Original Article

FORECASTING OF NEW CASES OF TB, USING BOX-JENKINS APPROACH

MUHAMMAD IMRAN[*], JAMAL ABDUL NASIR[**], SYED ARIF AHMED ZAIDI[***]
[*]Department of Statistics, MS Scholar
[**]Department of Statistics, Assistant professor, Islamia University, Bahawalpur, Pakistan
[***]Department of community medicine, Quaid-e-Azam Medical College Bahawalpur.

ABSTRACT:

OBJECTIVE:

Prediction through mathematical models allows us to better understand the development and the pattern of fatal diseases, a significant concern in the adaption of preventative measures. This study therefore aimed to uncover the trend and forecast the new TB incidences for the major districts of Punjab province, placed first in term of population as compared to other provinces of Pakistan.

METHODS:

The relevant data from 2001 to 2011 for the present study has been taken from the National TB Control program Ministry of National Health Service Government of Pakistan, Islamabad. A univariate modelling is used by taking X_t= (the number of new TB cases per quarter of four districts, namely Bahawalpur, Attock, Lahore and Rawalpindi were averaged) the total time series entities were 44.

RESULTS:

The main purpose of this study is to develop a forecasting model by incorporating the Box-Jenkins methodology for the new cases of Tuberculosis in major district of Punjab, Pakistan. Twenty four different ARIMA models have been attempted to forecast new cases of TB. Three measures namely Akaike, Hannan-Quinn and Schwarz are used to identify the efficient models. The final selected model, ARIMA (3, 1, 0) revealed the increment of new TB cases in major districts of Punjab—say for a 100 cases in 2011, 11 new cases are expected in 2012.

CONCLUSION:

A rising trend is expected in TB new cases in various districts of Punjab.

KEY WORDS: ARIMA model; Box Jenkins methodology; Forecast; Pakistan; Tuberculosis

INTRODUCTION:

Tuberculosis (TB) has been present in humans since olden time. The earliest explicit detection of M. tuberculosis involves evidence of the disease in the remains of bison dated to approximately 17,000 years ago. Researchers have found tubercular decay in the spines of Egyptian mummies dating from 3000–2400 BC.[1] Phthisis is a Greek word for consumption, an old term for pulmonary tuberculosis; around460 BC, ancient Greek physician identified phthisis as the most widespread disease of the times. It was said to involve fever and the coughing up of blood, which was almost always fatal. Although the pulmonary form associated with tubercles was established as pathology by Dr. Richard Morton in 1689, due to the variety of its symptoms. TB

Corresponding Author:
Muhammad Imran, Department of Statistics,
MS Scholar, Islamia University, Bahawalpur,
E-mail: Imranshakoor84@yahoo.com

was not identified as a single disease until the 1820s, and was not named tuberculosis until 1839 by J.L.Schönlein.[2]
The economic conditions correlate with new cases of TB. Ninety five percent of TB cases and 98% of TB deaths take place in low and middle income families.[3] The highest numbers of active TB cases are found in developing countries.[4-5] In 2012 there are an estimated 8.6 million new cases of TB and 1.3 million people died from the disease every year. [6] The highest incidence is seen in those countries of Africa, Asia, and Latin America with the lowest gross national products like Afghanistan, India, Pakistan, and Mozambique Cambodia.[6] Pakistan is a developing country of Asia. Its present population is estimated to be 183 million, spreads over an area of 796,095 sq. km. The population growth rate is 1.6% with life expectancy is about 63 years[7] and the average literacy rate is 58%. Health expenditure (public sector) is 0.08 % while total health sector investment is 3.9 % of GDP.[8] Pakistan is divided into the five provinces Punjab, Sindh, Baluchistan, Khyber Pakhtunkhwa and Gilgit baltistan.
Pakistan ranks 6th most populous country and also included among the highest burden countries (HBCs) of TB in the world.[6] Pakistan contributes about 55% of tuberculosis burden in the Eastern Mediterranean Region.[9] According to WHO, the incidence of sputum positive TB cases in Pakistan is 80/100,000 per year and for all types it is 177/100,000.[5] National TB Control Program reported that the case detection rate is 64% and an estimated incidence rate of 230/100,000 population.[10] TB is responsible for 5.1 percent of the total national disease burden in Pakistan.[9] To overcome this major health problem, The National TB Control Program (NTP) Pakistan adopted DOTS strategy in 1995 after the declaration of TB as a global emergency by the World Health Organization in1993, and now in Pakistan the DOTS achieved almost 100% coverage.[9]

BOX-JENKINS APPROCH IN MEDICAL FEILED:

From medical perspective the time series models play a significance role in disease prediction. Fazekas (2005)[11] used the autoregressive integrated moving average (ARIMA) models to analyze the Hungarian mortality rates. Sham et al. (2014) [12] used the Box-Jenkins methods to develop models which forecast the hand, foot and mouth disease in Sarawak, Malaysia. Soebiyanto and Kiang (2010)[13] used the ARIMA modelling approach on influenza transmission and found that the good agreement between actual and forecasted data. Promprou et al. (2006) [14] used Box-Jenkins modelling and found that ARIMA (1, 0, 1) was an adequate model to forecast the haemorrhagic fever cases in Southern Thailand. Permanasari et al (2009)[15] SARIMA (9, 0, 14) (12, 1, 24)$_{12}$ was selected as the most appropriate model for prediction of zoonosis Incidence in human. Gyasi-Agyei et al. (2014)[16] used the data of tuberculosis occurrence from January 2001 to March 2013 in the Ashanti Region and forecast for the period from April 2013 to April 2015 under the establish ARIMA (1, 0, 0) model. A similar study made in Iran, Moosazadeh et al. (2014)[17] applied the seasonal autoregressive integrated moving average (SRIMA) model on 84 month time series data set of tuberculosis cases and found that Box-Jenkins and SAIMA models suitable for predicting its prevalence in Future. Moosazadeh et al. (2013)[18] used seasonal ARIMA model (0, 1, 1) (0, 1, 1)$_{12}$ to predict TB cases in the North of Iran.

MATERIAL AND METHODS:

The relevant data from 2001 to 2011 with total 44 time series observations, for the present study has been taken from the National TB Control program (NTP)

Ministry of National Health Service Government of Pakistan, Islamabad. A univariate modelling is used by taking $X_t = \left(\frac{B+A+L+R}{4}\right)$, the number of new TB cases per quarter of four districts, namely Bahawalpur, Attock, Lahore and Rawalpindi are averaged. Where 'B', 'A', 'L' and 'R' denotes the new TB cases in Bahawalpur, Attock, Lahore and Rawalpindi districts respectively per quarter.

BOX-JENKINS METHODOLOGY:

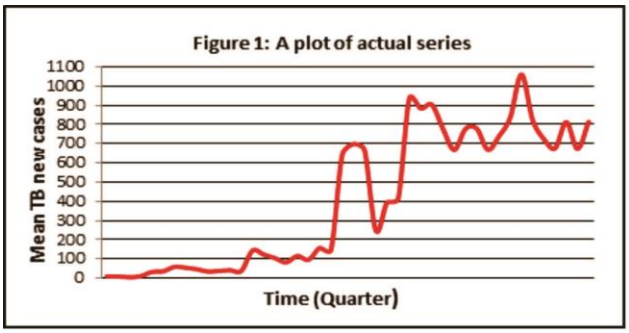
The Box-Jenkins (BJ) Methodology: It is a systematic procedure of establishing an adequate model using integrated autoregressive moving average (ARIMA) time series models.[19] A model consist of three parameters one autoregressive (p), second differencing order (d) and third moving average order(q). The general equation of the autoregressive moving average model is shown in equation (1)

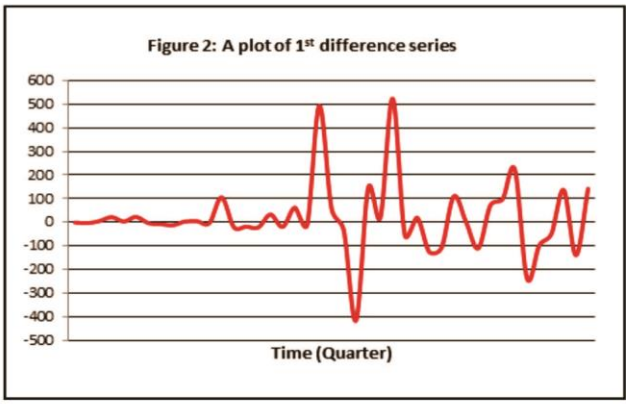
$$X_t = c + \phi_1 X_{t-1} + \ldots\ldots\ldots\ldots + \phi_P X_{t-P} + e_t - \theta_1 e_{t-1} - \ldots\ldots\ldots - \theta_q e_{t-q} \ldots\ldots\ldots (1)$$

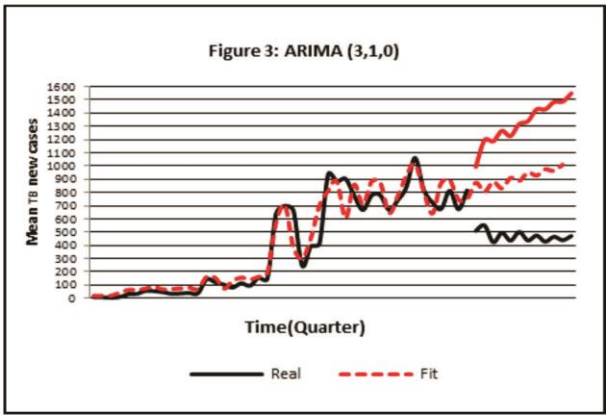
Where "ϕ" and "θ" are the autoregressive and moving average parameters to be estimated. "X's" and the "e's" are the original series and residuals respectively. The residuals assumed to follow a normal probability distribution. It consist four iterative steps Identification, estimation of parameters, diagnostic checks and finally forecast.[19-21]A procedure is enhanced step by step.

IDENTIFICATION:

In the identification we find out the appropriate values of p, d, and q by using the plot of auto correlation function (ACF) and partial auto correlation function (PACF). Before obtaining the suitable value of p and q; the first step to check the series is stationary or not. Augmented Dickey Fuller Test is used to verify the stationarity.[22,23] When d=1 the series become stationary. The plot of original series is shown in Figure1 while the first differential is presented in Figure 2 by taking quarterly time period along x-axis and mean new cases of TB along y-axis. The details of Augmented Dickey Fuller test for stationary including trend and constant is shown in table 1.







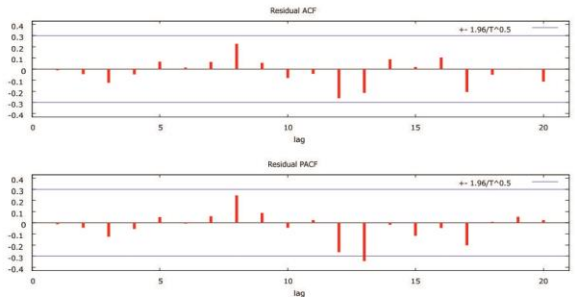


Figure 4: ACF and PACF plot

ESTIMATION:

After the identification of the parameters. The next stage is to estimate the parameters of the autoregressive and moving average terms included in the model.

Here we use statistical packages Minitab 15.0 and Gnu Regression, econometrics and time-series library (gretl)[24] software to execute the various combinations of the ARIMA parameters.

Table 1: A detail of Augmented Dickey Fuller test for stationary

Test	Test statistic	p-value	Difference Level
ADF	-1.69	0.75	d=0
ADF	-7.26	0.00	d=1

DIAGNOSTICS:

Having chosen a particular ARIMA model and having estimated its parameters, we next see whether the chosen model is adequate or not. This is called Diagnostic checking (Goodness of Fit). The fit is good because the p-value at various lags in Table 3 is greater than the level of significance (5%). According to this the model with least AIC value will be selected. We executed 24 ARIMA models and choose that model which has minimum AIC (Akaike Information Criterion). M14 is considering as an appropriate model, the models and corresponding AIC, HQ and SBIC values and the Ljung-Box Statistics p-value at various lags are shown in Table 2.

Table 2: The detail of ARIMA models with model selection criterion and p-value at various lags

Models	Parameters			Ljung-Box Statistics P-value at various lag			Model selection criterion		
	P	d	q	Lag12	Lag24	Lag36	AIC	HQ	SIC
M1	1	1	0	0.004	0.008	0.175	558	560	563
M2	1	1	1	0.017	0.03	0.356	555	558	562
M3	0	1	1	0.004	0.007	0.171	558	560	563
M4	0	1	2	0.004	0.007	0.171	557	559	564
M5	0	1	3	0.633	0.485	0.935	544	547	552
M6	1	1	2	0.003	0.003	0.092	544	547	552
M7	1	1	3	0.532	0.367	0.88	545	549	556
M8	2	1	0	0.002	0.005	0.146	560	563	567
M9	3	1	1	0.215	0.299	0.867	545	549	556
M10	3	1	2	0.198	0.277	0.857	547	551	559
M11	3	1	3	0.103	0.068	0.503	548	553	562
M12	3	1	4	0.068	0.261	0.81	546	552	562
M13	2	1	1	0.005	0.073	0.544	557	560	566
M14*	3	1	0	0.288	0.371	0.901	543	546	552
M15	2	1	3	0.157	0.05	0.408	547	551	559
M16	2	1	2	0.003	0.036	0.418	554	558	565
M17	2	1	4	0.188	0.098	0.581	544	548	557
M18	4	1	1	0.058	0.217	0.762	546	550	558
M19	4	1	0	0.213	0.303	0.869	545	549	555
M20	0	1	4	0.476	0.169	0.638	546	550	556
M21	1	1	4	0.22	0.431	0.925	546	550	558
M22	4	1	2	0.135	0.228	0.823	549	554	563
M23	4	1	3	0.11	0.172	0.74	548	554	564
M24	4	1	4	0.058	0.217	0.762	549	555	566

*Lowest value of AIC, SIC and HQ

Table 3: Final Estimates of Parameters with95% confident limits

Variable	Coefficients
Constant	18.3914[-2.160,38.94]
$\phi 1$	-0.1187 [-0.349,0.111]
$\phi 2$	-0.0504 [-0.283,0.182]
$\phi 3$	-0.5904 [-0.816,-0.364]

RESULTS AND DISCUSSION: After choosing the best model ARIMA (3, 1, 0), a quarterly average number of TB new cases for Bahawalpur, Attock, Lahore and Rawalpindi can be obtained by the selected ARIMA model. Table 3 gives the detail of final estimate of the parameters along with constant and 95% confidence interval limits. A model predicted plot of TB new cases with actual and forecast with 95% confidence interval limits is presented in figure 3. While the plot of ACF and PACF are given in Figure 4. The autocorrelation function (ACF) at all lags is well in inside the 95% critical limits ($1.96\frac{1}{\sqrt{44}} = 0.30$) while the partial auto correlation function (PACF) at lag 13 is outside the limits in Figure 4. But it is not a matter of concern because about 5% of the spikes fall a short distance beyond the 95% critical limits due to chance. The average annual TB new cases in four districts for 2011 are 2976.50 while the predicted cases are 3316.71 for 2012. The expected increment is 11.43% in TB new cases in 2012 as compared to 2011.

By substituting the final estimates of parameters, the model equation is given below in equation (2).

$$X_t = X_{t-1} - 0.1187(X_{t-1} - X_{t-2}) - 0.0504(X_{t-2} - X_{t-3}) - 0.5904(X_{t-3} - X_{t-4}) + 18.3914 \quad (2)$$

CONCLUSION:

A rising trend is expected in TB new cases in various districts of Punjab namely, Bahawalpur, Attock, Lahore and Rawalpindi. ARIMA (3,1,0) is an effective mathematical model for prediction of mean TB new incidences/quarter. This study will serve as a guide to understand the development and trend of new TB occurrence. On the other hand availability of statistical finding about infectious fatal disease (TB) can be handy to the policy planner as well as the health department of Pakistan to take steps in the improvement of TB vaccinations and prevention in the Punjab as well as other provinces.

ACKNOWLEDGMENT:

Authors would like to special thanks Ms. Maryam Samad and National TB Control program for accessing to data for this study.

REFERENCES:

1. E.Crubezy, H. D. A. Multiple bone Tuberculosis in a child from Predynastic Upper Egypt (3200 BC). *International Journal of Osteoarchaeology* 2009*; 1-12.*
2. Availableat:http://en.wikipedia.org/wiki/tuberculosis (accessed in February, 2014).
3. World Health Organization Tuberculosis report, Geneva 2000
4. Sandro et al. Impact of Insecurity, the Aids Epidemic, and Poverty on Population Health: Disease Patterns and Trends in Northern Uganda. *Am. J. Trop. Med. Hyg* 2001; 64**:** 214-221.
5. World Health Organization Tuberculosis report, Geneva 2001
6. Organization, W. H. Global Tuberculosis Report 2012. Geneva: World Health Organization.

7. Pakistan Demographic and Health Survey (2012-13). Preliminary Report, National Institute of Population Studies Islamabad, Pakistan.
8. Pakistan Economic Survey (2012-2013). Ministry of Finance, Government of Pakistan.
9. National TB control program (2012). Ministry of National Health Service, Regulation and Coordination Government of Pakistan.
10. National TB Control Program Annual Report (2011).Ministry of National Health Services, Regulation & Coordination Islamabad, Pakistan.
11. FAZEKAS M. time series models on medical research. periodica polytechnica ser el eng 2005; 49(3-4):175-181.
12. Noraishah Mohammad Sham IK, Mahendran Shitan, Munn-Sann Lye: time series model on hand, foot and mouth disease in sarawak, malaysia. asian pacific journal of tropical disease 2014; 4(6): 469-472.
13. R. P. Soebiyanto RKK: modeling influenza transmission using environmental parameters international archives of the photogrammetry, remote sensing and spatial information science 2010; xxxviii: 330-334.
14. S. Promprou, M. Jaroensutasinee and K. Jaroensutasinee: Forecasting Dengue Haemorrhagic Fever Cases in Southern Thailand using ARIMA Models. Dengue Bulletin 2006; 30:99-106.
15. Adhistya Erna Permanasari DRARaPDDD: Prediction of Zoonosis Incidence in Human using Seasonal Auto Regressive Integrated Moving Average (SARIMA) International Journal of Computer Science and Information Security 2009;5(1):103-110.
16. Kwame Asare Gyasi-Agyei, A. G.-A. a. W. O.-D. Mathematical Modeling of the Epidemiology of Tuberculosis in the Ashanti Region of Ghana. *British Journal of Mathematics & Computer Science* 2014; 4(3): 375-393.
17. Moosazadeh et al: Forecasting Tuberculosis Incidence in Iran Using Box-Jenkins Models *Iran Red Crescent Med J* 2014; 16(5): 1-6.
18. Mahmood Moosazadeh NK, Abbas Bahrampour. Seasonality and Temporal Variations of Tuberculosis in the North of Iran. Tanaffos 2013; 12(4):35-41.
19. Box, G. P. & Jenkins, G. M. 1976. Time series analysis: forecasting and control, San Francisco, Calif., Holden-Day.
20. Box, G. E. P., Jenkins, G. M. & Reinsel, G. C. 1994. Time series analysis: forecasting and control, Englewood Cliffs, N.J., Prentice Hall.
21. Chatfield, E. The Analysis of Time Series: An Introduction, Chapman and Hall Ltd., London, 1991.
22. David A. Dickey, W. A. F. Distribution of the estimators for Autoregressive time series with a unit root. *Journal of the American Statistical Association* 1979; 74: 427-431.
23. Dhrymes, P. J. 1998. Time series, unit roots, and cointegration, San Diego, Academic Press
24. Adkins LC (2007). Using gretl for Principle of Econometrics, 3rd edition: version 1.0 url http://www.LearnEconometrics.com/gretl/ebook.pdf

Submitted for publication: 27-08-2014

Accepted for publication: 02-10-2014