Forecasting seasonality and trend of tuberculosis prevalence in India using SARIMA-NNAR Hybrid model

Baikunth Kumar Yadav Baikunth  
Mahatma Gandhi Central University, Bihar, Motihari 845401

Sunil Kumar Srivastava Sunil  
Mahatma Gandhi Central University, Bihar, Motihari 845401

Pranveer Singh Pranveer (✉ mailpranveer@gmail.com )  
Mahatma Gandhi Central University, Bihar, Motihari 845401

Research Article

Keywords: Tuberculosis, Forecasting, Seasonality, Trend analysis, Hybrid model

Posted Date: November 4th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-2228940/v1

License: ☝️ 🔒 This work is licensed under a Creative Commons Attribution 4.0 International License.  
Read Full License
Abstract

Early detection of tuberculosis (TB) is very important for control and prevention. We aim to study the current status of tuberculosis prevalence in India by applying an appropriate model that can forecast TB incidence by analyzing the seasonality and trends using past time series data. Notified TB incidence data was extracted from open resources, i.e., Central Tuberculosis Division (CTB), Government of India from 2017 to 2019. A SARIMA model and a hybrid model combining SARIMA with Neural Network Autoregressive (SARIMA-NNAR) models were applied to fit the data and forecast the notified TB incidence. Counterfeiting performance parameters, MSE, RMSE, MAE, MPE, MASE, and MAPE were used to analyze the goodness fit of the models. Reported notified TB incidence data of 2020 were used to validate the models. Both models could reasonably predict and forecast the notified TB incidence, but a hybrid model demonstrated better results when compared to the individual models. In the hybrid model, the RMSE, MAE, and MAPE were (5260.359), (3910.648), and (2.080665) respectively whereas in the SARIMA model the corresponding values were (6712.889), (4863.659), and (2.644319) respectively. Therefore, the hybrid model was more effective in predicting the seasonality and trend of TB incidence than the individual SARIMA model. The hybrid model showed better TB incidence forecasting than the SARIMA. This model will help the government to develop better control strategies for the overall management, control, and eradication of the disease.

1 Introduction

Tuberculosis (TB) is a highly contagious disease caused by Mycobacterium tuberculosis [1]. TB is the tenth leading cause of death globally in the background of HIV-AIDS co-infection [2]. TB spreads from one person to another through aerosol droplets when people who are sick with TB, expel bacteria into the air by coughing and sneezing. When it affects the lungs, it is referred to as pulmonary TB (PTB) and when other parts of the body are involved, it is referred to as extrapulmonary TB (EPTB) [2]. In 2019, approximately 10 million people were diagnosed with TB and around 1.4 million people died from this disease worldwide [2]. In the 1990s, World Health Organization (WHO) came up with a twin strategy of directly observed treatments (DOTs) and short-course chemotherapy to combat TB. Despite worldwide concerted efforts to combat TB, it remains a major global health issue, particularly in lower and middle incomes countries or developing countries [3]. These countries also share the additional burden of multidrug-resistant (MDR), and the extensive drug-resistant (XDR) TB along with coinfection with human immunodeficiency virus (HIV) leading to acquired immunodeficiency syndrome (AIDS) [4]. As per the WHO report the highest TB incidence has occurred in southeast Asia, Africa, and the Western Pacific with (44%), (25%), and (18%) cases-loads respectively, while the Eastern Mediterranean, the United States of Americas (USA), and Europe show smaller percentage (8.2%), (2.9%), and (2.5%), respectively[2]. India is the highest TB burden country with an estimated 26% of the infectious population in 2019. The 30 most infectious countries that have the highest TB incidence (81% of global TB cases) were reported in 2019 out of which the top five countries are India, Indonesia, China, Philippines, and Pakistan [2].
World Health Organization (WHO) along with United Nations Organizations (UNO) in their 2014 and 2015 chapters have resolved to eradicate the TB by 2020 under its ‘End TB Strategy’ and the ‘UN Sustainable Development Goals (SDGs)’ programs. However, only European, and African regions to some extents were successful in achieving these goals [2]. The quantum of TB incidence show variation at the national level that ranged from (> 5) cases to (< 500) cases per 100000 population. The synergistic efforts globally have led to the reduction of TB incidence rates by 2% per year, however, the milestone of 20% global reduction from 2015 to 2020 is far away to reach [2].

India is home to around one-third of the global TB cases that also includes various drug-resistant cases and EPTB. The Indian government has launched an ambitious and massive TB control drive rechristened as the Revised National TB Control Program (RNTCP, also RNTCP-I) in 1997. By 2006, it has entered into the second phase (RNTCP-II) and covered the whole of India. The second phase (RNTCP-II) was launched to consolidate the gains made in the first phase (RNTCP-I); to address the issues of TB/HIV, and MDR-TB, and to cover the private sector. In 2020, RNTCP was renamed as National Treatment Elimination Program (NTEP) with a holy-grail to eliminate TB by 2025. RNTCP uses WHO’s recommended Directly Observed Treatment Short Course (DOTS) and has covered over a billion people in 632 districts/reporting units. RNTCP is entrusted to carry out the WHOs and UNs recommended 5-year National TB Strategic Plans. The Joint TB Monitoring Mission (JTMM) has brought national and international experts together under the umbrella of RNTCP to review the progress, challenges, plans, and efforts of RNTCP to control TB in India [5].

Seasonality is one very important characteristic of TB that helps in predicting the TB epidemics. The seasonality effect in TB epidemics helps in utilizing the existing TB data more efficiently and effectively and assists in incorporating and analyzing the environmental and social factors influencing the TB epidemics. There are innumerable studies that have analyzed the impact of the seasonal pattern of TB across countries [6]. For instance, China observed a higher peak of TB incidence in the summer season [7], South Africa has reported higher TB incidence in late winter and early spring season[8], Pakistan reported the peak in the third and fourth quarter of the year [9], Iran reported higher TB incidence in spring and summer season [10], Portugal has reported peak of TB incidence in March and December [11] and India reported its peak season in spring and winter [12, 13]. These studies revealed that the seasonal pattern of notified TB incidence increase in the late winter and early spring season [14, 15].

Thorpe and co-workers (2004) have pioneered the study that has elucidated the trends and patterns of TB in India. This group has utilized data from RNTCP and calculated seasonal variations for both PTB and EPTB, which peaked from April to June and dropped during October to December in northern Indian districts while no such trend was reported in the southern districts. Regions in the north India show the highest seasonal variations while the southern parts showed low or no seasonal variations. The seasonality trend was strongest in the children which decreases towards the older age people [12].

Some retrospective studies were conducted based on the seasonality and trend of notified TB incidences [8, 9, 13]. Various models have been applied to forecast the notified TB incidence to analyze the trend and
predict the trajectory of the TB epidemics in the country [6, 16]. 

The main aim of this study is to compare the efficacy of the existing seasonal autoregressive integrated moving average (SARIMA) model with the combination of the SARIMA and the Neural Network Autoregressive model, i.e., SARIMA-NNAR hybrid model to forecast and assess the seasonality and trend of the notified TB incidences in India. The SARIMA model is largely applied in the case of infectious disease to predict the pattern of the epidemic as the model of choice globally [17–19]. However, this model largely relies on linear information which is a limitation [11, 20–22]. There are other models such as autoregressive integrated moving average (ARIMA) basic model [23], general regression models, grey models [23], Markov chain model [24], and the neural network model [25] that can be applied for forecasting the infectious disease trajectory. The submitted study has adopted SARIMA and SARIMA-NNAR hybrid model to forecast the seasonality and trend of notified TB incidences in India based on the linear and non-linear information.

2 Materials And Methods

2.1 Data Collection

The monthly prevalence data of Notified Tuberculosis from January 2017 to December 2019 were collected from the official website of the Central Tuberculosis Division (CTB), Government of India, and data were incorporated in Excel 2019, to build a time-series database. The database was analyzed separately for each State/province of India.

2.2 Model Development

This study aims to analyze the efficacy of the individual SARIMA and hybrid SARIMA-NNAR models with respect to the forecasting and prediction based on notified TB incidence data. The time-series data of tuberculosis incidence from January 2017 to December 2019 is shown in Fig. 1 for visualizing and exploring the pattern of TB over time. To identify the seasonality and trend in TB time series data, an additive decomposition was performed Fig. 2. Additive decomposition also estimates the seasonal effects, which were shown as seasonal values. Seasonality adjustment was done to remove the seasonal effects and the plot was used to show whether the trend and seasonality exists Fig. 3.

This study is based on an autoregressive integrated moving average (ARIMA) model, which uses time-series data to predict the future TB incidence based on the past TB incidence. Seasonal autoregressive integrated moving average (SARIMA) is an extension of ARIMA that incorporates the time series data of the seasonal component. Furthermore, the Neural Network Auto-Regressive (NNAR) model is used for complex nonlinear relationships between the response variable and its predictors but it can be used as a complement of linear analysis [26]. In this study, the SARIMA-NNAR hybrid model was used for analyzing the trend of time series data of TB and its seasonal components that predicts the monthly TB incidences in India.

2.3 SARIMA Model
Smoothing was imposed on the time series to discern the underlying patterns due to high-frequency variations in the time series data. SARIMA was modelled with Box and Jenkin strategy including the augmented Dicky Fuller Test (ADF) and Ljung-Box portmanteau to check if the data is stationary or not. This was followed by the autocorrelation coefficient (ACF) and partial autocorrelation coefficient (PACF) of the stationary sequence to identify optional model parameters (P, D, Q, and p, d, q) to establish one or more alternative models Fig. 4. After the confirmation of the stationarity test, we have selected the Seasonal ARIMA, which is originated from ARIMA Model. Seasonal-ARIMA (SARIMA) model includes both non-seasonal (p, d, q) and seasonal (P, D, Q) factors to account for the seasonality of the time series data in a uniform pattern that repeats over the S period until the seasonal cycle changes again. In a SARIMA model, Auto-regressive (AR) and moving-average (MA) terms predict $x_t$ using data values and errors at times with lags that are multiples of S (the span of the seasonality) [6, 19]. The general SARIMA model includes non-seasonal and seasonal factors that are given below:

$$SARIMA (pdq) \ (PDQ)^s$$

1

Where, $p$ = non-seasonal AR order, $d$ = non-seasonal differencing, $q$ = non-seasonal MA order, while $P$ = seasonal AR order, $D$ = seasonal differencing, $Q$ = seasonal MA order, and $S$ = time span of repeating seasonal pattern.

The general mathematical SARIMA model is given as:

$$\varphi (B^s) \phi (B) (x_t - \mu) = \Theta (B^s) \theta (B) \epsilon_t$$

2

The non-seasonal components are:

$$AR : \phi (B) = 1 - \phi_1 B - \cdots - \phi_p B^p$$

3

$$MA : \theta (B) = 1 - \theta_1 B + \cdots + \theta_q B^q$$

4

The seasonal components are:

$$SeasonalAR : \varphi (B) = 1 - \varphi_1 B - \cdots - \varphi_P B^{PS}$$

5

$$SeasonalMR : \Theta (B) = 1 - \Theta_1 B + \cdots + \Theta_Q B^{QS}$$

6
In this equation, B represents a backward shift, $\epsilon_t$ stands for estimated residual error at $t$ for $\mu = 0$ and $\sigma^2$ and $x_t$ represents observed value at $t$ (1, 2, . . . , k), $\varphi$ is a vector for AR coefficient, $\theta$ is a vector of MA coefficients, $\Phi$ is a vector of seasonal AR coefficients, and $\Theta$ is a vector of seasonal MA coefficients. In a SARIMA model, seasonal values are removed in appropriate order from non-stationary data of time series. Seasonal differencing is defined as a difference between a value and a corresponding value with lag that is a multiple of $(S) = 12$ and it is expressed as $x_t = y_t - y_{t-s}$.

SARIMA model utilizes the autocorrelation function (ACF) and partial autocorrelation function (PACF) to check the reliability of time series data and to select the parameters (p, P, d, D, q, Q) of the components and should comply with the parametric and residual tests. ACF correlates with the previous time-series data of TB whereas PACF correlates with time-series data with its time series criterion lagged values. Akaike information criterion (AIC) and Bayesian information criterion (BIC) were also applied to select the best model for time series data of TB. Both AIC and BIC were penalized on the likelihood criterion. Lower values of AIC and BIC are considered as the best model selection for the time series of TB. Both the models SARIMA and SARIMA-NNAR were built in R-Program (v-4.0.5 Network Theory Ltd., Bristol, UK).

2.4 SARIMA-NNAR Model

Artificial neural networks are forecasting methods that are based on the general mathematical brain of the models. It is a self-organizing learning process. Generally, the learning process is used to adjust the weights of neural networks based on long and short stochastic function dependence upon the time series data. NNAR model applied on 3-years’ time series data of TB with the lag values that can be used as input variables for neural network autoregression. We considered only a feed-forward neural networks autoregressive model for forecasting the time series data of TB with one hidden layer, and we use the indication NNAR (p, k) where p indicates lagged inputs $x_t$ and k indicates nodes in one hidden layer. ANNAR (p, 0) model is equivalent to ARIMA (p, 0, 0) model, but without the restrictions on the parameters to make ensure stationarity. It also helps in adding last observation values in seasonal data from the same time as inputs. Inputs of this model is a learning process that employs the autoregressive neural network with the consideration of long and short term of the time series. The input of this model has $(y_{t-1}, y_{t-2}, \ldots , y_{t-s})$ and k neurons in the hidden layer [8].

SARIMA model was employed to examine the linear relationship of the time series data of TB while residual part nonlinear relation has been examined by the hybrid SARIMA-NNAR model [8, 27]. In this hybrid model, both linear and nonlinear sections are combined. Two input variables were selected for the estimated occurrence of TB at time $t$.

2.5 Accuracy Evaluation Methods

Comparison between the SARIMA and SARIMA-NNAR with their performance and simulation with the help of six parameters demonstrate the errors and goodness of fit of the models. The most common methods to evaluate the accuracy are mean square error (MSE), root means square error (RMSE), mean
absolute error (MAE), mean percent error (MPE), mean absolute scaled error (MASE), and mean absolute percentage error (MAPE).

General formulations of these methods are given below:

\[
MSE = \frac{1}{n} \sum_{t=1}^{n} (x_t - \hat{x}_t)^2
\]

\[
MAE = \frac{1}{n} \sum_{t=1}^{n} |x_t - \hat{x}_t|
\]

\[
MAPE = \frac{1}{n} \sum_{t=1}^{n} \frac{|x_t - \hat{x}_t|}{x_t}
\]

\[
MPE = \frac{1}{n} \sum_{t=1}^{n} \frac{A_t - F_t}{A_t}
\]

\[
RMSE = \sqrt{\frac{1}{n} \sum_{t=1}^{n} (x_t - \hat{x}_t)}
\]

Here \(x_t\) is the actual incidence, \(\hat{x}_t\) is estimated incidence, \(n\) is prediction number, \(A_t\) is the actual value of the portion being forecast, and \(F_t\) is a forecast value.

### 3 Results

A time-series analysis of notified TB incidences from 2017 to 2019 was conducted and a total of 6237046 incidences were reported over the years in the entire States and the Union territory of India. An average of 173251 TB incidence per month was reported. According to the Central Tuberculosis Division (CTB), Government of India, the annual notified TB incidence rate was 156 cases per 100000 population over the years. The highest notified incidence rate was observed in Chandigarh and Delhi with an annual case detection rate as 475 and 417 cases per 100000 population respectively. The notified TB incidence rate for the Indian States and Union Territories are given in Table 1 and also represented through the map of India Fig. 5. The average marginally increasing trend was 12.05% from the years 2017 to 2019 in India. During the study period, seasonal variation revealed March, April, May to be the peak months with incidence rates as 12.13%, 12.15%, 12.15% per 100000 populations respectively. The additive decomposition for the TB incidence series revealed that the seasonal pattern changed every 12 months Fig. 2 with the peak values in March, April, and May. Therefore, the notified TB incidence has shown that the seasonal trend is involved in seasonal indices.
Table 1
Annual case detection rate per 100000 population in Indian states during 2017 to 2019

<table>
<thead>
<tr>
<th>S. No</th>
<th>States &amp; UTs</th>
<th>Rate/100000 Population</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Andaman &amp; Nicobar Islands</td>
<td>150</td>
</tr>
<tr>
<td>2</td>
<td>Andhra Pradesh</td>
<td>172</td>
</tr>
<tr>
<td>3</td>
<td>Arunachal Pradesh</td>
<td>206</td>
</tr>
<tr>
<td>4</td>
<td>Assam</td>
<td>177</td>
</tr>
<tr>
<td>5</td>
<td>Bihar</td>
<td>92</td>
</tr>
<tr>
<td>6</td>
<td>Chandigarh</td>
<td>475</td>
</tr>
<tr>
<td>7</td>
<td>Chhattisgarh</td>
<td>144</td>
</tr>
<tr>
<td>8</td>
<td>Dadra and Nagar Haveli and Daman and Diu</td>
<td>148</td>
</tr>
<tr>
<td>9</td>
<td>Delhi</td>
<td>417</td>
</tr>
<tr>
<td>10</td>
<td>Goa</td>
<td>141</td>
</tr>
<tr>
<td>11</td>
<td>Gujarat</td>
<td>222</td>
</tr>
<tr>
<td>12</td>
<td>Haryana</td>
<td>206</td>
</tr>
<tr>
<td>13</td>
<td>Himachal Pradesh</td>
<td>221</td>
</tr>
<tr>
<td>14</td>
<td>Jammu &amp; Kashmir</td>
<td>81</td>
</tr>
<tr>
<td>15</td>
<td>Jharkhand</td>
<td>132</td>
</tr>
<tr>
<td>16</td>
<td>Karnataka</td>
<td>127</td>
</tr>
<tr>
<td>17</td>
<td>Kerala</td>
<td>70</td>
</tr>
<tr>
<td>18</td>
<td>Ladakh</td>
<td>123</td>
</tr>
<tr>
<td>19</td>
<td>Lakshadweep</td>
<td>41</td>
</tr>
<tr>
<td>20</td>
<td>Madhya Pradesh</td>
<td>193</td>
</tr>
<tr>
<td>21</td>
<td>Maharashtra</td>
<td>161</td>
</tr>
<tr>
<td>22</td>
<td>Manipur</td>
<td>84</td>
</tr>
<tr>
<td>23</td>
<td>Meghalaya</td>
<td>144</td>
</tr>
<tr>
<td>24</td>
<td>Mizoram</td>
<td>226</td>
</tr>
<tr>
<td>25</td>
<td>Nagaland</td>
<td>193</td>
</tr>
<tr>
<td>26</td>
<td>Odisha</td>
<td>117</td>
</tr>
</tbody>
</table>
The SARIMA model was applied to the time-series data of notified TB incidences with the characteristics of seasonal tendency. Based on the autocorrelation function plot and partial autocorrelation function plot the key parameters (p, P, d, D, q, Q) of the SARIMA Model were selected Fig. 4. The best model was generated from notified TB incidence data after applying auto.arima() function, was SARIMA (1, 0, 0)(1, 1, 0)_{12} where 12 is monthly time series data. The most preferred model was selected based on minimum values of AIC and BIC. The result is summarized in Table 2. The estimates and standard error of SARIMA (1, 0, 0) (1, 1, 0)_{12} model parameters and their significant values are given in Table 2.

### Table 2

<table>
<thead>
<tr>
<th>Measurements</th>
<th>Models</th>
<th>Estimates</th>
<th>Standard Error</th>
<th>z - Value</th>
<th>p-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-Seasonality</td>
<td>Ar1</td>
<td>0.4267</td>
<td>0.1981</td>
<td>2.1531</td>
<td>0.05</td>
</tr>
<tr>
<td>Seasonality</td>
<td>Sr1</td>
<td>-0.7452</td>
<td>0.1350</td>
<td>-5.5172</td>
<td>0.00</td>
</tr>
<tr>
<td>Coefficient</td>
<td></td>
<td>2272.9359</td>
<td>178.6762</td>
<td>12.7210</td>
<td>0.00</td>
</tr>
</tbody>
</table>

In the SARIMA-NNAR hybrid model, the hybrid model NNAR (3, 1, 2)_{12} was obtained by using nnetar and forecast.nnetar function on time series data of the notified TB incidence. The simulation accuracy of the NNAR model was determined by applying a smoothing factor (α) from the range 0.1 to 1.0. The lowest ME, RMSE, MAE, MPE, MAPE, MASE values were obtained in the hybrid model at the smoothing factor α = 0.1.

SARIMA (1, 0, 0)(1, 1, 0)_{12} was compared with the SARIMA-NNAR (3, 1, 2)_{12} hybrid model, and goodness of fit was predicted. The values for ME, RMSE, MAE, MPE, MAPE, and MASE for the hybrid model were
16.224, 5260.359, 3910.648, -0.077, 2.080, and 0.140 respectively, which is lower than the single SARIMA model having the values for the ME, RMSE, MAE, MPE, MAPE, and MASE to be 311.885, 6712.889, 4863.659, 0.033, 2.644, and 0.175 respectively given in Table 3. Finally, the monthly incidence of TB was predicted for 2020 using SARIMA and Hybrid models and compared with the actual notified TB incidences as given in Table 4. All the accuracy evaluation parameters of the hybrid model were found to be lower than the single SARIMA model and gave prediction closer to the actual TB incidences reported in 2020. However, forecast and prediction of the earlier model curves Fig. 6 & Fig. 7 show that monthly TB incidence of India is showing marginally increasing trend having the seasonal effect of notified TB incidences. A seasonal pattern showing the higher TB incidence in March, April, and May was also observed. We found that yearly notified TB incidence in India shows a marginally increasing trend.

**Table 3**

Comparisons of Predictive Performance Measures Among scale-dependent errors on both models

<table>
<thead>
<tr>
<th>Models</th>
<th>ME</th>
<th>RMSE</th>
<th>MAE</th>
<th>MPE</th>
<th>MAPE</th>
<th>MASE</th>
</tr>
</thead>
<tbody>
<tr>
<td>SARIMA</td>
<td>311.885</td>
<td>6712.889</td>
<td>4863.659</td>
<td>0.033</td>
<td>2.644</td>
<td>0.175</td>
</tr>
<tr>
<td>SARIMA-NNAR</td>
<td>16.224</td>
<td>5260.359</td>
<td>3910.648</td>
<td>-0.077</td>
<td>2.080</td>
<td>0.140</td>
</tr>
</tbody>
</table>

**Table 4**

Comparison between the reported notified TB incidence and forecast of TB incidence cases for 2020

<table>
<thead>
<tr>
<th>Time</th>
<th>Reported TB Cases</th>
<th>SARIMA</th>
<th>SARIMA-NNAR</th>
</tr>
</thead>
<tbody>
<tr>
<td>January 2020</td>
<td>197092</td>
<td>205656.7</td>
<td>216650.8</td>
</tr>
<tr>
<td>February 2020</td>
<td>213898</td>
<td>208842.5</td>
<td>211051.6</td>
</tr>
<tr>
<td>March 2020</td>
<td>169335</td>
<td>233244.1</td>
<td>226817.5</td>
</tr>
<tr>
<td>April 2020</td>
<td>83705*</td>
<td>248195.1</td>
<td>221273.8</td>
</tr>
<tr>
<td>May 2020</td>
<td>120825</td>
<td>258247.4</td>
<td>224473.4</td>
</tr>
<tr>
<td>June 2020</td>
<td>157368</td>
<td>241365.6</td>
<td>209595.4</td>
</tr>
<tr>
<td>July 2020</td>
<td>140941</td>
<td>239259.0</td>
<td>217260.0</td>
</tr>
<tr>
<td>August 2020</td>
<td>121920</td>
<td>222112.5</td>
<td>198533.3</td>
</tr>
<tr>
<td>September 2020</td>
<td>140891</td>
<td>223292.4</td>
<td>210521.8</td>
</tr>
<tr>
<td>October 2020</td>
<td>150639</td>
<td>228686.0</td>
<td>198387.6</td>
</tr>
<tr>
<td>November 2020</td>
<td>141650</td>
<td>216776.7</td>
<td>213780.0</td>
</tr>
<tr>
<td>December 2020</td>
<td>176641</td>
<td>217428.1</td>
<td>196691.8</td>
</tr>
</tbody>
</table>
*Due to COVID-19 all the OPD were not functioning properly therefore notified TB data becomes decline.

4 Discussion

TB is as old as the history of mankind. Since the time immemorial, TB is one of the significant killers globally [27]. Understanding the epidemiology of TB might enhance our knowledge and unravel the mystery behind its origin and spread. This will equip the stakeholders with better strategies to manage and control the key health infrastructures for effective control.

In this regard, we proposed the SARIMA model and the SARIMA-NNAR hybrid model to predict TB incidences in India. SARIMA (1, 0, 0)(1, 1, 0)_{12} and NNAR (3, 1, 2)_{12} models provide the best fit for time series forecasting of notified TB incidences from 2017 to 2020 in India. Both the models could simulate the TB time series data well, however, the hybrid model which incorporates both the linear and non-linear components performed better than the SARIMA model. There is a reason to believe that the hybrid models containing more data characteristics than the non-hybrid models do better forecasting[8]. The current study has presented the trend of notified TB cases in recent years in India using the SARIMA-NNAR hybrid model, which clearly shows seasonal variations in TB incidence that show periodicity in India. These incidents demonstrated a crest in the months from April to June and trough in the months from October to December. Similar, findings were reported from other workers as well [13, 28]. This means that the healthcare measures adopted for the control and the management of TB need to be intensified during the period from October to December. The finding of this study has revealed that the marginally increasing trend of TB notification is 12.05% from the years 2017 to 2019. Among these years, the higher peaks were observed in March, April, and May. According to Mathematical Models, the predicted outcomes indicated that the notified TB incidence in India will slightly increase in the future. A similar study was conducted in China applying a hybrid SARIMA (0, 1, 1) (0, 1, 1)_{12} GRNN model predicting the effects of TB notification based on the seasonality and trend[7]. A Chinese group has studied the TB notification database using SARIMA (1, 0, 0) (1, 0, 1)_{12} model and have proposed a hybrid model for the better assessment of seasonality and trend forecasting of TB notification rate when compared to SARIMA [29]. ARIMA model has found widespread and successful application in disease outbreak modelling. This has further led to the SARIMA and various other hybrid models.

Seasonality analysis reveals that the TB transmission appears to be higher during the winter months. This could be attributed to the winter overcrowding, low airflow, increased humidity, and reduced winter sunlight (antibacterial UV rays) coupled with poor ventilation. An additional contributing factor may be the seasonal variations in the immune function and the pattern of healthcare-seeking behaviour influenced by the intense cold season that forces people to stay indoors [28–30]. Some studies have also indicated a positive association between ambient air pollution and the risk of TB as the concentration and the retention of air pollutants such as particulate matter and other poisonous gases vary with the ambient temperature [30–33].
A similar study conducted in India revealed a high incidence of TB notification in March, April, and May [10, 19]. A meta-analysis of TB has been conducted based on a case-control study among healthy and infected individuals showing higher serum vitamin D in healthy individuals as compared to the serum vitamin D levels in the infected individuals [34].

Additionally, individuals having the co-infection of HIV-AIDS have a higher possibility of contracting TB due to compromised immunity. South Africa is one of the highest HIV/AIDS-affected countries, which coupled with lower socio-economic status leads to higher TB incidences [8]. A positive association between ambient concentrations of CO and NO$_2$ and risk of pulmonary TB among residents of northern California was established indicating the role of air pollution in causing TB [35].

The major strength of the study is the incorporation of nationwide data of a big country like India with its huge geo-climatic and socioeconomic diversity known to have an impact on the notified cases of TB. Current study is unarguably the largest spatio-temporal epidemiological study of notified TB cases in India. However, there are several limitations to this study as well, especially the non-availability of data from those individuals and populations that have no access to the healthcare professionals (vast swathes of tribal populations in India) and hence the data that we have used is mostly from the public and private sector, which may not reflect the actual data. Using notification date and not the date of diagnosis may also alter the seasonality pattern observed in the present study. Lack of climatic data, data on population migration, and other geo-climatic and, demographic data that is directly associated with the population were not included in this mathematical model. India is also one of the highest multidrug-resistant TB (MDR-TB) burden nations in the world. However, data related to the MDR/XDR-TB were not available to incorporate in the model. The non-availability of data on the huge geo-climatic diversity in India and its impact on TB incidence restricts its measurement through its incorporation in the model. TB is often referred to as a poor man’s disease. However, the lack of data on socio-economic indices and their impact on TB cannot be analyzed and modelled. Finally, both models have utilized data from 2017-19 and verified against only one year of data of TB prevalence. Hence, the outcomes need to be interpreted cautiously and should be re-examined with additional time-series data using a strong mathematical model.

## 5 Conclusions

This study is based on time-series data of notified TB cases from 2017–2019 in India. We have proposed the hybrid SARIMA-NNAR (3, 1, 2)$_{12}$ model based on the SARIMA Model. The results of the combined hybrid model are more efficient and effective compared to the single-generating SARIMA model. All the parameters were used to analyze the epidemic model. Since the model indicates that the result of notified TB prevalence in India will not decline in the forthcoming years, therefore there is an essential need to implement public health strategies for the prevention and control of TB transmission in India. Overall, this model could be applied successfully after the inclusion and adjustment of relevant parameters to forecast and predict the outbreak of diseases. Limitations inherent in the non-hybrid forecasting models can be compensated by developing hybrid models. The major takeaway from the study is that the health
care providers need to adopt this model as an early warning system that will allow the alerted and timely response in terms of enhanced surveillance and mobilization of healthcare resources.

**Declarations**

**Authorship contribution statement**

**Baikunth Kumar Yadav**: Methodology, Software, Data curation, Writing original draft. **Sunil Kumar Srivastava**: Methodology, Investigation, review and editing. **Pranveer Singh**: Conceptualization, Methodology, Supervision, Writing, review and editing.

**Acknowledgment**

UGC fellowship to BKY is gratefully acknowledged.

**Declaration of competing interest**

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

**Ethical Statement** The submitted work should be original and has not been published elsewhere in any form or language to our knowledge.

**Consent Statement** Not applicable.

**Data Availability** The Data for this study is owned by publicly available at the Central Tuberculosis Division (CTB), Government of India through data repository https://tbcindia.gov.in for notified monthly TB incidence from 2017-2019 were downloaded as excel sheet for Time Series analysis.

**References**

2. WHO (2020) Global Tuberculosis Report
5. TBFACTS.ORG. https://tbfacts.org/


11. Bras AL, Gomes D, Filipe PA, ... (2014) Trends, seasonality and forecasts of pulmonary tuberculosis in Portugal. ... journal of tuberculosis ...


15. Naranbat N, Nymadawa P, Schopfer K, ... (2009) Seasonality of tuberculosis in an Eastern-Asian country with an extreme continental climate. European Respiratory ...


23. Zhu JM, Tang LH, Zhou SS, Huang F (2007) Study on the feasibility for ARIMA model application to predict malaria incidence in an unstable malaria area. ... ji sheng chong xue yu ji ...


32. You S, Tong YW, Neoh KG, et al (2016) On the association between outdoor PM2.5 concentration and the seasonality of tuberculosis for Beijing and Hong Kong. Environmental Pollution


35. Geneé S. Smith, 1 Stephen K. Van Den Eeden, 2 Cynthia Garcia, 3 Jun Shan, 2 Roger Baxter, 2 Amy H. Herring, 4, 5 David B. Richardson, 1 Annelies Van Rie, 1 Michael Emch, 1, 5 6 and Marilie D. Gammon1 (2016) Air Pollution and Pulmonary Tuberculosis: A Nested Case–Control Study among Members of a Northern California Health Plan. Environ Health Perspect. https://doi.org/http://dx.doi.org/10.1289/ehp.1408166.

Figures
Figure 1

Monthly reported cases of Notified Tuberculosis prevalence data from 2017 to 2019.
Figure 2

Additive decomposition of monthly time series cases of Tuberculosis prevalence data.
Figure 3

Showing the Seasonally adjusted Tuberculosis incidence.
Figure 4

Time-Series (Autocorrelation Function) ACF & (Partial Autocorrelation Function) PACF plot.
Figure 5

Annually reported notified TB cases per 100000 in Indian states & UTs.

Figure 6

Forecast from SARIMA model to applied on Notified Tuberculosis case prevalence.
Figure 7

Forecast from SARIMA-NNAR to applied on Notified Tuberculosis cases prevalence.