Forecasting incidence of tuberculosis cases in Brazil based on various univariate time-series models

Rhuan Carlos Martins Ribeiro [a], Thaynara Araújo Quadros [b], John Jairo Saldarriaga Ausique [b], Otavio Andre Chase [a], Pedro Silvestre da Silva Campos [a], Paulo Cerqueira dos Santos Júnior [a], José Felipe de Almeida [a], Glauber Tadaiesky Marques [a]

[a] Federal Rural University of Amazonia (UFRA), Cyberspatial Institute (ICIBE), Brazil
[b] Federal University of Pará (UFPA), Science and Technology Park of Guamá (PCT-Guamá), Center for the Valuation of Bioactive Compounds in the Amazonia (CVCBA), Brazil

Abstract

Tuberculosis (TB) remains the world's deadliest infectious disease and is a serious public health problem. Control for this disease still presents several difficulties, requiring strategies for the execution of immediate combat and intervention actions. Given that changes through the decision-making process are guided by current information and future prognoses, it is critical that a country's public health managers rely on accurate predictions that can detect the evolving incidence phenomena. of TB. Thus, this study aims to analyze the accuracy of predictions of three univariate models based on time series of diagnosed TB cases in Brazil, from January 2001 to June 2018, in order to establish which model presents better performance. For the second half of 2018. From this, data were collected from the Department of Informatics of the Unified Health System (DATASUS), which were submitted to the methods of Simple Exponential Smoothing (SES), Holt-Winters Exponential Smoothing (HWES) and the Integrated Autoregressive Moving Average (ARIMA) model. In the performance analysis and model selection, six criteria based on precision errors were established: Mean Square Error (MSE), Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percent Error (MAPE) and Theil's U statistic (U1 and U2). According to the results obtained, the HWES (0.2, 0.1, 0.1) presented a high performance in relation to the error metrics, consisting of the best model compared to the other two methodologies compared here.

Keywords: Forecasting, Holt-Winters, Tuberculosis, Time Series, Univariate Models.

1. Introduction

Tuberculosis (TB) is a contagious, often fatal infectious disease caused by the Mycobacterium tuberculosis bacillus, which targets the lungs (pulmonary TB) and can sporadically settle in other organs (extrapulmonary TB). It is transmitted through air contaminated with the antigen and easily contracted by inhalation. According to Drain, et al., 2018, the limited understanding of the clinical and pathogenic spectrum of M. tuberculosis infection is one of the factors that make TB the leading cause of infectious agent mortality globally. In this sense, the World Health Organization's Global Tuberculosis Report (WHO, 2018) presented data for the year 2017, approximately 10 million people developed the disease and 1.3

million died.

Recent research indicates that about 23% of the world's population are at risk of acquiring the infection in its latent form and developing the disease over a lifetime. However, the progress of research for the control of M. tuberculosis is still low in most cases, because the bacillus is multidrug resistant (MDR-TB) or extensively resistant (XDR-TB) to the drugs used (BORISOV, et al., 2017). A study by Stein et al. (2018) show that Human Immunodeficiency Virus (HIV) infection and low body mass index contribute to a faster evolution of TB. Similarly, the overall treatment effectiveness rate for MDR-TB is <55% (FALZON, et al., 2017) and for XDR-TB, the mortality rate is up to 80% because it is resistant, at least four of first and second-line drugs (PIETERSEN, et al., 2014). Alternatively, Merker et al. (2015) consider that the cause of this resistance may be related to inadequate treatment of the disease, resulting as an evolutionary pressure to the etiological agent.

In 2017, the first global meeting at the end of TB was held, with the aim of devising combat strategies, focused mainly on early diagnosis and effective treatment, as these measures can prevent millions of new cases. This conference subsequently resulted in the Moscow Declaration, which is a global commitment to take urgent action to combat TB, including enough financial mobilization in research aimed at prevention, diagnosis and treatment. At the World Health Assembly in May 2018, WHO member states pledged to accelerate action against TB based on the Moscow Declaration (WHO, 2018). In Brazil, the Ministry of Health has developed a national plan to suppress tuberculosis as a public health problem. The document was published in 2017 and its main objective is the early diagnosis and the guarantee of continuous treatment without interruption in the first six months.

In order to provide support for the development of actions in combat and its prevention, related to the high mortality rate associated with TB, there is an urgent need to predict and monitor the epidemiology of this disease. Aiming at actions, it is necessary the epidemiological approach through time series, which consists of the ordered observations of a variable, following a regular and successive time interval (ATKINSON, et al., 2015; PAULINO, et al., 2017), that allow for elaboration of predictive models to monitor the epidemiological evolution of TB, and in the study of these temporal alterations different data series models have been used (NASEHI, BAHRAMPOUR, et al., 2014; KILICMAN and ROSLAN, 2009; ZHENG, ZHANG , et al., 2015; CAO, WANG, et al., 2013; DUBE, 2015). Moreover, the determination of the best model for forecasting will depend directly on the nature and behavior of the data, however, since the disease presents different forms of temporal distribution around the world, different models are able to perform the prognosis with relative precision quality. For example, Abdullah et al., 2012, showed that the double exponential smoothing technique is the best model for predicting TB in the state of Kelantan - Malaysia. However, Zheng, et. al., 2015, modeled the TB forecast in the Xinjiang - China region using the autoregressive integrated moving average (ARIMA) method as the basis.

This text describes what we consider to be the best univariate model against three proposed methods, based on the characteristics of TB diagnosis in Brazil. Data were collected on the platform of the informatics department at Brazilian unified health system (DATASUS) from January 2001 to June 2018. After tabulating the information, an application has been used by three prediction models, consisting of: Simple exponential smoothing (SES), Holt-Winters exponential smoothing (HWES) and the ARIMA method, we also present an analysis of the significant errors contained in these models and the importance of obtaining

minimal errors. Finally, the data adjusted for the first half of 2019 were predicted by the model observing the highest prediction capacity.

2. Methodology

2.1 Date Collection

Data for the monthly incidence of TB were obtained from DATASUS from January 2001 to June 2018, totaling 210 markings. In approximately 25 years of operation, DATASUS has developed over 200 systems that directly assist the Ministry of Health in the process of building and strengthening the Brazilian Unified Health System (SUS). Its data storage structure can store the health information for the entire Brazilian population due to its connection with all 27 state health secretariats and other SUS related centers, agencies and institutions. All cases of TB were verified by clinical or laboratory diagnosis. So, it is believed that the TB case reporting data from this department is relatively reliable for modeling and obtaining the results contained in this research. Table 1 presents the historical data obtained and used in this work.

Table 1. Cases diagnosed with TB in Brazil from January 2001 to June 2018.

Year	Month											
	Jan.	Feb	Mar.	Apr.	May.	Jun.	Jul.	Aug.	Sep.	Oct	Nov.	Dec.
2001	8.088	6.542	8.095	7.305	7.656	6.805	6.987	8.065	6.697	7.524	6.874	6.627
2002	8.013	7.346	7.961	8.771	7.784	6.713	7.746	8.330	7.654	8.200	7.532	6.809
2003	8.115	7.985	7.474	7.975	8.024	7.082	7.949	7.585	8.136	8.521	7.624	7.303
2004	7.574	6.743	8.574	8.089	7.763	7.314	7.825	8.232	7.901	8.016	7.753	7.196
2005	7.432	6.748	8.441	7.958	8.019	7.781	7.282	8.391	7.690	7.186	7.587	7.541
2006	7.466	6.767	8.256	6.901	7.699	7.094	7.220	7.835	6.873	7.089	6.751	6.209
2007	7.436	6.294	8.214	7.373	7.481	6.714	7.283	7.669	6.801	7.550	6.786	6.178
2008	7.459	6.729	7.359	7.717	6.992	7.001	7.777	7.912	7.720	7.643	6.863	6.650
2009	7.180	6.507	8.231	7.584	7.245	6.799	7.513	7.450	7.478	7.359	7.088	6.793
2010	6.987	6.471	8.373	7.113	7.039	6.701	7.178	7.463	7.216	7.273	7.096	7.236
2011	7.173	7.335	7.548	7.619	7.765	7.008	7.138	8.062	7.427	7.052	7.439	6.994
2012	7.435	6.830	7.936	6.981	7.649	6.861	7.317	8.060	6.785	7.679	7.010	6.362
2013	7.502	6.209	7.102	7.696	6.996	6.953	7.379	7.931	7.377	7.789	6.939	6.335
2014	7.641	7.027	6.743	7.256	7.169	6.320	7.511	7.118	7.502	7.465	6.903	6.469
2015	7.118	6.161	7.840	6.877	6.880	6.916	7.527	7.468	7.197	7.344	7.280	6.847
2016	7.037	6.772	8.054	7.363	7.320	7.548	7.029	7.750	7.101	6.663	7.187	6.993
2017	7.460	6.799	8.693	6.776	8.094	7.408	7.220	8.099	7.440	7.793	7.445	7.014
2018	7.815	6.774	7.887	8.122	7.848	7.625						

Source: DATASUS, 2019.

2.2 Simple Exponential Smoothing Model (SES)

The SES Method is a formalization of Machine Learning based on similarity and involves smoothing out

random fluctuations of time series data. The use of this technique is significant to predict data without trend or seasonal pattern, i.e.: When the data pattern found is close to horizontality (JERE, KASENSE and CHILYABANYAMA, 2017). The mathematical formulation SES model for time series data ($\bar{L}_t$), is shown below:

$\bar{L}_t = \alpha_1 L_t + (1-\alpha_1)\bar{L}_{t-1}$, $0 < \alpha < 1$ and $t > 0$.

Where α is the smoothing constant, L_t is the raw data of the series and $\bar{L}_t$ is the smoothed or output data. For the h-step-ahead prediction equation, we have:

$$\hat{L}_{t+h} = \bar{L}_t.$$

And h assumes values: h = 1,2,3.. (TULARAM & SAEED, 2016).

2.3 Autoregressive Integrated Moving Average Model (ARIMA)

The stochastic models popularized by Box-Jenkins in the early 1970s, known as ARIMA, involve a transformation of data to stabilize variance. The "I" in ARIMA indicates that the dataset is transformed through differentiation culminating in the stationarity of time series that, after the completion of modeling, will be the results integrated in order to make predictions and final estimates. (DRITSAKIS & KLAZOGLOU, 2019).

The function that represents this model is called ARIMA (p, d, q), constituted by the order of the autoregressive model (AR) (p), order of differentiation (d) and the moving average structure order (MA) (q). Thus, we have as expressions: AR, MA e ARMA:

$$\text{AR model: } \hat{L}_t = \theta_1 L_{t-1} + \theta_2 L_{t-2} + \cdots + \theta_p L_{t-p} + \mathcal{E}_t = \sum_{i=1}^{p} \theta_i L_{t-i} + \mathcal{E}_t$$

$$\text{MA model: } \hat{L}_t = \phi_1 \mathcal{E}_{t-1} + \phi_2 \mathcal{E}_{t-2} + \cdots + \phi_q \mathcal{E}_{t-q} = \sum_{i=1}^{q} \phi_i \mathcal{E}_{t-i} \ ,$$

and

$$\text{ARMA model: } \hat{L}_t = \sum_{i=1}^{p} \theta_i L_{t-i} + \mathcal{E}_t + \sum_{i=1}^{q} \phi_i \mathcal{E}_{t-i}.$$

Where θ_t is the autoregression parameter at time t, $\mathcal{E}_t$ is the error term at time *t*, and ϕ_t is the moving average parameter at time t.

2.4 Holt-Winters Exponential Smoothing Model (HWES)

The model HWES considers three smoothing equations, in which one comprises the level and the tendency and the other to the seasonality (RIBEIRO, et al., 2019). When an increase in the seasonal amplitude is required, the series presents a difference between the highest and the lowest demand point in the cycles grows with time, and the multiplicative model becomes adequate. Thus, when the seasonal amplitude is constant, it means that the largest and smallest points in the cycles are independent of the temporal variation, and the model to be used is the additive (HOLT, 2004; WINTERS, 1960).

Table 2. Comparative equations for the multiplicative and additive Holt-Winters models.

	Additive Holt-Winters	Multiplicative Holt-Winters
Level	$\bar{L}_t = \alpha(Y_t - \hat{S}_{t-s}) + (1-\alpha)(\bar{L}_{t-1} + \hat{B}_{t-1})$	$\bar{L}_t = \alpha\frac{Y_t}{\hat{S}_{t-s}} + (1-\alpha)(\bar{L}_{t-1} + \hat{B}_{t-1})$
Trend	$\hat{B}_t = \beta(\bar{L}_t - \bar{L}_{t-1}) + (1-\beta)\hat{B}_{t-1}$	$\hat{B}_t = \beta(\bar{L}_t - \bar{L}_{t-1}) + (1-\beta)\hat{B}_{t-1}$
Seasonality	$\hat{S}_t = \gamma(Y_t - L_t) + (1-\gamma)\hat{S}_{t-s}$	$\hat{S}_t = \gamma\left(\frac{Y_t}{L_t}\right) + (1-\gamma)\hat{S}_{t-s}$
Forecast	$F_{t+m} = (\bar{L}_t + \hat{B}_t m) + \hat{S}_{t-s+m}$	$F_{t+m} = (\bar{L}_t + \hat{B}_t m)\hat{S}_{t-s+m}$

In Table 2, S is the seasonality length, $\bar{L}_t$ is the series level, $\hat{B}_t$ is the trend, $\hat{S}_t$ is the seasonal component, F_{t+m} is the forecast for period m, Y_t is the observed value and α, β and γ are exponential parameters of the level, trend and seasonality, respectively.

2.5 Model-Selection Criteria

Although a predictive technique may occasionally be appropriate for any database, univariate analysis of time series models, requires simultaneous evaluation to identify which provides one of the best results for error minimization. (TULARAM & SAEED, 2016; BILLAH, KING, SNYDER, & KOEHLER, 2006). Thus, we applied six metrics for model selection. A common assumption among many such selection criteria requires the following parameters to be considered the most appropriate: Mean Square Error (MSE), Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percent Error (MAPE) and Theil's U statistic. Two Theil U statistics named U1 and U2 were defined and, like all selection criteria, these metrics exhibit advantages and disadvantages as well as specific conditions for applicability. Table 3 shows the calculation of each criterion exposed.

Table 3. Model Evaluation Metrics.

Criteria	Formula	Criteria	Formula
MSE	$\frac{1}{n}\sum_{i=1}^{n}\mathcal{E}_i^2$	RMSE	$\sqrt{\text{MSE}}$
MAE	$\frac{1}{n}\sum_{i=1}^{n}\lvert\mathcal{E}_i\rvert$	MAPE	$\frac{1}{n}\sum_{i=1}^{n}\left(\left\lvert\frac{\mathcal{E}_i}{x_i}\right\rvert\right) * 100$
U_1	$\frac{\text{RMSE}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}x_i^2} + \sqrt{\frac{1}{T}\sum_{i=1}^{n}\hat{x}_i^2}}$	U_2	$\frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n-1}\left(\frac{\hat{x}_{i+1} - x_{i+1}}{x_i}\right)^2}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n-1}\left(\frac{x_{i+1} - x_i}{x_i}\right)^2}}$

In Table 3, $\mathcal{E}_i = x_i - \hat{x}_i$, where x_i represents the observed actual value at the moment i, $\hat{x}_i$ is the predicted value at the moment i and n is the observation number used in the calculation.

The first evaluation criterion is the mean squared error, or MSE, for the entire set of data. The MSE attaches more weight to significant errors and the main limitation of this approach is that it overstates substantial errors. This valuation criterion also provides limited information about an overestimation or underestimation of actual forecast value. The second evaluation criterion, or RMSE, preserves the units in the estimation variable and, to some extent, this approach is more sensitive to a large number of errors occurred. However, the ability to compare different time series is limited by this criterion. In contrast, the third criterion, the MAE, determines the magnitude of the error and is an important issue for a precise set of predictions, which defines how close the predictions are to actual results, but does not consider the current direction on predictions. Elsewhere, the fourth criterion, or MAPE, allows the comparison of distinct time series data without defining the relationship or percentage error. This last information is significant in instances where the measured variables are too large.

The fifth and sixth criteria, U1 and U2, are more difficult metrics to use than MAPE. The first, or U1, provides a range of values that varies from 0 to 1. The closer U1 is to 0, the more accurate the forecast will be. When confronted with alternative predictions, the model with the lowest U1 value is considered best and thus selected. On the other hand, U2 makes relative comparisons based on random walk models (NEWAZ, 2008) and prediction models (Naive model). The Naive model (TAHERI & MAMMADOV, 2013) can be described as the current prediction model applied based on an indiscriminate walking process, i.e. it assumes that demand in the next period will be equal to demand in the most recent period. When U2 stabilizes in the unit, the Naive method is considered equally useful for forecasting and U2> 1 indicates that the prediction model would work better than the Naive approach.

3. Results and Discussion

3.1 Preliminary data analysis

Data analysis was performed using Minitab®v.18 computer software and commercial Microsoft Excel®2016. Minitab has built-in functions which can determine the best model parameters spontaneously and the data series being the only necessary correspondent for this system. Figure 1 shows that the distribution pattern indicates that data assume a low trend over the years, but the seasonality is quite clear in the visual observation.

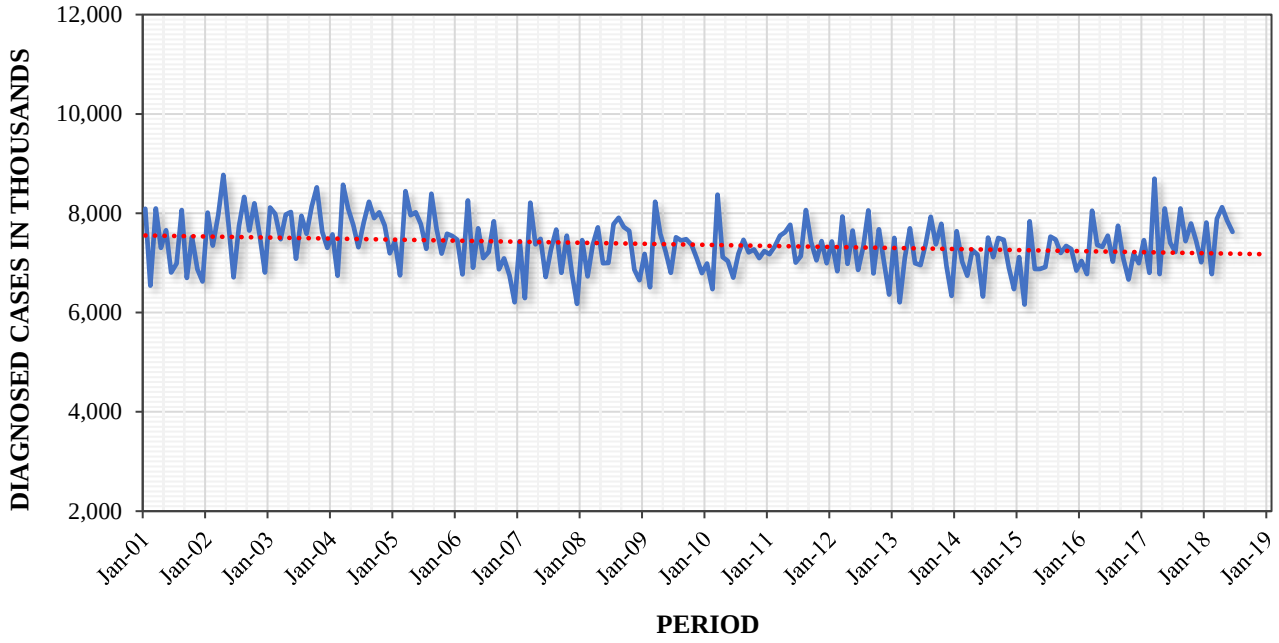

Figure 1. Primary data series from January 2001 to June 2018.

Regarding the cases analyzed, the highest annual percentage increase of diagnosed cases compared to the previous year was 6.41% between 2001 and 2002. This percentage of growth observed consists of 5,594 cases diagnosed in 2002 more than in 2001. However, between 2005 and 2006 a negative evolution was found, -6.40%, which means that in 2006 the number of people diagnosed with TB decreased, this percentage is equivalent to 5,896 fewer than the previous year, and this was the year with the greatest difference in case reduction compared to the years studied. The monthly average of diagnoses was 7,369, with a standard deviation of approximately 528.83. The month with the highest diagnosis rate was in April 2002 with 8,771, and the lowest was February 2015 with 6,161.

3.2 Modeling Results (SES, ARIMA e HWES)

The results from the application of the SES, ARIMA and HWES methods are shown in Figures 2, 5 and 6, respectively. These Figures contain only part of the original data series (from January 2015), also have the adjustments made by the models, the forecasts accompanied by the 95% confidence intervals and the real data. The results of the different models are substantial, regardless of which the metrics is used to qualify the accuracy of their predictions.

The SES model uses just one parameter α_1, therefore, the appropriate response value is determined by minimizing the error with respect to α_1. Results indicated that the value of $\alpha_1 = 0.07362$ is the best parameter measure for this prediction model. Therefore, the equation governing this first modeling scenario takes the following form: $\bar{L}_t = 0.07362\, L_t + 0.9264\, \bar{L}_{t-1}$.

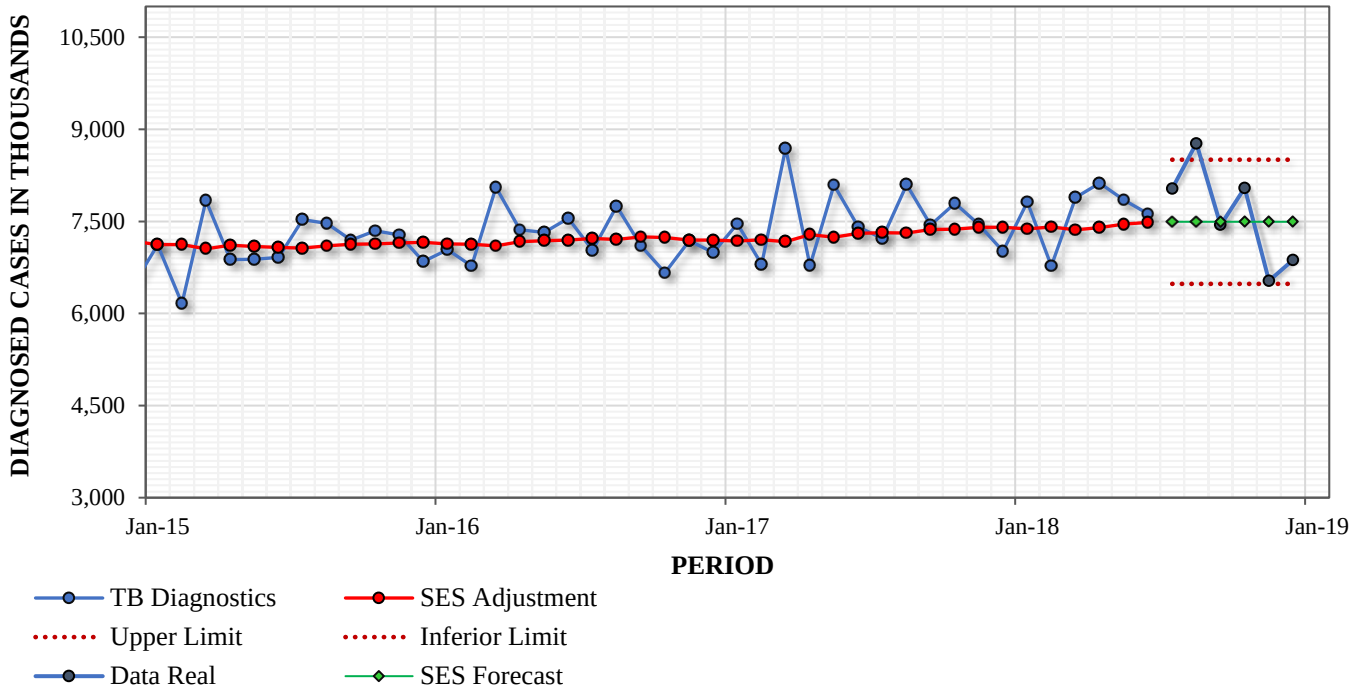

Figure 2. The results of the SES model forecast for TB cases diagnosed in Brazil from July 2018 to December 2018.

Figure 2 presents the result obtained using the SES model. It is observed that the forecast (Green Line) consistently does not follow the behavior for the actual data of the period studied (Blue Line). It is also noted that the real value for the month of August was outside the 95% CI, so that in this preliminary visual analysis the SES model does not satisfactorily predict the actual TB data.

In relation to the SES model, ARIMA includes an explicit statistical model for the irregular component of a time series, which allows non-zero autocorrelations in the irregular component, revealing the potential of this data processing for predictions. ARIMA consists of three main steps: the first focuses on model identification; the second on parameter estimation; and, lastly on diagnostic verification through predictions. However, it is considered that the initial phase of identification of the ARIMA model happens through the visual verification of the stationary behavior of the series. When the need for conversion from a non-stationary to a stationary time series is certified, the differentiation process consisting of an important part of the adaptation method of an ARIMA model is submitted (DOBRE & ALEXANDRU, 2008). In this sense, a time graph is plotted in Figure 3(a) where d=0 is equivalent to the series data without differentiation and in Figure 3(b) d=1 to the differentiated data. By displaying the Autocorrelation (ACF) and Partial Autocorrelation (PACF) graphs in Figure 4, it is noted that the TB data are stable at d=1, which is a need for series differentiation.

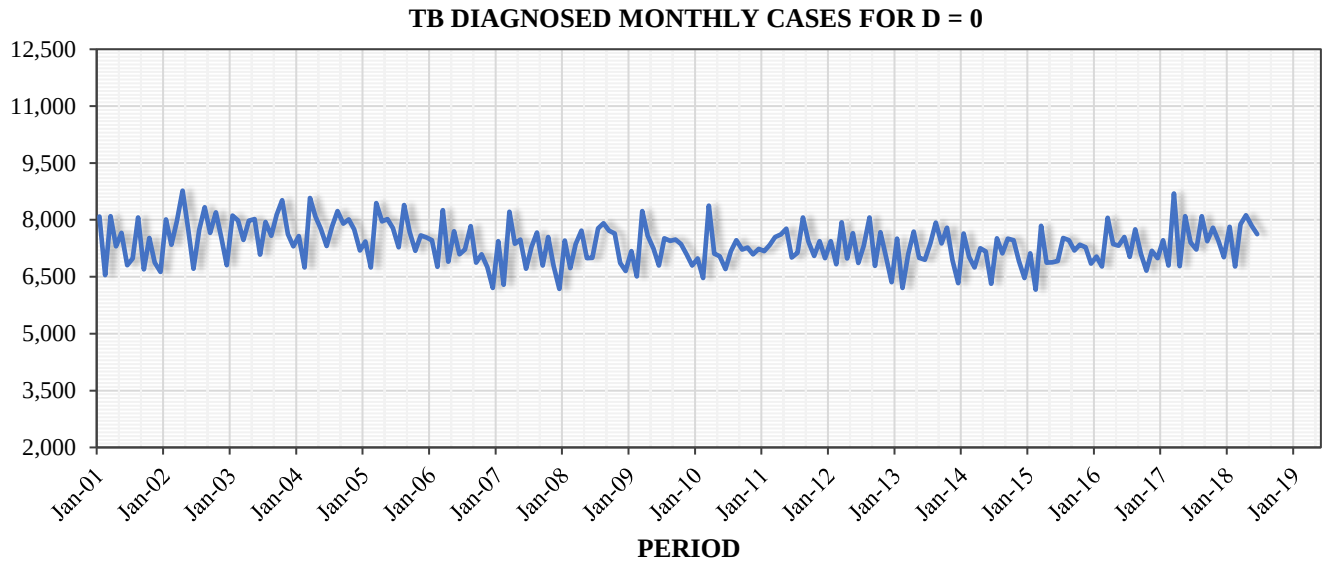

Figure 3 (a). TB diagnostic time data at d=0.

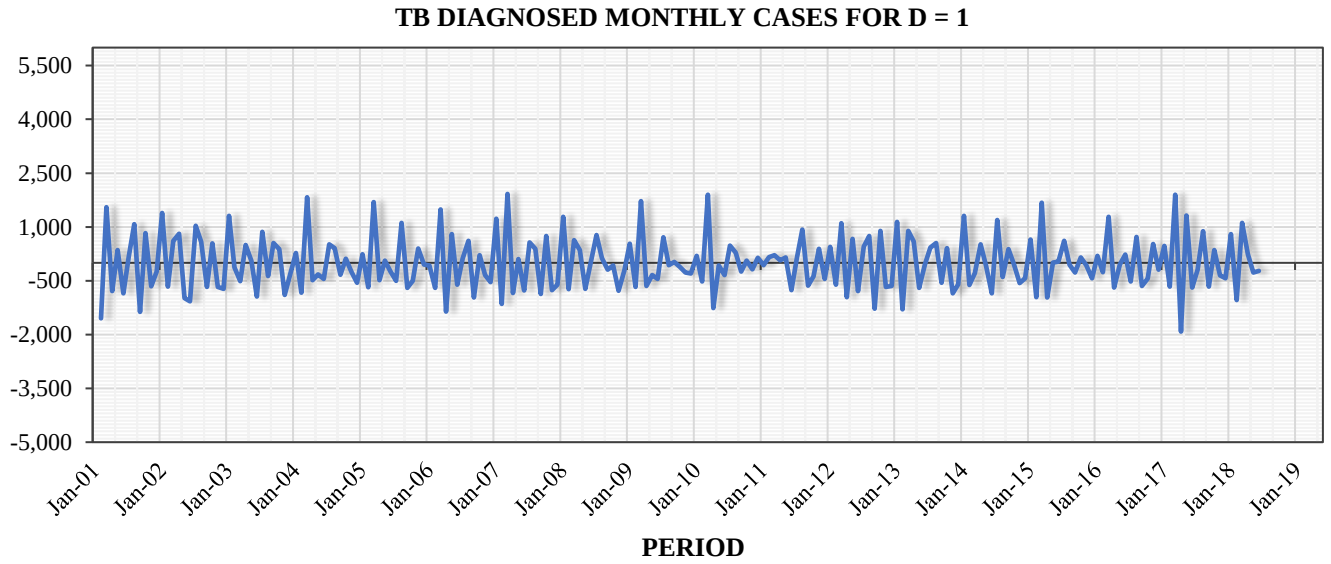

Figure 3(b). TB diagnostic time data at d=1.

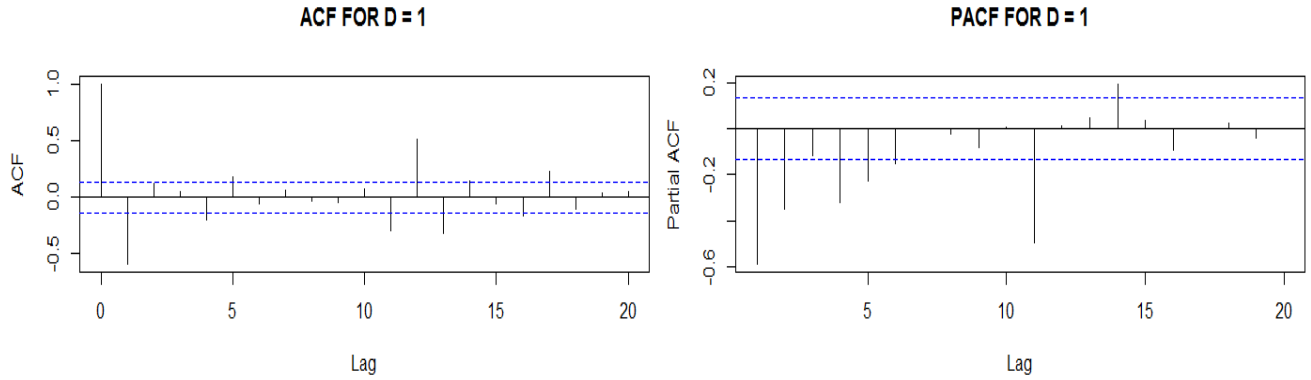

Figure 4. ACF and PACF graphs for d=1.

The time series of differences (Figure 3b) and the ACF and PACF graphs at d = 1 confirm that there is a

need to assign a stationary pattern to the mean and variance of the series by differentiation, thus an ARIMA model (p, 1, q) is probably the most appropriate, so proper verification is required to establish which AR and MA terms are appropriate. We see from the correlogram that the self-correlation in lag 1 (-0.591) exceeds the limits of significance and is negative, after this delay the autocorrelations decrease to zero, although perhaps too abruptly for an ARIMA model (0,1,1) is to be satisfactory. The partial correlogram shows that the partial autocorrelations in lags 4 (-0.319) and 5 (-0.228), are negative delays that exceed the limits of significance before the partial autocorrelations decrease to zero. These observations allow to visualize what would represent an approximation of a suitable model, in this case it would be an ARIMA (5,1,1). Moreover, it should be considered, at this point, that the ARIMA model with the smallest measurement error specific to the Akaike Information Criterion (AIC) be the best model for forecasting (SILUYELE & JERE, 2016). Thus, Table 4 shows the comparison of AIC errors between the proposed models.

Table 4. AIC statistical measures for selected ARIMA models.

Attempt Model	(4,1,5)	(0,1,5)	(1,1,1)	(5,1,1)	(1,1,5)	(0,1,2)	(0,1,3)	(1,1,2)
AIC	3171.86	3199.88	3204.01	3201.84	3201.1	3204.47	3205.85	3205.86
Attempt Model	(0,1,4)	(2,1,2)	(0,1,1)	(1,1,3)	(5,1,0)	(4,1,0)	(3,1,0)	(2,1,0)
AIC	3206.05	3207.48	3207.63	3208.43	3212.93	3223.87	3244.79	3245.56

Table 4 highlights the ARIMA model (4,1,5) which contains the lowest value for the (AIC). It is noteworthy the importance of this selection criterion, which considers both the graphical behavior of the autocorrelations and the error parameter common to the proposed model. The RStudio Software version (1.1.463) has been used as a statistical tool to obtain the main metrics and parameters of this model (Table 5). The coefficients and *p* values of the terms were used, as well as the equation that governs this second modeling scenario, highlighting (in bold) the corresponding high significance terms (*p*-value).

$$\hat{L}_t = 2{,}7 - 0{,}296\, L_{t-1} - 0{,}2954\, L_{t-2} - 0{,}7066\, L_{t-3} - 0{,}571\, L_{t-4} + 0{,}840\, \mathcal{E}_{t-1} - 0{,}351\, \mathcal{E}_{t-2} - 0{,}377\, \mathcal{E}_{t-3} + 0{,}361\, \mathcal{E}_{t-4} + 0{,}199\, \mathcal{E}_{t-5}$$

Table 5. ARIMA (4,1,5).

Variable	Coeficient	*p*-value
Constant	2,7	0,797
AR (1)	-0,296	0,011*
AR (2)	-0,2954	0,001*
AR (3)	-0,7066	0,000*
AR (4)	-0,571	0,000*

MA (1)	0,840	0,000*
MA (2)	-0,351	0,015*
MA (3)	-0,377	0,076
MA (4)	0,361	0,110
MA (5)	0,199	0,186
Sigma^ estimated	208354	-
log likelihood	-1575.93	-

*Note: *p*-value < 0.05.

Table 5 shows information on the significance of the parameters used in this model, it is observed that all AR terms are significant and the MA (1) and MA (2) also have an appreciated significance value in view of the *p*-values of these terms, found below 0.05%. Therefore, to model identification and parameter estimation, the model can then be checked by means of predictions (Figure 5).

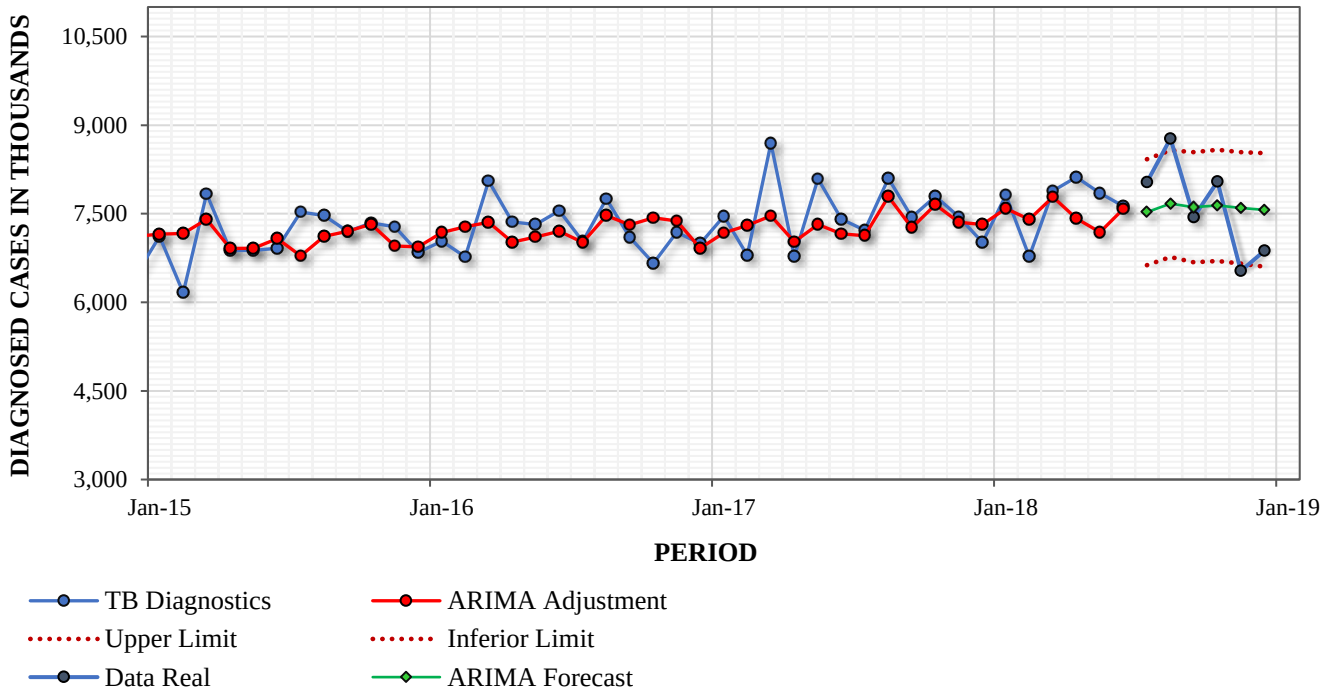

Figure 5. ARIMA (4,1,5) forecast results for TB cases diagnosed in Brazil from July 2018 to December 2018.

When compared with the ARIMA (4,1,5) prediction model, the HWES model performs better. Both additive and multiplicative HWES cases were evaluated and all parameters were stipulated and not estimated, which ranged from 0.4 to 0.1. The lower weights produce a smoother line and the larger ones a sharper line. Thus, using lower weights for noisy data, we obtain smoothed values that do not fluctuate along with noise (RIBEIRO, et al., 2019). In addition, a precision measure was chosen to compare the twenty-four HWES models obtained and find the one that best fits the TB data. Therefore, the measure

chosen was MSE and in Table 6 is showed all models and their MSE measurements.

Table 6. MSE Precision Parameters for Selecting the Best HWES Model.

HWES Additive Models						
Attempt Model	(0.2;0.1;0.1)	(0.2;0.2;0.1)	(0.3;0.1;0.1)	(0.2;0.1;0.2)	(0.3;0.2;0.1)	(0.2;0.2;0.2)
MSE	148,689	152,466	153,722	158,173	159,771	162,880
Attempt Model	(0.4;0.1;0.1)	(0.3;0.1;0.2)	(0.3;0.2;0.2)	(0.4;0.2;0.1)	(0.4;0.1;0.2)	(0.4;0.2;0.2)
MSE	163,068	163,432	170,540	170,741	172,593	181,247
HWES Multiplicative Models						
Attempt Model	(0.2;0.1;0.1)	(0.2;0.2;0.1)	(0.3;0.1;0.1)	(0.2;0.1;0.2)	(0.3;0.2;0.1)	(0.2;0.2;0.2)
MSE	150,141	154,356	155,511	159,816	161,613	164,824
Attempt Model	(0.3;0.1;0.2)	(0.4;0.1;0.1)	(0.3;0.2;0.2)	(0.4;0.2;0.1)	(0.4;0.1;0.2)	(0.4;0.2;0.2)
MSE	165,352	165,382	172,528	173,184	175,091	183,902

In Table 6, is highlighted (in bold) the HWES model that presented the lowest MSE. According to this selection criterion, the case of HW chosen is the additive and the parameters for predicting TB diagnoses should be α=0.2, β=0.1 and γ=0.1. This α value is relatively low, indicating that the model produces a current moment level estimate based on recent observations and some observations in the more distant past. The β value indicates that the slope estimate is constantly updated during the series and differs from its initial value in an intuitive form, as the level varies greatly over the time series, this value suggests that the slope β of the trend has also changed constantly. Furthermore, the low value of γ allows estimating the seasonal component at the moment is based on relatively recent observations and on more distant ones. Therefore, the equations constituting the model and governing this third modeling scenario culminate in:

$$\bar{L}_t = 0.2\left(Y_t - \hat{S}_{t-s}\right) + 0.8\left(\bar{L}_{t-1} + \hat{B}_{t-1}\right),$$

$$\hat{B}_t = 0.1\left(\bar{L}_t - \bar{L}_{t-1}\right) + 0.9\,\hat{B}_{t-1},$$

and

$$\hat{S}_t = 0.1\left(Y_t - L_t\right) + 0.9\,\hat{S}_{t-s}.$$

Consequently, the model described above was used to generate the predictions contained in Figure 6. It is observed that HWES is superior to SES and ARIMA (4,1,5) at a visual analysis of the prediction and adjustment results. The projected forecast line tends to follow the actual data and even satisfactorily forecasts the value for October 2018.

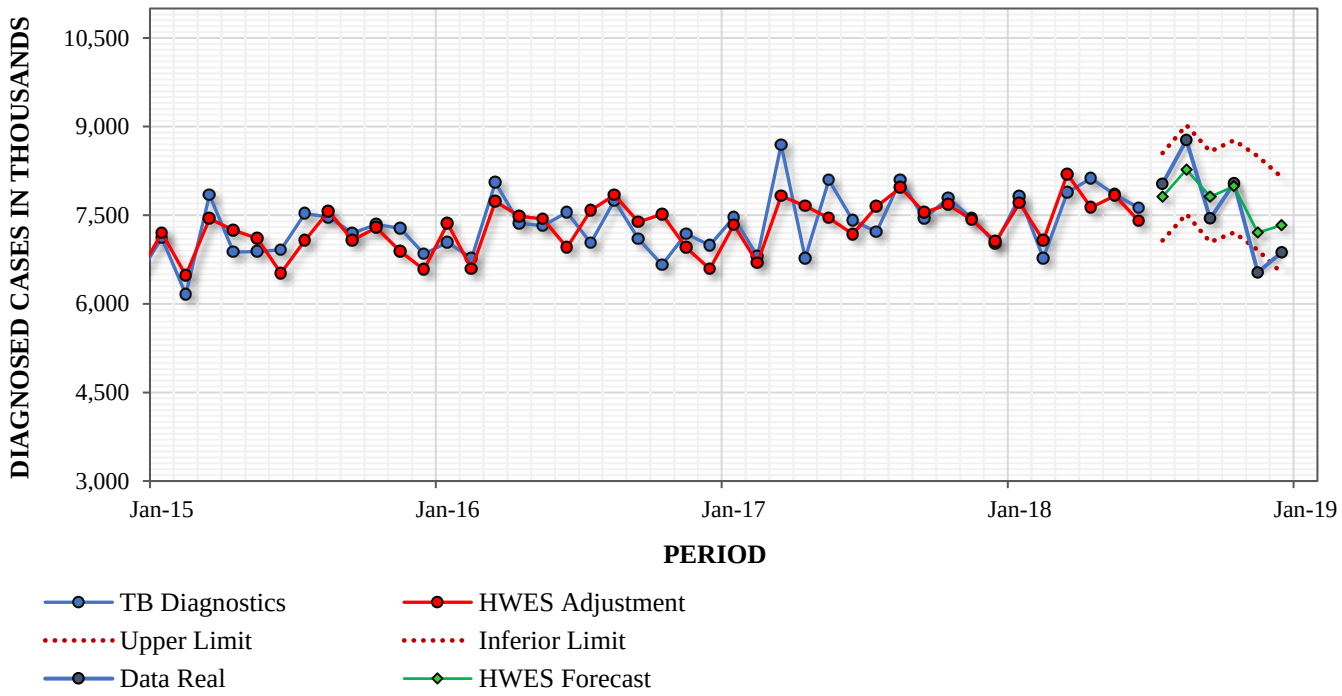

Figure 6. HWES model prediction results for TB cases diagnosed in Brazil from July 2018 to December 2018.

Although exponential smoothing methods are useful for forecasting, without considering assumptions about correlations between successive time series values, HWES showed better performance against the other two methodologies. This shows that only in some cases is it possible to make a better predictive model taking into account the correlation of data. This difference becomes clearer when is observed in Table 7 which contains the six model precision measurements and the HWES (0.2, 0.1, 0.1) is superior to the others.

Table 7. Selection criteria of the best model among the three used in this study (SES, ARIMA and HWES).

Measures of accuracy	SES	ARIMA	HWES
MSE	262,131	198,381	148,689
RMSE	512	445.4	385.6
MAE	413	352.4	302
MAPE	6	4.84	4
U_1	0.013	0.012	0.010
U_2	0.983	0.998	1.016
Model Ranking	3	2	1

Table 7 summarizes the three models for the precision measurements each presented. The results show that HWES (0.2, 0.1, 0.1) obtained the best performance against this precision evaluation. The ARIMA method is theoretically more robust from a mathematical point of view, so that a model meets important requirements regarding errors such as non-correlation and normal distribution with zero mean and constant

variance. However, although exponential smoothing methods do not use assumptions about correlations between consecutive time series values, in some cases it is possible to find a more predictive exponential smoothing model that best fits the original data, as in the present case.

In general, forecasts play an indispensable role in the entire policymaking process of public policy makers. Two perspectives are visualized when changes are achieved in the face of a decision-making process: current events and predictions of future events (TULARAM & SAEED, 2016). With this bias, a country's public health decision-makers rely on accurate forecasts to push their policies so that the end results are different from those predicted. Forecasts can also play a vital role in the development and expansion of control and intervention programs, as well as in the allocation of optimal material for such mechanisms (AKHTAR & MOHAMMAD, 2008).

The diagnosis of the temporal distribution of TB cases around the world is quite diverse, in some countries and regions TB that presents high seasonality of occurrence (MOHAMMED, AHMED, AL MOUSAWI, & AZEEZ, 2018; WUBULI, et al., 2017; KHALIQ, BATOOL, & CHAUDHRY, 2015). This results in the development of models that are capable of assessing peaks of occurrence and seasonal trends, such as the integrated seasonal moving average autoregressive (SARIMA), neural network autoregression (SARIMA-NNAR), heteroscedasticity conditioned autoregressive (ARCH) models, among others (AZEEZ, OBAROMI, ODEYEMI, NDEGE, & MUNTABAYI, 2016; ZHENG, ZHANG, ZHANG, WANG, & ZHENG, Forecast Model Analysis for the Morbidity of Tuberculosis in Xinjiang, China, 2015; MAO, ZHANG, YAN, & CHENG, 2018). Thus, we have a model for each kind of data behavior. A similar study was conducted by Nothabo Dube in Zimbabwe in 2015, a country still considered by the WHO with a high incidence rate of TB. In this case study, we analyzed the ARIMA, ARIMA-ARCH and HW models, and the performance of the ARIMA (2,1,1) model proved to be far more superior to others, becoming an ideal model for predicting the annual incidence of TB in Zimbabwe.

4. Final Considerations

In this work, the observed results show that, despite efforts by the Brazilian ministry of health to fight the disease, there will be no apparent improvement in the high incidence of diagnosed TB cases in Brazil, for the near future. TB is still a major global public health problem and is one of the most fatal infectious diseases in the world. In Brazil, the increase in patients coinfected with TB and HIV, as well as the emergence of drug-resistant strains such as MDR-TB and XDR-TB, increase the difficulty in preventing the disease. According to the results obtained in the modeling, the Holt-Winters Exponential Smoothing method can be used to forecast the cases of TB diagnoses in Brazil, due to the high performance found in relation to error metrics. As a result, the present study is an apparatus relevant for strategic planning involving detection of outbreaks, epidemics or substantial increases in TB cases at an early stage, in parallel with reducing treatment costs and optimizing the decision-making process in controlling this disease.

6. Acknowledgement

The authors are thankful to Brazilian National Council for Scientific and Technological Development for making it possible through provision of resources to come up with this research work.

5. References

[1] ABDULLAH, S. et al. Application of Univariate Forecasting Models of Tuberculosis Cases in Kelantan. ICSSBE, 2012.

[2] AKHTAR, S.; MOHAMMAD, H. G. Seasonality in pulmonary tuberculosis among migrant workers entering Kuwait. BMC Infectious Diseases, v. 8, n. 3, 2008.

[3] ATKINSON, R. W. et al. Fine particle components and health – a systematic review and meta-analysis of epidemiological time series studies of daily mortality and hospital admissions. J Expo Sci Environ Epidemiol. v. 25, pp. 208-214, 2015.

[4] AZEEZ, A. et al. 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, n. 13, p. 757, 2016.

[5] BILLAH, B. et al. Exponential Smoothing Model Selection for Forecasting. International Journal of Forecasting, n. 22, p. 239-247, 2006.

[6] BORISOV, S. E. et al. Effectiveness and safety of bedaquiline-containing regimens in the treatment of MDR- and XDR-TB: a multicentre study. Eur Respir J. v. 49, ed. 5, 2017.

[7] CAO, S. et al. A hybrid seasonal prediction model for tuberculosis incidence in China. BMC Medical Informatics and Decision Making, v. 13, n. 56, p. 1-7, 2013.

[8] DOBRE, I.; ALEXANDRU, A. A. Modelling unemployment rate using Box-Jenkins procedure. Journal of Applied Quantitative Methods, v. 3, n. 2, p. 156-166, 2008.

[9] DRAIN, P. K. et al. Incipient and Subclinical Tuberculosis: a Clinical Review of Early Stages and Progression of Infection. Clinical Microbiology Reviews, v. 31, n. 4, p. 1-24, 2018.

[10] DRITSAKIS, N.; KLAZOGLOU, P. Time series analysis using ARIMA models: an approach to forecasting health expenditures in USA. ECONOMIA INTERNAZIONALE / INTERNATIONAL ECONOMICS , Genova (Italy), v. 72, n. 1, p. 77-106, 2019. ISSN 2499-8265.

[11] DUBE, N. Application and Comparison of Time Series Methods on Tuberculosis Incidence Data: A case study of Zimbabwe 1990-2013. Faculty of Texas Tech University, 2015.

[12] FALZON, D. et al. World Health Organization treatment guidelines for drug-resistant tuberculosis, 2016 update. Eur Respir J. v. 49, ed. 3, 2017.

[13] HOLT, C. C. Forecasting seasonals and trends by exponentially weighted moving averages. International Journal of Forecasting, v. 20, p. 5-10, 2004.

[14] JERE, S.; KASENSE, B.; CHILYABANYAMA, O. Forecasting Foreign Direct Investment to Zambia: A Time Series Analysis. Open Journal of Statistics, v. 7, p. 122-131, Fvereiro 2017. <https://doi.org/10.4236/ojs.2017.71010>.

[15] KHALIQ, A.; BATOOL, S. A.; CHAUDHRY, M. N. Seasonality and trend analysis of tuberculosis in Lahore, Pakistan from 2006 to 2013. Journal of Epidemiology and Global Health, n. 5, p. 397 403, 2015.

[16] KILICMAN, A.; ROSLAN, U. Tuberculosis in the Terengganu region: Forecast and data analysis. Science Asia, v. 35, p. 392–395, 2009.

[17] MAO, Q. et al. Forecasting the Incidence of Tuberculosis in China Using the Seasonal Auto-regressive Integrated Moving Average (SARIMA) Model. Journal of Infection and Public Health, n. 11, p. 707-712,

2018.

[18] MERKER, M. et al. Transmission of Extensively Drug-Resistant Tuberculosis in South África. Nature Genetics. v. 47, 2015.

[19] MINISTRY OF HEALTH. Tuberculosis Free Brazil: National Plan to End Tuberculosis as a Public Health Problem. Health Surveillance Secretariat, Communicable Disease Surveillance Department - Brasilia, 2017.

[20] MOHAMMED, S. H. et al. Seasonal Behavior and Forecasting Trends of Tuberculosis Incidence in Holy Kerbala, Iraq. International Journal of Mycobacteriology, v. 7, n. 4, p. 361-367, 2018.

[21] NASEHI, M. et al. Forecasting tuberculosis incidence in iran using box-jenkins models. Iran Red Crescent Med J, v. 5, n. 16, p. 1-6, 2014.

[22] NEWAZ, M. K. Comparing the Performance of Time Series Models for Forecasting Exchange Rate. BRAC University Journal, v. 5, n. 2, pp. 55-65, 2008.

[23] PAULINO, J. S. et al. Predictive Models and Health Sciences: A Brief Analysis. International Archives Of Medicine, v. 10, 2017. <http://imedicalsociety.org/ojs/index.php/iam/article/view/2271>.

[24] PIETERSEN, E. et al. Long-term outcomes of patients with extensively drug-resistant tuberculosis in South Africa: a cohort study. The Lancet, v. 383, 5–11, April, 2014.

[25] RIBEIRO, R. C. M. et al. Holt-Winters Forecasting for Brazilian Natural Gas Production. International Journal for Innovation Education and Research, v. 7, n. 6, p. 119-129, 2019.

[26] SILUYELE, I.; JERE, S. Using Box-Jenkins Models to Forecast Mobile Cellular. Open Journal of Statistics, v. 6, p. 303-309, 2016.

[27] STEIN, C. M. et al. Resistance and Susceptibility to Mycobacterium tuberculosis Infection and Disease in Tuberculosis Households in Kampala, Uganda. American Journal of Epidemiology, v. 7, n. 187, p. 1477–1489, 2018.

[28] TAHERI, S.; MAMMADOV, M. Learning the Naive Bayes classifier with optimization models. International Journal of Applied Mathematics and Computer Science, v. 23, n. 4, p. 787–795, 2013. <https://doi.org/10.2478/amcs-2013-0059>

[29] TULARAM, G. A.; SAEED, T. Oil-Price Forecasting Based on Various Univariate Time-Series Models. American Journal of Operations Research, v. 6, p. 226-235, 2016. <http://dx.doi.org/10.4236/ajor.2016.63023>.

[30] WINTERS, P. R. Forecasting sales by exponentially weighted moving averages. Management Science, v. 6, p. 324– 342, 1960.

[31] WORLD HEALTH ORGANIZATION. Global Tuberculosis Report. World Health Organization. Geneva. 2018.

[32] WUBULI, A. et al. Seasonality of active tuberculosis notification from 2005 to 2014 in Xinjiang, China. PLoS ONE, v. 12, n. 7, p. 1-12, 2017.

[34] ZHENG, Y. et al. Forecast Model Analysis for the Morbidity of Tuberculosis in Xinjiang, China. PloS ONE, v. 10, n. 3, p. 1-13, 2015.