DOI: http://dx.doi.org/10.26483/ijarcs.v8i7.4424

ISSN No. 0976-5697

Volume 8, No. 7, July – August 2017

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

# DEVELOPMENT OF A SARIMA MODEL TO FORECAST TUBERCULOSIS DETECTION RATE IN THE DIBRUGARH DISTRICT OF ASSAM, INDIA

A. N. Patowary
College of Fisheries
Assam Agricultural University, Nagaon
Raha-782103, Assam, India

M. P. Barman
Department of Statistics
Dibrugarh University, Dibrugarh-786004
Assam, India

***Abstract:*** Predicting the incidence of tuberculosis (TB) plays an important role in planning health control strategies for the future, developing intervention programs and allocating resources. The techniques of time series analysis and forecasting have become a major tool in predicting different health issues. In this context, an attempt has been made to fit seasonal autoregressive integrated moving average (SARIMA) model to the historical data of quarterly TB detection rate in Dibrugarh district of India for the period of 2001-2011, a total of 44 data points. We investigated and found that ARIMA $(0,0,0)\times(1,1,0)_4$ model is suitable for the given data set.

***Keywords:*** Tuberculosis, Seasonal Autoregressive Integrated Moving Average (SARIMA), Additive Decomposition, Q-Q plot, normwn.test

## 1. INTRODUCTION

Tuberculosis (TB) is caused by bacillus Mycobacterium tuberculosis is one of the major health problem around the globe which causes morbidity to millions of people every year. [16] In the year 2015, it is estimated that about 10.4 million people developed TB globally. TB is fatal in the sense that it is one of the leading causes of death worldwide. TB ranked one of the top 10 causes of death globally with about 1.5 million people died due to the disease in the year 2013 and 1.8 million in 2015 [16]. Past research showed that malnutrition, poverty, deprivation, overcrowding etc. are some of the factors associated with the infection of the disease [14], [8]. These may be the reasons why the incidence and mortality of TB is very high in poor countries than in comparison to developed countries. The burden of disease is abysmal in case of India with highest incidence of TB cases. Six countries (Indonesia, China, Nigeria, Pakistan and South Africa) together with India constitute 60% of the total incidence of TB [16]. In the year 2013, out of the estimated global incidence of 9 million TB cases, about 2.1 million were estimated to have occurred in India. WHO estimated that in 2015 the incidence of TB in India was 2.8 million i.e., 217 per 1,00,000 people. Mortality due to TB is also very high in India which was estimated to be 19 per 1,00,000 people in the year 2013 [16], [9]. The World Health Organization declared TB to be global health emergency in 1993. To check the burden of the disease in India National TB Control programme was launched in the year 1962 which failed to achieve its goal. A new TB control programme under the name of Revised National TB Control Programme (RNTCP) was initiated with new strategies to fight against this deadly disease in the year 1997. Under this programme, WHO recommended treatment strategy of Directly Observed Treatment Short Course (DOTS) was launched to provide proper treatment to the patients. Present trends showed decline in the burden of the disease which may be due to the proper implementation of DOTS. From reports it can be observed that there is a 55% reduction in TB prevalence rate in 2013 in comparison to 1990. A decreasing trend can also be observed in case of mortality also as mortality reduced about 50% in 2013 in comparison to 1990 [9]. Though the burden of the disease has been decreasing still it is very high in comparison to other parts of the world. The state of Assam is one of the states in India who has topped the tally in case of incidence of TB. The proper estimation and projection of the burden of TB is of very much important for proper handling of the disease. It is also important for reducing the socio-economic burden of a country as previous research showed that it mainly effect economically productive age groups which in turn bring various social issues [13], [11]. One of the important aspects in reducing the burden of the disease is to increase the detection rate of the disease so that the infected people get the proper medical treatment.

Previous attempts to estimate the burden of TB in India are based on indirect methods characterized by substantial uncertainty and lack of substantial detail. More accurate estimation and projection of tuberculosis burden in India is needed to guide policy making, to improve assessment of Government efforts and to understand global trends in the incidence of TB. According to WHO, the burden of the disease in India is considerably higher than previously estimated [16] . Considering the high burden of the disease in India, it is of great importance to project its future trends which will provide valuable information about the likely occurrence in future. It will help the planners in formulating proper policy to fight against the disease. The most important issue here is to choose an appropriate projecting method. Thus, this research work is initiated to assess whether the seasonal autoregressive integrated moving average (SARIMA) model can be used to project the TB detection rate.

Attempts were made for predicting the prevalence of TB by using different time series models around the globe some of which are mentioned here. [2] used SARIMA model and a combined model of (SARIMA) model and a neural network auto-regression (SARIMA-NNAR) model to analyzing and predicting the TB data in Eastern Cape of South Africa. Similarly, [3] used hybrid seasonal prediction model for

Tuberculosis in incidence in China. Also, [15] studied chronological Tuberculosis data by predictive time series models. Further, [17] developed time series model for forecasting morbidity of Tuberculosis in Xinjang, China. Moreover, [10] predicted the incidence of Smear Tuberculosis cases in Iran using time series analysis. The researchers could find some of literature in this direction conducted in India. [7] analyzed seasonality of Tuberculosis in Delhi. Similarly, [12] analyzed seasonality of Tuberculosis across Indian states and union territories. Moreover, [4] assessed the seasonality in TB in rural West Bengal and to developed SARIMA model to TB data. Nevertheless, as far the knowledge of the researchers there are no works available for analyzing and predicting detection rate of TB cases in Assam. Keeping all these points in mind, this research work is initiated to develop a seasonal autoregressive integrated moving average (SARIMA) model to quarterly TB detection rate data in Dibrugarh district of Assam, India, for the period of 2001-2011, a total of 44 data points, so that the model can be used in future for forecasting TB cases in Dibrugarh.

The organization of the paper is as follows: In Section 2, we have given the sources of data and described the complete methodology undertaken in this study. We have given developing of SARIMA model to the data in Section 3. Conclusion is given in Section 4.

## 2. SOURCE OF DATA AND METHODOLOGY OF THE STUDY

The collected data is purely secondary in nature. It is taken from Revised National Tuberculosis Control Programme (RNTCP), Dibrugarh. In Assam, Dibrugarh was the place where the Revise National Tuberculosis programme was first implemented in the year 1998-1999. After the successful implementation in Dibrugarh district, RNTCP programme now covers all the 23 district of Assam. One of the reasons for considering the detection rate of TB for Dibrugarh district is that it is one of the highest in comparison to the other districts of Assam and it will provide valuable information about the number of future TB patients. While calculating the detection rate the projected population based on census population of 2001 is used.

The methodology and the theorems propounded by Box and Jenkins (1970) called the Seasonal Autoregressive Integrated Moving Average (SARIMA) has been used in this study. This is an advance technique of forecasting which requires long seasonal time series data. This model decomposes historical data into an Autoregressive (AR) process, where there is a memory of past values, an Integrated (I) process, which accounts for stabilizing or making the data stationary plus a Moving-Average (MA) process, which accounts for previous error terms making it easier to forecast. The multiplicative seasonal autoregressive integrated moving average (SARIMA) model, of Box and Jenkins (1970) is given by [6]

$$\Phi_p(B^S)\phi(B)\nabla_S^D\nabla^d X_t = \mu + \Theta_Q(B^S)\theta(B)e_t \tag{2.1}$$

where $e_t$ is the usual white noise process. The general model is denoted by ARIMA (p, d, q) (P, D, Q)$^S$. The ordinary autoregressive and moving average components are represented by the following polynomials $\phi(B)$ and $\theta(B)$ of orders p and q, respectively,

$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - .... - \phi_p B^p \tag{2.2}$$

$$\theta(B) = 1 + \theta_1 B + \theta_2 B^2 + .... + \theta_q B^q \tag{2.3}$$

and the seasonal autoregressive and moving average components are represented by the following polynomials $\Phi_P(B^S)$ and $\Theta_Q(B^S)$ of order P and Q respectively,

$$\Phi_P(B^S) = 1 - \Phi_1 B^S - \Phi_2 B^{2S} - ... - \Phi_P B^{PS} \tag{2.4}$$

$$\Theta_Q(B^S) = 1 + \Theta_1 B^S + \Theta_2 B^{2S} + ... + \Theta_Q B^{QS} \tag{2.5}$$

Seasonal difference components are represented by:

$$\nabla^d = (1-B)^d \text{ and } \nabla_S^D = (1-B^S)^D$$

The steps involving in Box and Jenkins (1970) methodology are given below:

**Phase 1**

(a) Data Preparation: Transform data to stabilize variance and difference data to obtain stationary series.

(b) Model Selection: Examine autocorrelation function (ACF) and partial autocorrelation function (PACF) to identify potential models.

**Phase 2**

(a) Estimation: Estimate parameters in potential models. Select the best model using suitable criterion.

(b) Diagnostic: Check ACF and PACF of residuals. Examine residuals follow white noise or not. If it does not follow white noise, select another model by model selection criterion.

**Phase 3**

(a) If residuals follow white noise, use the model for forecasting.

## 3. RESULTS AND DISCUSSION

Data consisting of monthly TB detection rate for 11 years, starting on 1st quarter 2001 to 4th quarter 2011 of Dibrugarh district has been used in this study.

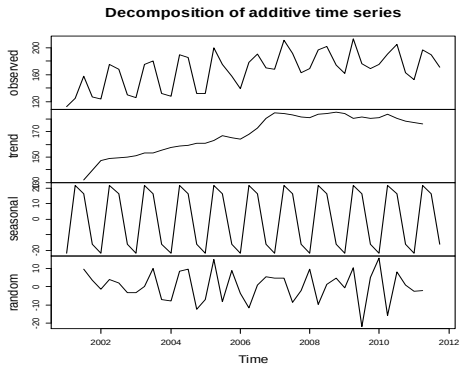


**Fig. 3.1: Decomposition of Time Series by Additive Method during 2001-2011**

We have plotted year in X-axis and observed TB detection rate in Y-axis in Fig.3.1. To examine the trend clearly, we decompose the data by additive decomposition method using the statistical software R as depicted in Fig. 3.1. From decomposition, a distinct significant upward trend of TB detection rate from 2001 to 2011 is observed. Likewise, it is also observed that the presence of strong seasonal cycle in the data set. Next, for developing SARIMA model, we divide our data set in to two portions. We build up the seasonal ARIMA model to quarterly TB detection rate data for the first portion dataset i.e. period 2001-2009 and match the forecasted value obtained from the model with original one for the years 2010 and 2011 for checking the validity of the selected model.

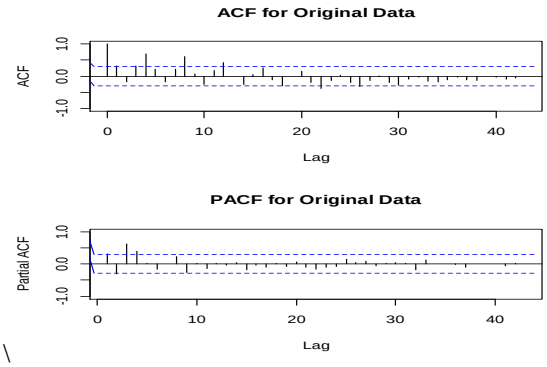


**Fig. 3.2: ACF and PACF of the original data**

Fig. 3.2 consists of plots of ACF and PACF taking 40 lag values in X-axis and autocorrelation values in Y-axis for the quarterly TB detection rate. The seasonal autocorrelation relationships are also shown in this display. As shown in Fig. 3.1 and Fig. 3.2, due to the presence of strong upward trend and seasonality which indicates that our data is non stationary. But, according to B-J methodology we must ensure that the time series being analyzed is stationary before we fit SARIMA model.

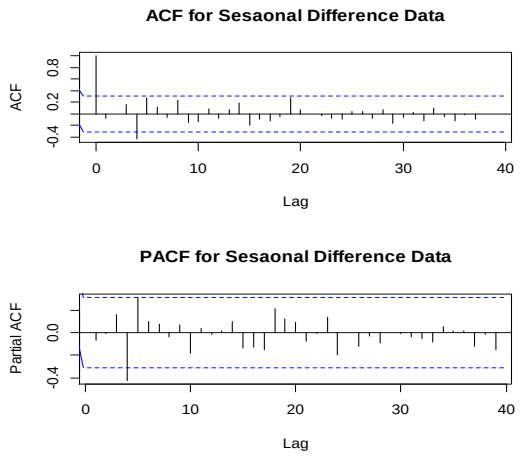


**Fig. 3.3: ACF and PACF of D=1 of original data**

Fig. 3.3 shows ACF and PACF of the of the detection rate data after taking a first seasonal difference i.e. D=1. From Fig. 3.3, it is observed that very little autocorrelation present in the series after taking first seasonal difference. Also, the p-value=0.1 of the KPSS test is greater than 0.05, so, we cannot reject the null hypothesis of level or trend stationary [5]. Further, in Fig.3.3, a highly significant spike at seasonal lag 4 in the PACF is observed which suggesting a seasonal AR(1) i.e. P=1. Moreover, there is no any significant information can be obtained from Fig. 3.3. Therefore, we acquire a rough idea that the ARIMA$(0,0,0)\times(1,1,0)_4$ might be feasible to our data. Also, automatic ARIMA function in R software i.e. auto.arima() function select the same model i.e. ARIMA$(0,0,0)\times(1,1,0)_4$ to the data. Therefore, our selected model is given by

$$(1-B^4)X_t = (1-\Phi_1 B^4)e_t \tag{3.1}$$

The maximum likelihood estimate of $\Phi_1$ obtained from R software is as follows:

Table 3.1: Estimated parameter value

| Parameter | s.e. | z-value | p-value |
|---|---|---|---|
| $\hat{\Phi}_1 = -0.4290$ | 0.1604 | -2.6799 | 0.0007 |

and then estimated model Eq. (3.1) is given by

$$(1-B^4)X_t = (1+0.4290B^4)e_t \tag{3.2}$$

The coefficient of the estimated model is highly significant because p-value is less than 0.05. Further, for checking, we have also considered other ARIMA models to our detection rate data with different combinations of p, d, q, P, D and Q and compared their performance using Akaike Information Criterion (AIC) [1] found that the ARIMA$(0,0,0)(1,1,0)_4$ model selected initially has the lowest AIC value. Thus, in the next step we go for the diagnostic checking of fitted ARIMA$(0,0,0)(1,1,0)_4$ model.

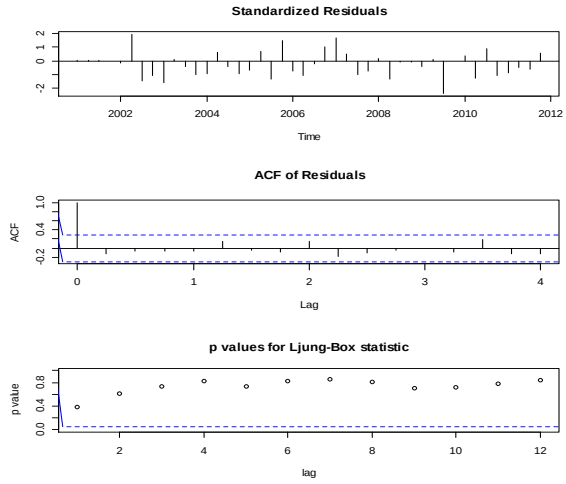


**Fig. 3.4: Diagonistic checking of residuals**

If the model fits well, the standardized residuals estimated from this model should behave as an i.i.d. (independent and identically distributed) sequence with mean zero and variance $\sigma^2$. Such a sequence is referred to as white noise. Fig. 3.4 displays a plot of the standardized residuals, the ACF of the residuals and the p-values of the Q-statistic at lag 1 through 12. From standardized plot of residuals, all residuals fall inside the limit of -3 and +3. Here, the Ljung-Box test statistic is 7.078, and the p-value is 0.7927, so we cannot reject the null hypothesis of independence in this residual series. Using the white noise test (from the normwn.test package in R: Perform a univariate test for white noise), we obtain the p-value of 0.0911 which means

that the residuals series is white noise (with mean 0 and variance $\sigma^2$). To be sure that the predictive model cannot be improved upon, it is also a good idea to check whether the forecast errors are normally distributed with mean zero and constant variance. The histogram of the forecast error is given in Fig. 3.5.

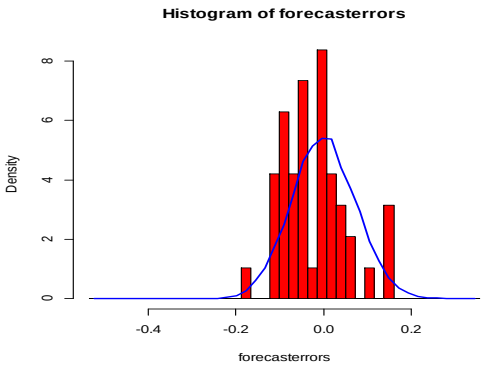


**Fig. 3.5: Histogram of residuals**

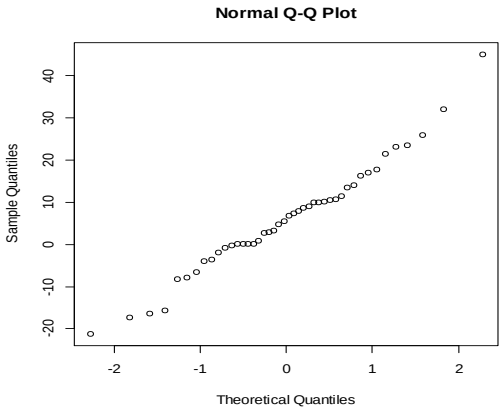


**Fig. 3.6: Q-Q plot of residuals**

Fig. 3.5 shows that the distribution of forecast errors is roughly centered on zero, and is more or less normally distributed. Also, we have drawn quantile-quantile (q-q) plot of residuals in Fig 3.6 which is the evidence of normality of residuals. Moreover, we have applied Komogorov-Smirnov (K-S) test to the residuals and found that D=0.1166 and p-value =0.5483 which indicates that the residuals follow normal distribution well.

**Table 3.2: Actual Observations with Forecasted values for 2010 and 2011**

| Year/Quarter | Actual Observations | Forecasted Observations |
|---|---|---|
| 2010-Q1 | 175 | 164 |
| 2010-Q2 | 190 | 209 |
| 2010-Q3 | 205 | 183 |
| 2010-Q4 | 163 | 170 |
| 2011-Q1 | 153 | 163 |
| 2011-Q2 | 196 | 209 |
| 2011-Q3 | 205 | 181 |
| 2011-Q4 | 171 | 170 |


**Fig 3.7: Actual vs Forecasted for 2010 and 2011**

Now, to check the accuracy of the fitted model, the actual observations are tabulated with forecasted values obtained from the model for the period 2010-2011 (8 quarters) in Table 3.2. We have also drawn the graph (Fig. 3.7) where red line represents actual TB detection rate whereas the black line represents forecasted TB detection rate for 2010 and 2011. In Fig. 3.7, it is observed that the forecasted TB detection rate from the model shows same picture as in the case of actual detection rate data for the years 2010 and 2011. Therefore, our selected model would be good fitted to the observed quarterly TB detection rate data in Dibrugarh for the period 2001-2009. Now, we refit the model ARIMA(0,0,0)×(1,1,0) to the whole dataset i.e. 2001-2011 and forecasted TB detection rate for the years 2012, 2013 and 2014 which are given in Table 3.3.

**Table 3.3: Forecasted TB detection rate in Dibrugarh for 2012, 2013 and 2014**

| Year/Quarter | Forecasted Detection Rate | Year/Quarter | Forecasted Detection Rate |
|---|---|---|---|
| 2012-Q1 | 163 | 2013-Q3 | 193 |
| 2012-Q2 | 193 | 2013-Q4 | 169 |
| 2012-Q3 | 196 | 2014-Q1 | 160 |
| 2012-Q4 | 168 | 2014-Q2 | 194 |
| 2013-Q1 | 158 | 2014-Q3 | 199 |
| 2013-Q2 | 195 | 2014-Q4 | 168 |

## 4. CONCLUSION

In this paper, the quarterly TB detection rate in the Dibrugarh region has been modeled by using SARIMA model. The estimation and diagnostic analysis results revealed that the selected model ARIMA$(0,0,0)\times(1,1,0)_4$ is adequately fitted to the historical data. The residual analysis, confirmed that there is no violation of assumptions in relation to the model adequacy. While comparing the observe detection rate with the expected estimated from the fitted model, both of them are found to be close. This shows that validity of the model is good. Thus, this estimated model can be used in future for projecting the detection rate.

## REFERENCE

[1] Akaike H. (1974): "A New Look at the Statistical Model Identification", IEEE Transactions on Automatic Control, Vol. 19, Issue: 6, pp.716-723.

[2] Azeez A., Obaromi D., Odeyemi A., Nedge J. and Muntabayi R. (2016): "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, Vol.13, pp.2-12.

[3] Cao S., Wang F., Tam W., Tsc L.A., Kim J. H., Liu J. and Lu Z. (2013): "A Hybrid Seasonal Prediction Model for Tuberculosis Incidence in China", BMC Medical Informatics and Decision Making, doi:10.118611472-6974-13-56.

[4] Chowdhuri R., Mukherjee A., Naska S., Adhikary M. and Lahiri S. K. (2013): "Seasonality of Tuberculosis in Rural West Bengal: A Time Series Analysis", International Journal of Health and allied Sciences, Vol2, Issue:2, pp.95-98.

[5] Gimeno R., Manchado B. and Minguez R. (2009): "Stationarity tests for Financial Time Series", Physica A: Statistical Mechanics and Its Applications, Vol. 269, Issue: 1 pp. 72-78.

[6] Hazarika J., Pathak B. and Patowary A. N. (2017): "Studying Monthly Rainfall over Dibrugarh, Assam: Use of SARIMA Approach", Mausam, Vol.68, No.2, pp.349-356.

[7] Kumar V., Singh A., Adhikary M., Daral S., Khokhar A. and Singh S. (2014): "Seasonality of Tuberculosis in Delhi, India: A Time Series Analysis", Tuberculosis Treatment and Research.

[8] Lienhardt C, Rowley J, Manneh K, Lahai G, Needham D and Milligan P. (2001): "Factors affecting time delay to treatment in a tuberculosis control program in a sub-Saharan African country: the experience of the Gambia", International Journal Tuberculosis and Lung Disease, Vol.5, No.3, pp.233-2399.

[9] Ministry of Health and Family Welfare (2015): "TB India 2015, Revised National TB Control Programme: Annual Status Report" Indian Medical Association for Central TB Division, Directorate General of Health Services, Ministry of Health and Family Welfare.

[10] Moosazadeh M., Khanjani N., Nasehi M. and Bahrampour A. (2015): "Predicting the Incidence of Smear Positive Tuberculosis Cases in Iran using Time Series Analysis", Iran Journal of Public Health, Vol.44, No.11, pp.1526-1534.

[11] Muniyandi M, Ramchandran R, Balasubramanian R. and Narayanan P. R. (2006): "Socio-economic dimensions of tuberculosis control: Review of studies over two decades from Tuberculosis Research Centre", The Journal of Communicable Disease, Vol. 39, No.3, pp.204-15.

[12] Narula P., Sihota P., Azad S. and Lio P. (2015): "Analyzing Seasonality of Tuberculosis across Indian States and Union Territories", Journal of Epidemiology and Global Health, Vol.5, Issue 4, pp.337-346.

[13] Rajeswari R, Balasubramanian R, Muniyandi M, Geetharamani S, Thresa X. and Venkatesan P. (1999): "Socio-economic impact of tuberculosis on patients and family in India" International Journal of Tuberculosis and Lung Disease, Vol.3, Issue:10, pp.869–77.

[14] Rieder H. L. (1999): "Epidemiologic Basis of Tuberculosis Control", 1st ed. International Union against Tuberculosis and Lung Disease, Paris, France 1999, pp.50–52

[15] Wah W., Das S., Earnest A, Lim L.K.Y., Che C. B., Cook A. R., Wang Y. T., Win K. M. K., Ong M. E. H. and Hsu L. Y. (2014): "Time Series Analysis of Demographic and Temporal Trends of Tuberculosis in Singapore, BMC Public Health, doi:10.118611471-2458-14-1121.

[16] World Health Organization (2016): "Global Tuberculosis Report 2016", Geneva, Switzerland, ISBN: 9789241565394

[17] Zheng Y. L., Zhang L. P., Zhang X. L., Wang K. and Zheng Y. J. (2015): "Forecast Model Analysis for the Morbidity of Tuberculosis in Xinjang, China", PLoS ONE, Vol. 10, No. 3, doi:10.13711/journal.pone.0116832.