www.ThePharmaJournal.com

The Pharma Innovation

ISSN (E): 2277-7695
ISSN (P): 2349-8242
NAAS Rating: 5.23
TPI 2023; 12(1): 180-184

www.thepharmajournal.com
Received: 20-12-2022
Accepted: 30-01-2023

Rohit Kundu
Department of Mathematics and Statistics, College of Basic Sciences and Humanities, Chaudhary Charan Singh Haryana Agricultural University, Hisar, Haryana, India

OP Sheoran
Department of Mathematics and Statistics, College of Basic Sciences and Humanities, Chaudhary Charan Singh Haryana Agricultural University, Hisar, Haryana, India

Sanjeev
Department of Mathematics and Statistics, College of Basic Sciences and Humanities, Chaudhary Charan Singh Haryana Agricultural University, Hisar, Haryana, India

Corresponding Author:
Rohit Kundu
Department of Mathematics and Statistics, College of Basic Sciences and Humanities, Chaudhary Charan Singh Haryana Agricultural University, Hisar, Haryana, India

Modelling of tuberculosis prevalence through Bayesian technique

Rohit Kundu, OP Sheoran and Sanjeev

DOI: https://doi.org/10.22271/tpi.2023.v12.i2c.18433

Abstract
Tuberculosis, or TB, is a serious global health problem that is the leading cause of death from a single infectious disease. There are various approaches used by the research workers for modelling TB prevalence and Bayesian technique is one of them. For this study, the number of notified cases of tuberculosis in India from 1999 to 2019 were analyzed, using data from annual reports published by the Ministry of Health and Family Welfare. To model the number of TB cases in India, the current study used a counting process, represented by the random variable $\{N(t), t \in [0, \infty)\}$, which counts the number of TB cases each year. The study utilized a Non-Homogeneous Poisson Process (NHPP), which allows for the average rate of TB cases to vary with time, in their analysis. The parameters of the posterior distribution were assumed to follow a uniform distribution, with certain hyperparameter values assigned to specific intervals of time. The Markov Chain Monte Carlo (MCMC) method was used to analyze the data. The proposed model was found to fit the data well, as demonstrated by the results of the chi-square test for goodness of fit.

Keywords: Tuberculosis, MCMC, NHPP, counting process

Introduction
India is a developing country and it faces several developmental challenges. One of the major challenges is the diseases which severely affects the quality of human resource and a significant amount of expenditure is done on curing the diseases. Tuberculosis disease is a major problem in developing and under-developed countries and case of India is not different. India accounts for 26% of global TB cases and 24% of the gap between projected TB incidence and the number of patients newly diagnosed in 2020 (WHO, 2021) [11]. TB is also a leading cause of death among HIV patients and is the 9th leading cause of death worldwide (WHO, 2021) [11].

Bayesian inference involves using data, which is assumed to be fixed, and treating unknown parameters as random variables. The goal is to determine the probability of a given parameter (θ) given a set of observed data (x). The Bayesian approach utilizes prior information about θ, in addition to the likelihood, to calculate the posterior distribution of the unknown parameter. This prior information can be either informative, based on external information about the distribution of the parameter(s) of interest, or non-informative, in the absence of such information. In the context of tuberculosis modeling, non-informative priors, such as Jeffreys prior, Improper Prior and uniform prior are often used, while informative priors are less common. Severity of the TB is measured by several measures and prevalence is one among them. The prevalence of TB can be estimated using Bayesian techniques with the help of Markov chain Monte Carlo (MCMC) simulation. This model can also be used to predict TB prevalence.

Bayesian technique is applied by several research workers for modelling of tuberculosis prevalence. Wallinga *et al.* (2006) [10] conducted a study in which they used Bayesian techniques to analyze TB prevalence in Africa, taking into account both spatial and temporal patterns in the data. They discovered that TB prevalence was higher in urban areas and in countries with a high HIV prevalence. Achcar *et al.* (2008) [1] used Bayesian techniques to analyze the prevalence of tuberculosis cases in New York City from 1970 to 2000. The study used a counting process with two change points during the period and modeled the data using non-homogeneous Poisson processes in the presence of the two change points. Gelman *et al.* (2010) [4] used Bayesian techniques to model TB prevalence in Russia, taking into account both spatial and temporal patterns in the data. The authors found that TB prevalence was higher in certain regions of Russia and that it was increasing over time.

Methodology

Data description

For this study, the number of notified cases of tuberculosis in India from 1999 to 2019 were analyzed, using data from annual reports published by the Ministry of Health and Family Welfare. The data reveals three trends over a period of 21 years: an increase in cases from 1999 to 2011, likely due to the ineffectiveness of TB control programs and the HIV epidemic; a decrease in cases from 2012 to 2015, attributed to the success of programs such as the Revised National Tuberculosis Control Programme (RNTCP); and another increase in cases from 2016 to 2019, leading to the implementation of the National Strategic Plan (2017-2025) and the TB-free campaign at the Delhi End TB Summit in 2018. There are two significant change points in the data, in 2012 and 2016, with the highest number of TB cases recorded in 2019. The number of notified and cumulative cases per year can be found in Table 1.

To model the number of TB cases in India, the current study used a counting process, represented by the random variable {N(t), t ∈ [0, ∞)}, which counts the number of TB cases each year. A stratified sample of size 277050, representing 10% of the total number of TB cases, was used. In present study uniform distribution was used to assign the time in number of days for the occurrence of each case between 1999 and 2019, totaling T=7670 days.

Table 1: Year-wise number of notified and cumulative TB cases

Sr. No.	Year	Year-1999	Notified cases	Cumulative cases
1.	1999	0	133918	133918
2.	2000	1	240835	374753
3.	2001	2	468360	843113
4.	2002	3	619259	1462372
5.	2003	4	906638	2369010
6.	2004	5	1188545	3557555
7.	2005	6	1294550	4852105
8.	2006	7	1400340	6252445
9.	2007	8	1474605	7727050
10.	2008	9	1517363	9244413
11.	2009	10	1533309	10777722
12.	2010	11	1522147	12299869
13.	2011	12	1515872	13815741
14.	2012	13	1467585	15283326
15.	2013	14	1410880	16694206
16.	2014	15	1443942	18138148
17.	2015	16	1423181	19561329
18.	2016	17	1754957	21316286
19.	2017	18	1827959	23144245
20.	2018	19	2155894	25300139
21.	2019	20	2404815	27704954

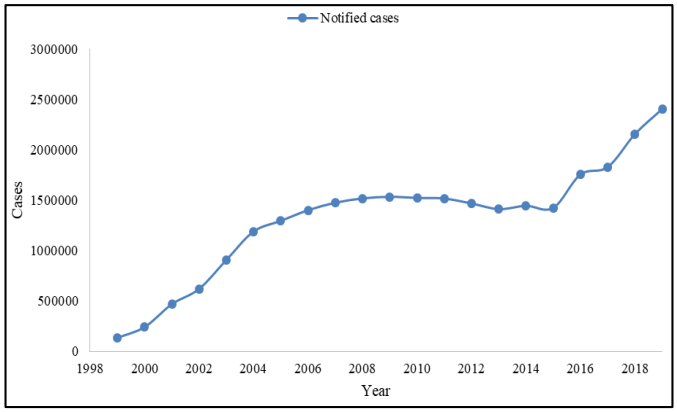

Fig 1: Number of Notified TB cases over the years

The number of notified and cumulative TB cases per year can be found in Table 1, and a graph of the progression of notified cases over the years can be found in Figure 1. The study utilized a Non-Homogeneous Poisson Process (NHPP), which allows for the average rate of TB cases to vary with time, in their analysis.

A Bayesian approach using Markov Chain Monte Carlo methods (Gelfand & Smith, 1990) [3] was employed to account for the presence of two change points in the data. This method has been utilized by other researchers to analyze homogeneous and non-homogeneous Poisson processes in the presence of change points (Raftery & Akman, 1986; Ruggeri & Sivaganesan, 2005) [7, 8].

The likelihood function

In this study, a power-law process with two change points was used to model the cumulative number of TB cases observed over time, represented by the function N(t). This process, known as a non-homogeneous Poisson process, assumes that the intensity function $\lambda(t) = dm(t)/dt = dE[N(t)]/dt$, where $m(t)$ is the mean value function (Cox & Lewis (1966)) and λ(t) follows a power law and can vary over time. The mean value function m(t) is used to describe this variability. The power-law process has been chosen for its ability to capture changes in the rate of TB case occurrence, which is important for understanding and predicting the spread of the disease. The intensity function for the overall process can be given as

$$\lambda\left(\frac{t}{\theta}\right) = \begin{cases} \lambda_1 = \frac{\beta_1}{\propto_1}\left(\frac{t}{\propto_1}\right)^{\beta_1 - 1} \text{ if } 0 < t < \zeta_1 \\ \lambda_2 = \frac{\beta_2}{\propto_2}\left(\frac{t}{\propto_2}\right)^{\beta_2 - 1} \text{if} \zeta_1 \le t < \zeta_2 \\ \lambda_3 = \frac{\beta_3}{\propto_3}\left(\frac{t}{\propto_3}\right)^{\beta_3 - 1} \text{if } t \ge \zeta_2 \end{cases} \quad (1)$$

Where $\theta = (\alpha_1, \alpha_2, \alpha_3, \beta_1, \beta_2, \beta_3, \zeta_1, \zeta_2)$

By letting $m_j(t) = m(t/\theta_j)$ the mean value function corresponding to the intensity function can be given as

$$m\left(\frac{t}{\theta}\right) = \begin{cases} m_1(\mathrm{t}) \text{ if } 0 < t < \zeta_1 \\ m_2(\mathrm{t}) - m_2(\zeta_1) + m_1(\zeta_1) \text{ if} \zeta_1 \le t < \zeta_2 \\ m_3(\mathrm{t}) - m_3(\zeta_2) + m_2(\zeta_2) - m_2(\zeta_1) + m_1(\zeta_1) \text{ if} \zeta_2 \le t < T \end{cases} \quad (2)$$

Where $m_1(\mathrm{t}) = \left(\frac{t}{\propto_1}\right)^{\beta_1}$, $m_2(\mathrm{t}) = \left(\frac{t}{\propto_2}\right)^{\beta_2}$ and $m_3(\mathrm{t}) = \left(\frac{t}{\propto_3}\right)^{\beta_3}$

The intensity function $\lambda_j(t)$ given in equation (1) can take on different forms. In the present study, a power-law process with two change points is used. This process is constant for $\beta_j = 1$, decreases for $\beta_j < 1$, and increases for $\beta_j > 1$, where $j = 1,2,3$. This type of relation between parametric forms is known as Weibull distributions (Kuo & Yang, 1996) [6].

In this study, data was collected on the number of notified cases of tuberculosis in India over a 21-year period, up until a total time T. The series of occurrences of cases was represented by t_i, with i ranging from 1 to n and the times falling in the sequence of $0 < t_1 < t_2 < \cdots < t_{N(\zeta_1)} < t_{N(\zeta_1)+1} < \cdots < t_{N(\zeta_2)} < t_{N(\zeta_2)+1} < \cdots < t_n < T, the$ likelihood function for the parameter vector θ, given the presence of two change-points ζ_1 and ζ_2, was represented by the equation (3).

$$L(\theta) = \prod_{i=1}^{N(\zeta_1)} \lambda_1(t_i) e^{-m_1(\zeta_1)} \times \prod_{i=N(\zeta_1)+1}^{N(\zeta_2)} \lambda_2(t_i) e^{-m_2(\zeta_2)+m_2(\zeta_1)} \times \prod_{i=N(\zeta_2)+1}^{N(T)} \lambda_3(t_i) e^{-m_3(T)+m_3(\zeta_2)} \quad (3)$$

The occurrence of tuberculosis can be modeled using a Poisson distribution, as it is a rare phenomenon. This is supported by the likelihood function shown in equation (3). Additionally, the sampling distribution for the intervals between occurrences, denoted by U_i, follows a density with a function of $f_{U/\theta}(t) = \lambda(t/\theta)\exp[-m(t/\theta)]$, $f_{U_2/U_1=s}(t) = \lambda(s+t/\theta)\exp[-m(s+t/\theta)+m(s/\theta)]$ and so on. The data for this occurrence, denoted by $D_T = \{n;\ t_1, \ldots, t_{N(\zeta_1)}, t_{N(\zeta_1)+1}, \ldots, t_{N(\zeta_2)}, t_{N(\zeta_2)+1}, \ldots, t_n, T\}$ includes the number of occurrences, n, the times of occurrence, $t_1, .., t\_n$, and the presence of two change points, ζ_1 and ζ_2.

Bayesian Analysis

In this analysis, the intensity function is described by equation (1) and is assumed to have two change points, which are normally distributed with known hyperparameters c_l and d_l (where $l = 1, 2$ and $\zeta_1 < \zeta_2$). The parameters α_j and β_j are uniformly distributed with known hyperparameters a_j, b_{1j}, and b_{2j} $(for\ j = 1,2,3)$. The values of b_{11} and b_{13} are assumed to be one in the intervals $0 < t < \zeta_1$ and $\zeta_1 \leq t < \zeta_2$, while in the interval $\zeta_2 \leq t < T$, the values of b_{21} and b_{23} are assumed to be 10. To get a decreasing function, the value of b_{12} is assumed to be zero and the value of b_{22} is assumed to be one. The parameter a_j is assumed to have a large value, and the prior independence among the parameters is also considered. The joint posterior distribution is given.

$$\prod\left(\frac{\theta}{D_T}\right) \propto \left(\frac{\beta_1}{\alpha_1}\right)^{N(\zeta_1)} \left(\frac{\beta_2}{\alpha_2}\right)^{N(\zeta_2)-N(\zeta_1)} \left(\frac{\beta_3}{\alpha_3}\right)^{N(T)-N(\zeta_2)} \times \left[\prod_{i=1}^{N(\zeta_1)} \left(\frac{t_i}{\alpha_1}\right)^{\beta_1-1}\right] \left[\prod_{i=N(\zeta_1)+1}^{N(\zeta_2)} \left(\frac{t_i}{\alpha_2}\right)^{\beta_2-1}\right] \left[\prod_{i=N(\zeta_2)+1}^{N(T)} \left(\left(\frac{t_i}{\alpha_3}\right)^{\beta_3-1}\right)\right] \times \exp\{-\left(\frac{\zeta_1}{\alpha_1}\right)^{\beta_1-1} - \left[\left(\frac{\zeta_2}{\alpha_2}\right)^{\beta_2} - \left(\frac{\zeta_1}{\alpha_2}\right)^{\beta_2}\right] - [\left(\frac{T}{\alpha_3}\right)^{\beta_3} - \left(\frac{\zeta_2}{\alpha_3}\right)^{\beta_3}]\} \quad (4)$$

Where $D_T = \{n;\ t_1, \ldots, t_n; T\}, 0 < \alpha_j < a_j, 0 < \beta_j < b_j, c_l < \zeta_1 < d_l, j = 1,2,3$ and $l = 1,2$.

Because the posterior distribution is complex, Markov Chain Monte Carlo (MCMC) methods were used to estimate its parameters. Specifically, the Gibbs sampling algorithm of the MCMC method was used. To use the Gibbs sampling algorithm, we need the full conditional posterior distributions, which are given.

$$\prod(\theta_j/\theta_{(j)}, D_T), j = 1,2, \ldots, K \text{ and } \theta_{(j)} = (\theta_1, \ldots, \theta_{j-1}, \theta_{j+1}, \ldots, \theta_K)$$

To perform this analysis, the WinBugs software was used. With this software, we can obtain the desired results by simply specifying the likelihood and prior distributions for the parameters.

Results and Discussion

In this study, the number of notified cases of tuberculosis in India between 1999 and 2019 was analyzed using data from the yearly report of TB in India released by the Central TB Division of the Ministry of Health and Family Welfare under the Revised National Tuberculosis Control Programme. This report contains information on various parameters related to tuberculosis, including the number of notified cases.

The parameters of the posterior distribution were assumed to follow a uniform distribution, with certain hyperparameter values assigned to specific intervals of time. The increasing function in the data was defined by the equation (1) and the hyperparameters $a_j = 100$ and $b_{11} = b_{13} = 1, b_{21} = b_{23} = 10$ were assigned to the intervals $0 < t < \zeta_1$ and $\zeta_1 \leq t < \zeta_2$, respectively. The decreasing function in the data was defined by the equation (1) and the hyperparameters $b_{12} = 0$ and $b_{22} = 1$ were assigned to the interval $\zeta_2 \leq t < T$. Additionally, prior distributions were assumed for the change points $(\zeta_1$ and $\zeta_2)$ with hyperparameters $c_1 = 3648, d_1 = 5102, c_2 = 5844,$ and $d_2 = 6572,$ corresponding to the estimated time periods of 2008-2013 and 2014-2018, respectively. It was assumed that the intervals of the prior distributions of the change points did not overlap.

The MCMC method was used to analyze the data. To ensure the reliability of the results, a burn-in sample was discarded initially. This allows the Markov chain time to reach the equilibrium distribution and avoids the possibility of over-sampling regions that have low probability under the equilibrium distribution due to the "bad" starting point of the chain. The Gibbs sampler was used to generate dependent samples, and to obtain independent samples, samples were taken at a constant interval using the Winbugs software. A burn-in sample of size 30000 was considered, and the Gibbs samples of size 100000 were simulated by selecting every 50th sample for each of the parameters to obtain approximately uncorrelated samples. To obtain the final sample of independent samples, every 50th sample was taken, resulting in a final sample of size 1400 for posterior estimation. The convergence of the Gibbs sampling algorithm was verified by analyzing plots of the simulated samples for each parameter to ensure that a stationary distribution was obtained in the final sample of 1400 simulated Gibbs samples.

In this study, estimates of the posterior median of various parameters $(\alpha_1, \alpha_2, \alpha_3, \beta_1, \beta_2, \beta_3, \zeta_1, and\ \zeta_2)$ were calculated and presented in a table 2. The mean value function was then obtained using these estimates and a line chart was created to display the pattern of observed and estimated TB cases over the years (figure 2). The observed and estimated TB cases are also presented in a table (table 3).

Table 2: Estimates of parameters of posterior distribution

Sr. No.	Parameter	Median	Standard deviation
1.	α_1	0.866	0.154
2.	α_2	159.230	17.820
3.	α_3	0.219	0.524
4.	β_1	0.892	0.025
5.	β_2	2.180	0.062
6.	β_3	0.785	0.111
7.	ζ_1	4748.000	92.300
8.	ζ_2	6939.00	5.289

$$m\left(\frac{t}{\theta}\right) = \begin{cases} \left(\frac{t}{0.8656}\right)^{0.8916} & \text{if } 0 < t < 4748 \\ \left(\frac{t}{159.23}\right)^{2.18} - \left(\frac{4748}{159.23}\right)^{2.18} + \left(\frac{4748}{0.8656}\right)^{0.8916} & \text{if } 4748 \le t < 6939 \\ \left(\frac{t}{0.2189}\right)^{0.7851} - \left(\frac{6939}{0.2189}\right)^{0.7851} + \left(\frac{6939}{159.23}\right)^{2.18} - \left(\frac{4748}{159.23}\right)^{2.18} + \left(\frac{4748}{0.8656}\right)^{2.18} & \text{if } 6939 \le t < 7670 \end{cases} \quad (5)$$

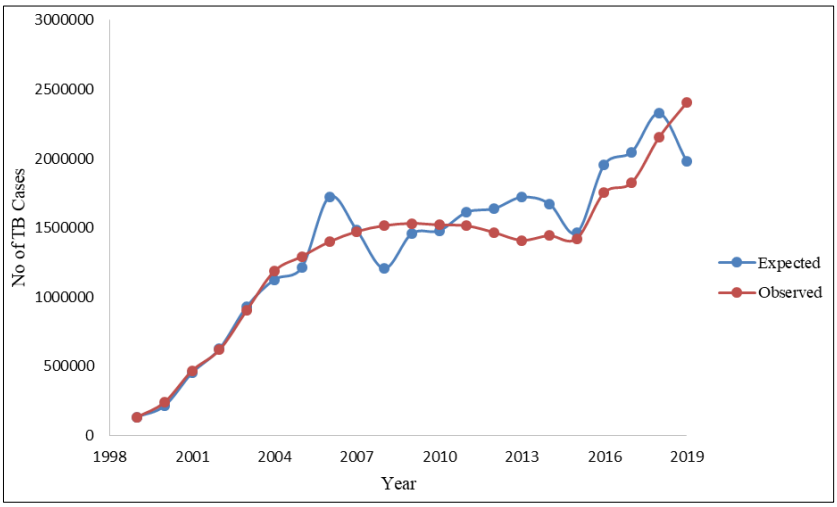

Fig 2: A chart showing number of observed and Expected number of TB cases over the years

Table 3: Number of observed and Expected TB cases over the years

Sr. No.	Year	Time	Observed	Cumulative (Obs.)	Expected	Cumulative (Exp.)
1	1999	365	133918	133918	135418	135418
2	2000	731	240835	374753	215366	350784
3	2001	1096	468360	843113	452714	803498
4	2002	1461	619259	1462372	626628	1430126
5	2003	1826	906638	2369010	933093	2363219
6	2004	2192	1188545	3557555	1127764	3490983
7	2005	2557	1294550	4852105	1215709	4706692
8	2006	2922	1400340	6252445	1724905	6431597
9	2007	3287	1474605	7727050	1483631	7915228
10	2008	3653	1517363	9244413	1208271	9123499
11	2009	4018	1533309	10777722	1459652	10583151
12	2010	4383	1522147	12299869	1480774	12063925
13	2011	4748	1515872	13815741	1614588	13678513
14	2012	5114	1467585	15283326	1640347	15318860
15	2013	5479	1410880	16694206	1722677	17041537
16	2014	5844	1443942	18138148	1675067	18716604
17	2015	6209	1423181	19561329	1465434	20182038
18	2016	6575	1754957	21316286	1956254	22138292
19	2017	6940	1827959	23144245	2048861	24187153
20	2018	7305	2155894	25300139	2327857	26515010
21	2019	7670	2404815	27704954	1982606	28497616

The data on notified cases of tuberculosis in India shows an increase in TB prevalence in the 2000s, possibly due to lack of awareness about the transmission and other factors contributing to the disease. After 2010, there appears to be two change points, possibly due to increased efforts by the government to control the spread of TB. However, the number of cases has continued to increase in recent years, with the highest number of cases reported in 2019. Data for the number of notified cases in 2020 and 2021 was not included in the analysis due to underreporting of TB cases due to the COVID-19 pandemic, as people may have hidden their other medical conditions, including TB, during this time. In addition, there is a positive correlation between the occurrence of TB and COVID-19 (Visca *et al.*, 2021) [9], which could further impact the number of notified cases. (India TB Report, 2021).

The proposed model was found to fit the data well, as demonstrated by the results of the chi-square test for goodness of fit. This study can be extended to other epidemiological data sets with more than two change points, and the use of MCMC methods makes it easier to estimate the parameters in the case of a non-homogeneous Poisson process with change points, a task that traditional inference techniques are not capable of. In this study, the intensity function was in the form of a power law process, but other parametric forms such as Gompertz growth or logistic growth could also be used.

References

1. Achcar JA, Martnez EZ, Rufino-Neto A, Paulino CD, Soares P. A Statistical Model Investigating the

Prevalence of Tuberculosis in New York Using Counting Processes with Two Change-Points. Epidemiology and Infection. 2008;136(12):1599-1605.
2. Cox DR, Lewis PA. Statistical Analysis of Series of Events. London: Methuen; c1966.
3. Gelfand AE, Smith AFM. Sampling based approaches to calculating Marginal Densities. Journal of the American Statistical Association. 1990;85:398-409.
4. Gelman A, King G, Niu X. A Bayesian model for tuberculosis prevalence in Russia. International Journal of Tuberculosis and Lung Disease. 2010;14(11):1456-1461.
5. India TB Report. New Delhi: Ministry of Health and Family Welfare; c2021.
6. Kuo L, Yang TY. Bayesian computation for non-homogeneous Poisson process in software reliability. Journal of the American Statistical Association. 1996;91(434):763-773.
7. Raftery AE, Akman VE. Bayesian analysis of a poison process with a change-point. Biometrika. 1986;73(1):85-89.
8. Ruggeri AE, Sivaganesan S. On modelling change-points in non-homogeneous Poisson processes. Statisitcal Inference for Stochastic Processes. 2005;8:311-329.
9. Visca D, Ong CWM, Tiberi S, Centis R, D'ambrosio L, Chen B, *et al*. Tuberculosis and COVID-19 interaction: A review of biological, clinical and public health effects. Pulmonology. 2021;27(2):151-165.
10. Wallinga J, Te Beest D, Agnandji S. A Bayesian spatial model for tuberculosis prevalence in Africa. Statistics in Medicine. 2006;25(16):2767-2782.
11. WHO. Global tuberculosis report 2021. Geneva: World Health Organization; c2021.