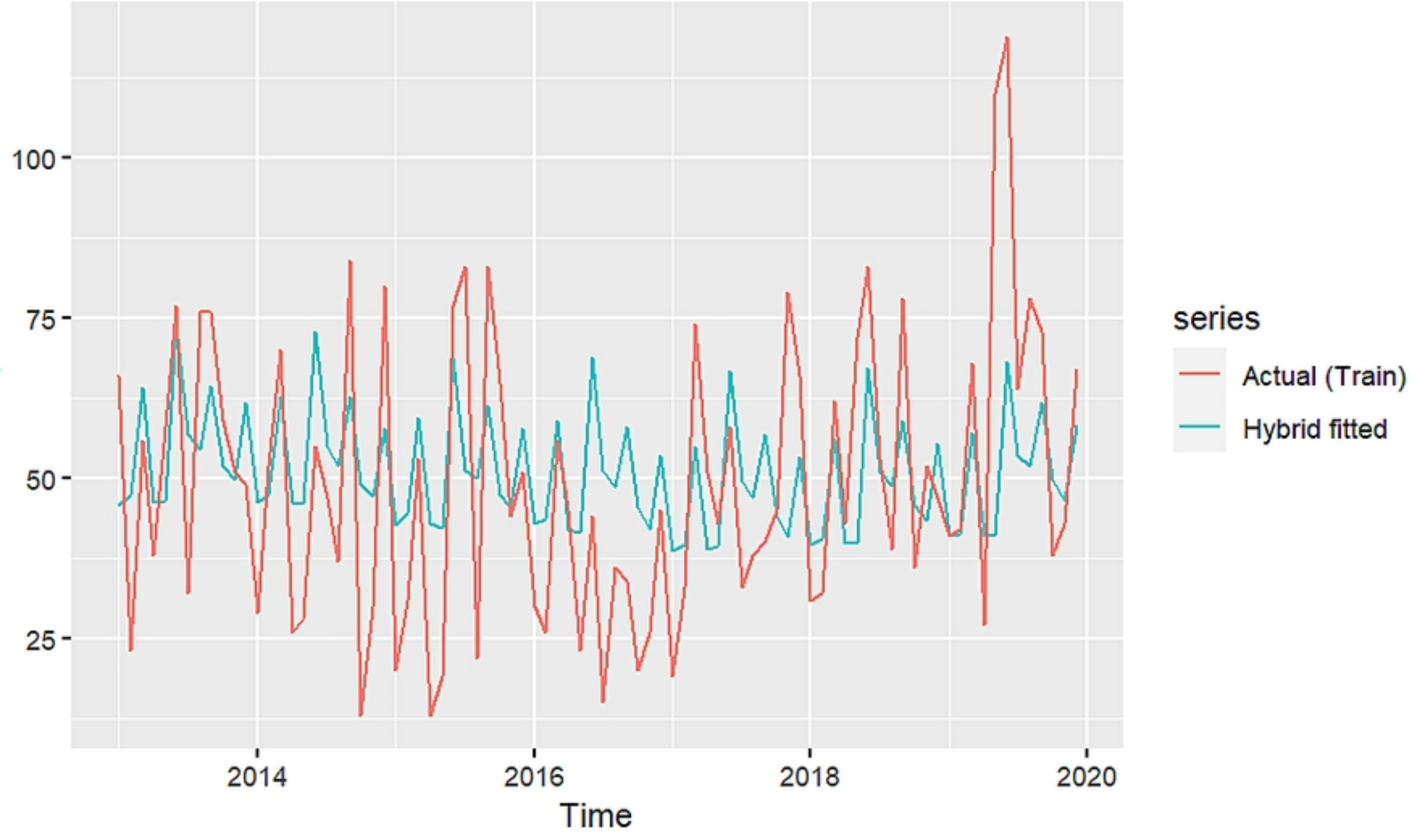
Comparison of ARIMA 24-month forecast versus actual (test dat



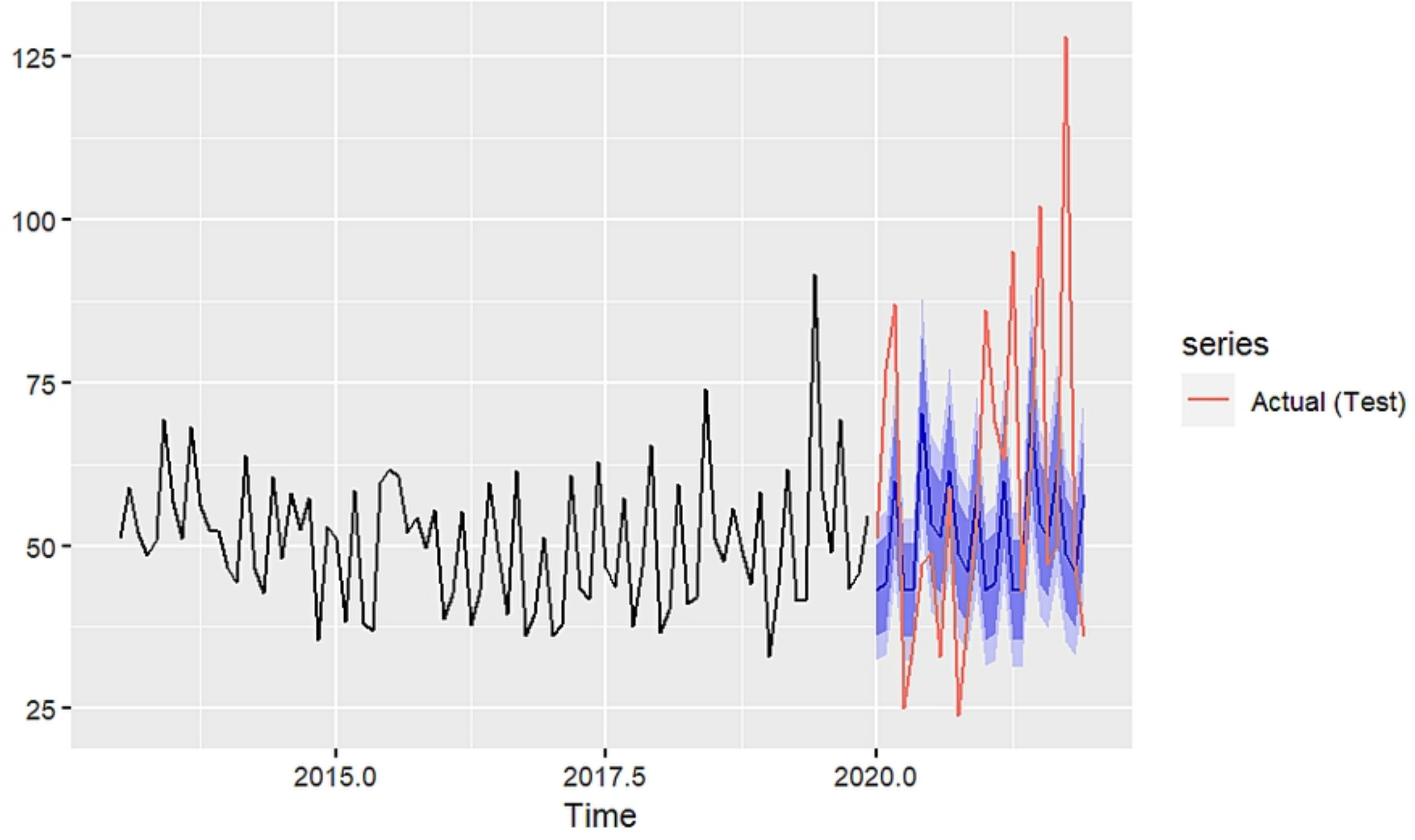
Comparison of NNAR (1,1,2) [12] predicted TB cases versus actua



Comparison of NNAR (1,1,2) [12] 24-month forecast TB cases ver



Comparison of ARIMA-ANN predicted TB cases versus actual TB



Comparison of ARIMA-ANN 24-month forecasted TB cases versu