

<https://doi.org/10.48047/AFJBS.6.14.2024.10522-10556>



African Journal of Biological Sciences

Journal homepage: <http://www.afjbs.com>



Research Paper

Open Access

Analysis of Tuberculosis Cases in India: An Inferential Approach using Scan Statistics and Regression Modelling

riyanka Subramani, Jitendra Kumar , Kalpanapriya Dhakshnamoorthy

Department of Mathematics, School of Advanced Sciences, Vellore Institute of Technology,
Vellore - 632 014, Tamil Nadu, India.

Email: piyamath54@gmail.com, jitendra.kumar@vit.ac.in, dkalpanapriya@vit.ac.in

Volume 6 Issue 14, Aug 2024

Received: 15 June 2024

Accepted: 25 July 2024

Published: 29 Aug 2024

doi: [10.48047/AFJBS.6.14.2024.10522-10556](https://doi.org/10.48047/AFJBS.6.14.2024.10522-10556)

Abstract

Tuberculosis remains a longstanding and pervasive health concern globally, particularly in India, which accounts for 27% of all TB cases worldwide. This study aims to identify and characterize TB hotspots to inform policy development. By distinguishing between static and dynamic hotspots, policymakers can tailor interventions effectively. Understanding these hotspots can guide awareness programs, monitoring efforts, and resource allocation. Additionally, leveraging technologies and statistical methods like scan statistics and regression modeling, along with software tools such as saTScan, R, and Python, enables efficient analysis. Notably, around 50 districts in northwestern India emerge as significant TB hotspots. Addressing these hotspots is crucial for optimizing resource allocation nationwide. This study underscores the urgent need for targeted policies and interventions to combat TB effectively in India.

Keywords: Tuberculosis, Scan statistics, Hotspots, Log-likelihood ratio, Relative risk, p-value.

1. INTRODUCTION

Tuberculosis (TB) is a global health challenge caused by the bacterium *Mycobacterium tuberculosis*. Despite medical and technological advancements, TB continues to affect millions of people worldwide, posing a threat, particularly in

resource-limited countries like India, Kenya, Tanzania, Pakistan, Indonesia and in many more countries. The present study focused on details investigations of TB cases in India at exploratory, descriptive and inferential levels. Historically, TB has spanned centuries and left a significant imprint on human history, earning titles like the "White Plague". The diseases impact extends beyond medical realms, influencing literature, art, and music. Evidence of TB has been found in ancient Egyptian mummies, and its cultural significance is evident in the Romantic era. In 1882, Robert Koch's discovery of the tubercle bacillus marked a pivotal moment in understanding the disease. Epidemiologically, TB persists as a global health emergency, with around 10 million cases and 1.4 million fatalities in yr.2019, particularly affecting low and middle-income nations due to factors like poverty, overcrowding, malnutrition, and weakened immune systems. Drug-resistant TB poses a formidable challenge, necessitating innovative treatment strategies. Pathogenetically, mycobacterium tuberculosis primarily affects the lungs but can impact various organs. The bacterium's distinctive cell wall composition enables it to elude immune reactions and establish persistent infections, forming granulomas that serve as potential reservoirs for reactivation. Clinical manifestations of TB range from asymptomatic latent infection to active pulmonary and extrapulmonary diseases. Diagnosis is crucial for successful management and containment, with traditional methods like sputum smear microscopy and chest X-rays complemented by molecular techniques like PCR. Treatment strategies have evolved from mid-20th-century antibiotics to second-line drugs and more intricate protocols due to the rise of drug-resistant strains. WHO-endorsed DOTS programs aim to enhance treatment adherence. Preventive efforts involve vaccination (BCG), infection control measures, and addressing social determinants. BCG may not provide complete protection, but it safeguards against severe childhood TB. In healthcare settings, proper ventilation and respiratory hygiene are crucial for preventing TB transmission. Long-term control requires addressing root causes like poverty, malnutrition, and HIV

co-infection. In summary, TB remains a significant global health obstacle with a rich historical legacy. Collaborative efforts and continuous research are crucial to understanding and mitigating the impact of this age-old challenge. The study in India aims to identify TB hotspots to frame strategic policies for combating the disease by 2030.

Even while there has been some statistical work on TB, it has not been specifically applied to Indian scenario. As such the prime objective of the present article is

- To do exploratory analysis of TB cases in India at district level.
- To characterize TB cases using descriptive statistics.
- To compute the most likely cluster (MLC), popularly known as hotspot using the method of scan statistics with the tool saTScan. It enables us to incorporate the spatio-temporal nature of TB cases in India.
- To test whether the hotspot is manmade or due to chance.
- To conduct the test of statistical significant z-test for diagnosing whether there is a need to frame a separate policy formulation or we should adopt a single but comprehensive common policy formulation effectively to address the issue of static and dynamic hotspots.

The primary aim is to conduct an in-depth analysis of tuberculosis (TB) cases in India. The second goal involves characterizing these cases through both absolute and relative descriptive statistics. The third objective is to identify and assess the nature (static or dynamic) of potential hotspots. The fourth objective is to conduct a test to confirm whether these hotspots are statistically significant (due to human activities) or merely due to chance. Our approach is to implement the methods of scan statistics and their properties, using the saTScan software for computational efficacy. This is an approach to optimize the resource allocations for TB eradication, in align with federal govt and global programme implementors(UNO). Another goal is to investigate whether identified hotspots are static or dynamic. Subsequent objectives include developing a sustainable model for TB cases and to support in finding the practical solutions of the problem. We also intend to construct a model for estimating and predicting future TB cases in India. Finally, we aim to validate the model using various diagnostic measures, employing software such as saTScan, MS Solver, R, G Numeric spreadsheet, MS Excel, and Python. The present work is relevant to the society and related stakeholders, serving as a roadmap for future studies and guiding policy decisions. The organizational chart (flow diagram) is presented in Figure 1.

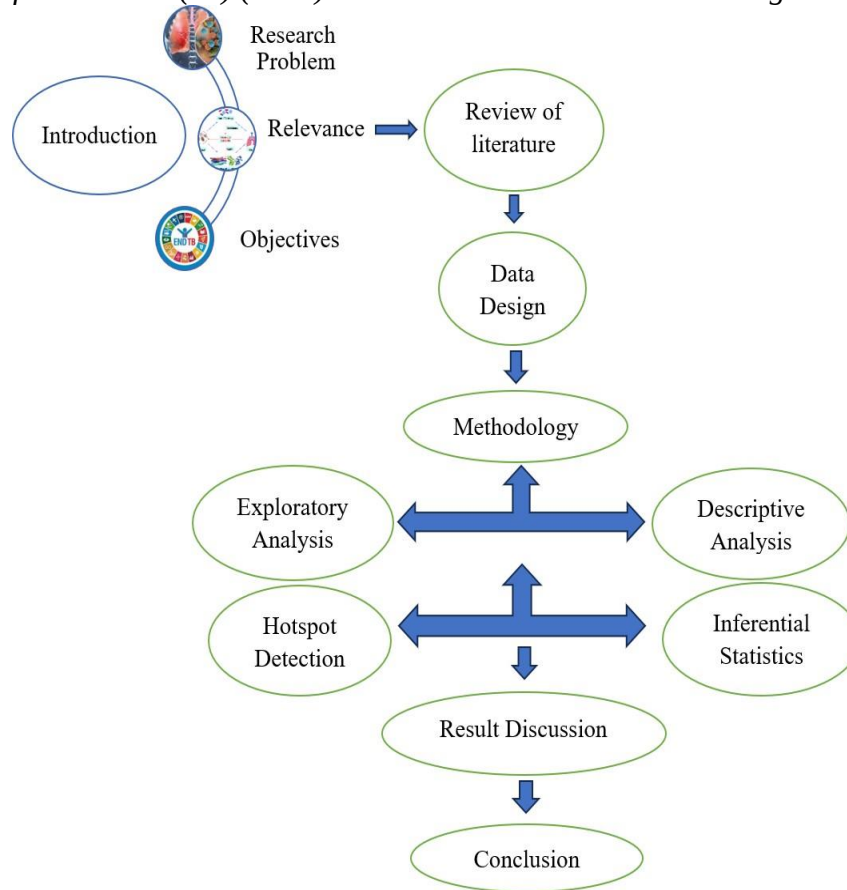


Figure 1: Workflow

2. LITERATURE REVIEW

Kulldorff [1], [15], [34] authored SaTScan, a free software programme for surveillance of any events that is a function of time, space or both. Finding clusters that the baseline process cannot explain is the key goal. These techniques are demonstrated using an epidemiological dataset, but they could also be used in other contexts. Kumar, J., and Anjali [5] stated the spatio-temporal aspect of suicide and suicidal ideation using satscan software to find hotspots in four major cities of Tamil Nadu, and it was found to be useful and relevant for control and monitoring bodies in organising and reducing such a precious loss of human life in the state of Tamil Nadu. Similar but not identical concepts were utilised in Bhatia, U. et al., [6] (2022)'s work on computer vision-based drowsiness image detection. In the Indian state of Haryana, the key hotspots were found by Sukhija and Kumar, J, et al., in the yr.2017 [8] did statistical analysis and determined the hotspot of rape cases in the state of Haryana. They have also conducted an investigation to compare how well different hotspot mapping approaches predict future spatial crime patterns. Reddy, G.V., et al., (2023) [9] work is relevant in prediction and analysis of the stock market prices and used in the study is inferential, it has been proven to be useful in the analysis of financial time series. In order to anticipate and analyze the stock market, time series modeling and forecasting are essential. Although, Kumar J. (2014) [7] has successfully implemented the method of scan statistics in finding out the hotspots various cognizable crimes in the NCT of Delhi, his study was more general in nature and scope. Yang L, et al., [12] approached an model which is ineffective for estimating the level of danger in high risk cluster locations. Time is a significant confounder that has been shown to have a direct impact on the identification of TB high-risk areas. Zulu [13] and Xia [14] had identified the high-risk locations for malaria, and HIV/AIDS the GIS application and scan

statistical approach may be used to detect the geographical, temporal, and spatiotemporal distribution of malaria. Meng-Yang Liu and many authors [16]-[33] significant spatial, temporal, and spatio-temporal clusters of reported TB cases at the prefecture level were found by their investigation. The findings of the spatio-temporal clustering revealed that SS(+) TB clustering time was concentrated before the yr.2010, but SS(-) TB clustering time was mostly concentrated after the yr.2010. This finding is valuable for prioritising resource allocation in high-risk times and locations and for developing effective TB preventive and control strategies. Cameille Frievant, et al., [35] provides three new spatial scan statistics for multivariate functional data that can be used to find abnormally high concentrations of air contaminants at extremely small scales while accounting for their associations. Their methodology is developed from a multivariate extension of the Wilcoxon test statistic, an adaption of the Hotelling T^2 -test statistic, and a functional multivariate analysis of variance. The methods were examined in a simulated environment before being used on the dataset for air pollution. Garrison et.al., [10] the French physician Gaspard-Laurent Bayle of Vernet identified the widespread "military" TB in the yr.1810. TB is a generalised illness that not only affects the lungs but also the other parts of the human body and can be clinically diagnosed by coughing and breathing difficulties, fever, purulent expectoration and breathing. Daniel TM [43] found the first effective treatment for tuberculosis was the development of the sanatorium cure, which was first described in the yr.1854 in Hermann Brehmer's doctoral dissertation, "Tuberculosis is a Curable Disease," by a student of botany who had TB and had recovered after visiting the Himalayan Mountains. Brehmer later established a facility in Gorbardsdorf, a mountain village surrounded by fir trees, to treat patients by providing them with constant access to fresh air and a healthy diet. Dye C [11] examined and explained that TB is still a significant public health issue, a multifaceted approach centred on enhancing medication therapy, diagnostic tools, and prevention is necessary. Kaandal, S., et al.,[36] used a scan statistic approach and identified the hotspots of cancer incidences and mortality in India. Anjali, et al., [42] presented the time series modelling approach on various factors responsible for suicidal hotspots using comprehensive approach in which she had attempted to prove the superiority of the multivariate VARMA model over ARIMA model. The World Health Organisation (WHO) has pledged to eradicate tuberculosis by the yr.2050, hence an effective plan is needed to achieve this. In India, efforts to combat tuberculosis have evolved over the years through various programs. Here are some of the key programs for tuberculosis eradication in India, along with their launch years: National Tuberculosis Control Program (NTCP) - Launched in the yr.1962, this was one of the earliest initiatives to control tuberculosis in India. Revised National Tuberculosis Control Program (RNTCP) - Launched in the yr.1997, this program replaced the NTCP and introduced the internationally recommended Directly Observed Treatment Short-course (DOTS) strategy. It aimed to provide free diagnosis and treatment for tuberculosis patients across the country. National Strategic Plan for Tuberculosis Elimination - Launched in the yr.2017, this plan set ambitious targets to eliminate tuberculosis in India by the yr.2025, five years ahead of the global target. It introduced innovative strategies and technologies to enhance tuberculosis control efforts. National Tuberculosis Elimination Program (NTEP) - in the yr.2018, the RNTCP was renamed the NTEP, reflecting the country's commitment to achieving tuberculosis elimination. These programs have been instrumental in India's ongoing efforts to combat tuberculosis and reduce its burden on public health.

3. DATA DESIGN

Data for this research has been sourced from reliable and officially available secondary channels, including government entities at various levels (central, state, regional, districts, quasi-government and semi-government) as well as numerous public and private websites. To ensure the accuracy and validity of the dataset, we deliberately opted against conducting

primary surveys for data acquisition. This decision was driven by the realising that, as researchers, we would have limited control over respondents, and the information shared might lack scientific validity. Simultaneously, our goal is to optimize both time and cost inputs. We have extensively utilized portals and reports such as Ni-Kshay, NHFS-5, annual reports on TB cases [2]-[4], and others as authentic sources. While we aim to tap into the digital richness of expansive information reservoirs, our focus remains on acquiring information that aligns with our computational needs to meet our research objectives. For computational efficiency, we have processed and structured a dataset tailored to our requirements. The data refinement process involved tasks like data cleaning, mining, and enhancing flexibility. Despite the laborious and extensive efforts involved in reshaping the data, we undertook normalization using methods such as z-score and box-plot as part of the exploratory and descriptive processing.

4. RESEARCH DESIGN

Primarily our research design is a hybrid of exploratory, descriptive and causal. We have attempted to execute spatial as well as temporal data sets for finding out the TB hotspot and coldspot. Needless to say, our research design enables us to derive the inferential conclusion and provide an easy way out to develop an effective policy framework to become successful in dealing with total eradication of TB cases before the yr.2030. The geographical unit selected for the present study is the districts of India.

5. METHODOLOGY

The present work adopted statistical methods analysed at three different levels briefly explained as exploratory, descriptive and inferential analysis.

5.1. Exploratory Analysis

The first level of analysis is exploratory which is aiming at presenting the surface attributes, at the stage of exploratory analysis we are mainly concerned with the general characteristics of the dataset in which we will do the basic statistical operations involving data processing like finding out the missing values, and arranging the data, developing a table of the information which is suitable and structured format, finding out an outliers, converting raw data into normalised data using z-score, visualizing the general trend of patterns to understand the monotonicity (increasing or decreasing). Also, at this stage we are concentrating at various data scale namely nominal, ordinal, interval and ratio and segregate the variable and parameters accordingly to plan the suitable measures for computation. The data logic set forth at this stage will provide the bed rocks upon which the descriptive and inferential analysis is built up. Our focus is to know and explore the surface level visualisation (general pattern). For that we are encouraged to draw various plots and charts. We have done with trend fitting in order to have an idea about the TB cases progress with years. Our geographical unit for research focus is district level across the country. We have a strong dataset consists of 624 districts of India almost cover entire states and union territories. During data processing we have selected an outlier district of TB cases in India of the yr.2020-2023 for purpose of effective data analysis.

5.2. Descriptive Analysis

The descriptive analysis is mainly concerned with specific characterisation using various absolute relative tools of descriptive statistics. For example, coefficient of variation (CV) is purely a relative tool whereas sample mean, sample variance etc are the absolute measures. So far as basic characteristics are concerned, we have used absolute measures for

the purpose of comparison, relational measures and model building we have deployed the relative tools. A fundamental field of study in statistics is descriptive statistics. It illustrates the link between variables in a particular sample and is frequently used to clean up and compile dispersed data, which is essential for conducting research and generating inferential statistical comparisons. In statistical modelling, descriptive statistics are frequently employed. The prior understanding of the probability experiment that produced the observed data is expressed by a statistical model.

5.2.1. Mean

One of the measures of central tendency is the mean, which is the arithmetic average of the total of values in the data divided by the numbers of observations shown in the equation 1.

$$\bar{X} = \frac{\sum_{i=1}^n (X_i)}{n}; i = 1, 2, 3 \dots n. \quad (1)$$

5.2.2. Variance and Standard Deviation

The variance and standard deviation of a distribution can be used to show how dispersed it is. In other words, variance and standard deviation show how closely each observed result resembles the mean.

The following equation 2 is used to determine the sample variance:

$$s_x^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}; i = 1, 2, 3 \dots n, \quad (2)$$

where n is the number of observations, $\bar{X}$ is the mean value of all observations, s_x^2 is the sample variance, and X_i is the value of one observation.

According to the following equation 3, the sample standard deviation is the square root of the sample variance and is defined below.

$$s_x = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}}; i = 1, 2, 3 \dots n. \quad (3)$$

It is important to note that s_x^2 and s_x are respectively unbiased estimator of population variance σ_x^2 and the population standard deviation σ_x . The statistics S_x^2 and S_x are respectively the biased estimator of population variance σ_x^2 and the population standard deviation σ_x and it is presented by the below equation.

$$S_x^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2; i = 1, 2, 3 \dots n,$$

$$S_x = \sqrt{\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2}; i = 1, 2, 3 \dots n.$$

5.2.3. Coefficient of Variation

It enumerates the features of the data. The coefficient of variation by equation 4 enables us to develop consistency ranking of TB cases for districts under the study during the years 202023. The district with the lowest C.V. values will be considered as more consistent and highly persistent in occurrences of TB cases.

$$\text{Coefficient of Variation} = \frac{\sigma}{\mu} * 100; \mu \neq 0. \quad (4)$$

5.2.4. Hotspot Analysis

The program SaTScan Kulldorff [1] was utilized to implement and calculate the estimations of several scan statistics parameters. For instance, we used a Monte Carlo simulation procedure to calculate the log-likelihood ratio value for each circle and acquire the p-values. In the current study, we examined and pinpointed the hotspots of TB cases dispersed throughout documented and recorded districts in India for the years 202023 using the discrete Poisson model. Let x stand for the group of zones that each circle contains then scan statistics has a mathematical definition as shown below

$$L(x) = \left(\frac{c_x}{\mu_x} \right)^{c_x} \left(\frac{c - c_x}{c - \mu_x} \right)^{c - c_x}, \quad \text{if } c_x > \mu_x, \quad (5)$$

where C is a total cases of tuberculosis,

C_x are the total cases in district x that occurred in district y,

μ_x are the expected cases in district x.

5.3. Inferential Statistics

The inferential analysis consists of drawing the inferentials out of the results obtained through various tools. At this stage mainly we have focussed upon estimation and testing. We estimate the parameters involved in the model and subsequently conducted the statistical testing for the purpose of validation of estimated values of the parameters. Here finding out the estimated value may include point as well as interval estimation will play an important role in our decision-making process. In statistical science the computation of confidence interval and predictive interval plays main role and we have attempted to execute the same.

Unbiased is a property associated with finite number of observations, n.

$$\hat{\Theta}_n = \hat{\Theta}_n(x_1, x_2, x_3, \dots, x_n).$$

A statistic is said to be an unbiased estimate of parameter Θ if $E(\hat{\Theta}_n) = \Theta$, we know that in sampling from a population with mean μ and variance σ^2 . $E(\bar{x}) = \mu$ and $E(S^2) = \sigma^2$ but $E(S^2) \neq \sigma^2$.

If $E(\hat{\Theta}_n) > \Theta$, $\hat{\Theta}_n$ is said to be positively biased and if $E(\hat{\Theta}_n) < \Theta$, it is said to be negatively biased. The amount of bias $b(\Theta)$ being given by $b(\Theta) = E(\hat{\Theta}_n) - \Theta$.

In consistency, if $\hat{\Theta}_n$ is an unbiased estimate of Θ_n with variance σ_n^2 and $\Theta_n \rightarrow \Theta$ and $\sigma_n \rightarrow 0$ as $n \rightarrow \infty$, then $(\hat{\Theta}_n)$ is a consistent estimate of Θ . More precisely,

$$E(\hat{\Theta}_n) \rightarrow \Theta,$$

$$V(\hat{\Theta}_n) \rightarrow 0, \text{ as } n \rightarrow \infty.$$

then $\hat{\Theta}_n$ is a consistent estimator of Θ .

Efficiency or Efficient Estimators, in statistical science there might exists multiple unbiased estimators then which one is most efficient is a question to be answered.

For example, in sampling from a normal population $N(\mu, \sigma^2)$, where σ^2 is known sample mean $\hat{x}$ is an unbiased and consistent estimator of μ . Also sample median (M_d) is an unbiased estimator too.

The question arises it should be proved with sample mean or sample median.

To handle such a scenario we precisely define the efficient estimator as follows:

$$E(M_d) = \mu, \\ V(M_d) \rightarrow 0, \text{ as } n \rightarrow \infty.$$

If the two consistent estimators say $\hat{\Theta}_1, \hat{\Theta}_2$ of a certain parameter Θ_n we have $V(\hat{\Theta}_1) < V(\hat{\Theta}_2)$, for all n, then $\hat{\Theta}_1$ is more efficient than $\hat{\Theta}_2$ for all sample size.

For example, $V(\hat{x}) = \frac{\sigma^2}{n}$; for all n.
and $V(M_d) = \frac{\pi^2 \sigma^2}{2n} = 1.57 \frac{\sigma^2}{n}$; for large n.

Since, $V(\hat{x}) < V(M_d)$. We conclude that for normal distribution sample mean is more efficient estimator for μ than the sample median for large n.

In most efficient estimator, if $\hat{\Theta}_1$ is the most efficient estimator with variance V_1 and $\hat{\Theta}_2$ is any other estimator with variance V_2 then the efficiency E of $\hat{\Theta}_2$ is defined as $E = \frac{V_1}{V_2}$, maximum value of E is 1.

$$E = \frac{V(\hat{x})}{V(M_d)} \\ = \frac{\frac{\sigma^2}{n}}{\frac{\pi^2 \sigma^2}{2n}} \\ = \frac{2}{\pi} \\ E = 0.637.$$

Mean is 63.7% more efficient than median in a normal distribution population.

When an estimator includes all of the data from the sample pertaining to a parameter, it is deemed adequate for that parameter. More specifically, the estimator $T=t(x_1, x_2, \dots, x_n)$ of a parameter Θ , is based on a sample of size n from the population with density $f(x, \Theta)$ such that the conditional distribution of $x_1, x_2, \dots, x_n$ are supplied. If T is unaffected by Θ , then T serves as an adequate estimate for Θ . Martin Kulldorff is the creator of two popular software programs: SaTScan, a free software application for disease monitoring in hospitals and other settings, and the widely used TreeScan data mining program [1]. The spatial-temporal scanning analysis software SaTScanTM version 10.0.2, developed by Information Management Services Inc., can be downloaded on the internet at <http://www.satscan.org>. The method is based on a scanning window with a moving column, where the base of the column represents a geographic area and the height of the column indicates time. The actual incidence and theoretical incidence inside and outside of the scanning window are utilized, respectively, to create the logarithmic likelihood ratio (LLR). Based on the overall incidence and the number of individuals in the window, the theoretical incidence for each scanning window may be computed.

Out of all the scanning Windows, the window with the biggest LLR is chosen since it has the strongest clustering. The additional windows with a statistically significant LLR value make up the secondary clusters in order to more precisely define the regional clustering distribution of the TB notification rate's features.

Regression modelling for TB cases at district levels: We have attempted to implement multiple linear regression model consists of average TB cases during the period 2020-23. The model is being represented by equation 6,

$$Y = \theta_0 + \theta_1 X_1 + \theta_2 X_2 + \theta_3 X_3 + \cdots + \theta_n X_n + \epsilon_i; \quad n = 1, 2, 3, \dots \quad (6)$$

k = number of independent regressors in the model, such that p = k+1; where p denotes the number of parameters in the model, ϵ_i is the error term. The response and Exploratory variables of TB cases are shown in the Table 1.

Table 1: Response and Exploratory Variables

Variables	Exploratory variables
y	Number of TB cases
x_1	Population Density
x_2	Temperature
x_3	Relative Humidity
x_4	Per Capita Income
x_5	Malnutrition
x_6	NO ₂ [Air Pollutants]
x_7	PM(2.5) [Air Pollutants]
x_8	O ₃ [Air Pollutants]
x_9	SO ₂ [Air Pollutants]
x_{10}	CO [Air Pollutants]

For estimating the parameter of the model illustrated above we have implemented the least square method (LSM). As per Gauss Markov's theorem, we will be able to estimate these θ 's known as BLUE (Best Linear Unbiased Estimator) that is likely to satisfy the criteria of unbiasedness, consistency, efficiency, and sufficiency as defined above.

For purpose of proper statistical analysis and drawing the inferentials we need to compute the following measures.

$$R^2 = \frac{SSR}{SST} = \frac{\text{Sum of squares due to Regressors}}{\text{Sum of squares due to total}},$$

$$R^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}, \quad (7)$$

where R^2 is known as coefficient of determination,

y_i is observed value,

$\hat{y}_i$ is fitted or predicted or estimated value for any given x_i ,

$\bar{y}$ is average of observed value y_i .

In order to address the issue of overfitting a model we can prefer the adjusted R^2 and it is defined by 8,

$$R^2_{adjusted} = 1 - \frac{(1 - R^2)(n - 1)}{(n - p)} = 1 - \frac{(1 - R^2)(n - 1)}{(n - k - 1)}, \quad \text{where } p = k + 1. \quad (8)$$

Empirically, $R^2_{adjusted} < R^2$.

The degree of overfitting(overloading) of the model will be reflected through the difference between the equation 7 (R^2) and the equation 8 ($R^2_{adjusted}$).

$$\Delta = R^2 - R^2_{adjusted} >$$

The role of error analysis play vital role in regression modelling especially to diagnose the problem of multicollinearity, autocorrelation and homoscedasticity of the model using different approach. Further more, error analysis enable us to find out the suitable solution of these problems.

$$Error = \epsilon_i = (y_i - \hat{y}_i).$$

we want $\epsilon \rightarrow 0$ for the best fitted model. The standardise error is computed from the errors for the purpose of assessment and generalisation of the model under the use.

$$Standardised\ Error = \frac{Error}{Standard\ Deviation\ of\ Error} = \frac{\epsilon_i}{SD\ of\ \epsilon_i}$$

$$SD\ of\ \epsilon_i = \sqrt{\frac{1}{n-1} \sum_{i=1}^n \epsilon_i^2}$$

The Akaike Information Criterion (AIC) is given by the equation 9 used to assess each model's relative superiority for a particular data set when comparing statistical models fitted by maximum likelihood (ML) to the same data, often for non-nested models, to statistical models fitted by other methods. As a result, the statistic penalizes for the number of predictors included in the model and considers model parsimony: It is obvious that the second term is a deviation and the first is a penalty for the number of parameters. The statistical model with the lowest AIC value is the most successful.

$$AIC = n \ln(SSE) - n \ln(n) + 2p. \quad (9)$$

Using the Akaike information criterion, A statistical model's comparative excellence for a particular set of data is evaluated. The AIC provides an unbiased method for identifying the most economical model among several rival models. When compared to the AIC values of rival models, the meaningless AIC values for a specific set of data become meaningful. The optimal (ideal) model for the given data set is the one with the lowest AIC value among rival models. The AIC penalizes the addition of parameters since it is used to choose a model with few parameters that fits the data well.

The Schwarz Bayesian information criterion (SBIC), shown in the equation 10 a distinct estimate of model fit for a given set of data across several non-nested model types, has the following formula:

$$SBIC = n \ln(SSE) - n \ln(n) + p \ln(n). \quad (10)$$

When additional variables are added to a model, model selection statistics like the residual sum of squares always go lower. This is known as overfitting, and it is solved by Mallows's C_p (11).

$$C_p = \frac{SSE_p}{\hat{\sigma}^2} + 2(p+1) - n. \quad (11)$$

5.4. Assessing the quality of prediction

The quality of prediction is pertinent for any scientific model and we have attempted to understand at comprehensively. R^2 of the regression is the measure of fit of regression to the sample data, it is not considered as adequate measure the regression equations ability to estimate $\mu(\frac{y}{x})$ (the conditional mean of y, given x) to predict new responses, the standard R^2 over estimates the quality of future (out of sample) prediction. We must have possible means of assessing prediction quality and the first one is the data splitting method. In this method the data set is partitioned into two parts (groups). One group of n_1 data is used to estimate (fit) possible model used for forecasting. The second group of n_2 data points called hold out data or validation sample.

For each candidate model, prediction errors, $(y_i - \hat{y}_i)$, are computed for all n_2 observations in the validation sample. In the measure of forecast accuracy based on forecast errors can be computed as follows.

MSD/MSE (Mean Square Deviation/Mean Square Error)

$$MSD = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n}. \quad (12)$$

MAD (Mean Absolute Deviation)

$$MAD = \frac{\sum_{i=1}^n |y_i - \hat{y}_i|}{n} \quad (13)$$

MAPE (Mean Absolute Percentage Error)

$$MAPE = \frac{\sum_{i=1}^n \frac{(|y_i - \hat{y}_i|)}{|y_i|}}{n} * 100. \quad (14)$$

The above 3 equations (12-14) are commonly computed measures. The decision rule (guiding principle) can be presented as models with smaller MSD, MAD, MAPE are better for prediction purpose. MAPE is a relatively measure and shows percentage deviations of forecast from the original datasets. Though, average relative measure prefers under estimated values and as such is somewhat biased, still considered relevant for an effective analysis. The advantage of using such a approach is that the models are tested on the data that are not used in the fitting (model estimation process). This approach provides an independent assessment of models predictive ability. We compute PRESS statistic by the equation 15 and the PRESS residuals by 16 for the prediction model. R^2_{adj} maintains its value at the target R^2 of the model, despite increasing the variable, while R^2_{PRESS} decreases as variable increases. From this we can summarise the characteristics of R^2_{PRESS} as

- R^2_{PRESS} is consistently less than R^2 .
- R^2_{PRESS} value falls as the number of extraneous predictors in the model rises, in contrast to R^2_{adj} , which keeps its value at the target R^2 with growing numbers of redundant predictors.
- When there are too many irrelevant predictors in the model, R^2_{PRESS} may become negative.
- When additional predictors are included in the model, the value of R^2_{PRESS} falls more quickly for datasets with lower R^2 values.

A second method of assessing prediction quality is to use PRESS statistic. PRESS stands for Prediction Sum of Squares and is defined as

$$PRESS = \sum_{i=1}^n (y_i - \hat{y}_{i-1})^2, \quad (15)$$

where $\hat{y}_{i-1}$ represent the prediction obtained from a model estimated with one of the sample observations deleted. If there are n-observations in the sample, then there are n-different prediction $\hat{y}_{i-1}$. The quantities $(y_i - \hat{y}_{i-1})$ often are called PRESS residuals because they are similar to the actual residuals $(y_i - \hat{y}_i)$. The prediction of sum of squares is also equal to the error sum of squares (SSE). We use the statistic,

$$R^2_{PRESS} = 1 - \frac{PRESS}{SST}. \quad (16)$$

Here, the guiding decision rule is larger values of R^2_{PRESS} (or smaller values of PRESS) suggest the models of greater predictive ability.

5.4.1. Prediction Interval versus Confidence Interval

The prediction interval is always wider than the confidence interval because of the added uncertainty involved in predicting and individual response [37]. There exists a mathematical relationship between different variances represented as s^2_p , s^2_e , s^2_m respectively as shown in below equation.

$$s^2_p = s^2_e + s^2_m,$$

The variance of the prediction for an individual value is equal to the variance from estimation the point on the regression line for $x = x_p$, $s^2_{\hat{y}_p}$ plus the estimate of the variation of the individual y values around the regression line s^2_e

$$s_p^2 = s_e^2 + \frac{1}{n} + \frac{(x_m - \bar{x})^2}{(n-1)s_x^2},$$

$$s_m^2 = s_e^2 + \frac{1}{n} + \frac{(x_m - \bar{x})^2}{(n-1)s_x^2},$$

$$s_e^2 = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n-p},$$

where p is the number of parameter in the model.

A $(1 - \alpha)100\%$ prediction interval for y_p is computed using the below equation.

$$(\hat{y}_p - (t_{\alpha/2, n-2})s_p, \hat{y}_p + (t_{\alpha/2, n-2})s_p),$$

A $(1 - \alpha)100\%$ confidence interval at $x = x_m$ is computed using

$$(\hat{y}_p - (t_{\alpha/2, n-2})s_m, \hat{y}_p + (t_{\alpha/2, n-2})s_m).$$

5.5. Assumptions of the Multiple Linear Regression model & related violations

- The expected value of the disturbances is zero, means $E(e_i) = 0$.
- The disturbances have constant variance σ^2 (Homoscedasticity).
- The disturbances are normally distributed (Normality).
- The disturbances are independent (Autocorrelation).
- The independance of the regressors (Multicollinearity).

5.5.1. Problems & Solutions related to violating the assumptions:

The Effects of violations of each of these assumptions on the least square estimates of the regression coefficients are examined and possible correction for violations are implemented.

1. **In Regression modelling** the plot of the standardized residuals versus the fitted values provides an overall picture, while the plot of the standardized residuals versus each explanatory variable may help identity any violations specifically related to an individual explanatory variable, $(x_1, x_2, \dots, x_n)$. The systematic pattern are observed in the residual plot suggests a violation of the linearity assumption. The residual plots are used to determine whether linearity assumptions had been violated.
2. **Test for lack of fit & linearity:** This test is also known as pure error lack-of-fit test, we decompose the error sum of squares into two parts: the pure component and lack-of-fit component. These components are used to construct F-statistic to test the hypothesis.

H_0 : The relationship is linear,

H_1 : The relationship is not linear.

If we fail to reject the H_0 , the linear regression model is appropriate. On the other hand, if H_0 is rejected the linear model does not fit the data well, and we have to think some other functions to get better fit.

Decision Rule To conduct the F-test we follow the rule,

Reject H_0 : if $F > F(\alpha; w - k - 1, n - w)$,

Do not reject: H_0 if $F \leq F(\alpha; w - k - 1, n - w)$,

where k is the number of explanatory variables, n is the sample size and $w = n + u + v$, where u is the total number of repeated values of the observations in their study and v is the number of observations that repeats.

The pure error lack of fit test require that there are replications in the dataset for atleast one level of x -variable. So, we may conclude that a curvilinear model may fit the data better than the linear model in case if we reject H_0 .

3. **Model validation using p-value:** The following decision rule can be applied to validate the model reject H_0 if $p\text{-value} < \alpha$, do not reject H_0 if $p\text{-value} \geq \alpha$, where α is the level of significance.

4. **Correction for violations of the linearity assumptions:** The following methods are used to correct the linearity problem.

- log transformation of the x -variable
- log transformation of both x & y variable
- Transform y to $\sqrt{y}$
- Transform both x and y to $\sqrt{x}$ and $\sqrt{y}$
- Reciprocal transformation of x -variable
- Polynomial regression

Since, we do not have parabolic curve, we decided to drop the implementation of polynomial regression transformation. We may conclude that a random scatter in the residual plots indicates that the correct model has been fit to the data. If we observe any systematic patterns in the residual plot that indicate the violation of linear and suitable corrections recommended.

5.5.2. Accessing the constant variable assumption: (Homoscedasticity)

We wish to have the constant variance σ_e^2 for the model. To examine we plot the residuals $\hat{e}_i$ versus the regressors (x_i) used in the model. The residuals should appear scatter randomly about the zero line with no differences in the amount of variations in the residuals regardless the value of x . In a residual plot, nonconstant variance is identified by a "cone shaped" pattern or presence of any systematic pattern in the residuals, such a plot indicates heteroscedasticity (nonconstant variance). When the disturbance or error variance are not constant then the use of least square method suffers from following drawbacks:

- The estimates of the regression coefficients are no longer minimum variance estimates.
- The estimates of the standard errors of the coefficients are biased.

5.5.3. Test for nonconstant variance: (Heteroscedasticity)

We will use J.Szroeter test to detect the nonconstant variance. We develop the following null hypothesis

$$H_0 : \text{Variance is constant, } H_1 : \text{Variance is not constant.}$$

Szroeter's test statistic is

$$Q = \frac{6n}{n^2 - 1}^{1/2} h - \frac{n+1}{2}, \quad (17)$$

where n is the sample size and h is given below

$$h = \frac{\sum_{i=1}^n i e_i^2}{\sum_{i=1}^n e_i^2} \quad (18)$$

where e_i is the residual from the i^{th} -observation in the regression equation.

Decision Rule Reject H_0 : if $Q > Z_\alpha$,

Do not reject H_0 : if $Q \leq Z_\alpha$,

where α is the level of significance for the test, Z_α is chosen from the standard normal table with upper tail area α . While conducting Szroeter's test we assume that all the observations can be arranged in order of increasing variance. This further implies that, we assume that the variance increases with the value increasing in the exploratory variable. Once we complete the sorting of x-values in ascending order we do adjust y-values accordingly and then we run the regression to get the residuals e_i . Now we'll compute h from the above equation 18, then we will substitute in the equation 17 to get Q.

Corrections for nonconstant variance to stabilize the variance: There are a number of possible corrections for nonconstant variable all these requires a transformation of the dependent variable following are the commonly used method.

- Transform y to log y with natural log
- Transform y to $\sqrt{y}$, the square root transformation is appropriate when the dependent variable is count variable that follows poisson distribution.
- Transform y to log y and x to log x, ie., $\log y \sim \log x$
- Transform y to $\sqrt{y}$ and x to $\sqrt{x}$, ie., $\sqrt{y} \sim \sqrt{x}$
- Box-cox transformations, that can be used to transform the y-variable to stabilize the variance. Box-cox transformation can be defined as follows apply y^p in place of y, where $0 \leq p \leq 1$. So this implies that powers of y-variable are used in the place of original variable.

When $p=0$, we define the transformation to be natural log transformation and obtain maximum compression of the data when $p=1$, no compression of the data are necessary. Power between 0 & 1 provide flexibility in the amount of compression obtained [38].

5.5.4. Test for Normality:

The straightness of the line in the normal probability plot can be measured by computing a variety of test statistics such as Kolmogorov-Smirnov(KS) statistic, the Anderson-Darling(AD) statistic, the Ryan-Joiner(RJ) statistic. Computing H_0 & H_1

H_0 : Disturbances are normal, H_1 : Disturbances are nonnormal.

for testing the statistic we use computed p-value as our test statistic.

Decision Rule Reject H_0 : if p-value $< \alpha$, Do not reject H_0 : if p-value $\geq \alpha$, where α is the level of significance of test.

Corrections for nonnormality: There are a variety of possible corrections, one popular correction is to transform the dependent variable using box-cox transformation, as we have done for nonconstant variance. Again, we'll use y^p in place of y, $0 \leq p \leq 1$.

5.5.5. Assessing the assumption that the disturbance are independent (Problem of Autocorrelation):

1. **Autocorrelation:** For a time series data, the independence of the disturbance, e_i generally violated. It happens because an event in one time period may influence, the event in the next time period. The relationship is shown by following equation

$$e_i = \rho e_{i-1} + u_i,$$

where e_i is the disturbance in the i^{th} time period, e_{i-1} is the disturbance in the previous time period, ρ is called the serial correlation coefficient or autocorrelation coefficient, u_i is the new disturbances generated as shown in the following models

$$y_i = \beta_0 + \beta_1 x_i + e_i,$$

presence of autocorrelation forces the estimated standard errors of the coefficient to become biased but it does not affect least square estimates. Autocorrelation coefficient ρ determines the strength of the relationship between disturbances in a successive time period and it varies $-1 \leq \rho \leq 1$. For $\rho = 1$, we have very strong relationship, undesirable situation, $\rho = 0$, indicate truly independence, a desirable condition to validate the model assumption of autocorrelation. $\rho > 0$ indicates positive autocorrelation, and $\rho < 0$ indicates negative autocorrelation.

2. **Test for first order autocorrelation:** Durbin-Watson test, computing the H_0 and H_1 hypothesis,

$$H_0 : \rho = 0, H_1 : \rho > 0.$$

where ρ is the first order autocorrelation.

Durbin-Watson test statistic is given by the equation

$$d = \frac{\sum_{i=2}^n (\hat{e}_i - \hat{e}_{i-1})^2}{\sum_{i=1}^n \hat{e}_i^2}. \quad (19)$$

Decision Rule Reject H_0 , if $d < d_L(\alpha; n, k)$, do not reject H_0 , if $d > d_U(\alpha; n, k)$, where $d_L(\alpha; n, k)$ and $d_U(\alpha; n, k)$ are the critical values, which can be found in statistical table known as Durbin-Watson table. The critical value depends upon the level of significance of the test α , the sample size n and the number of regressors k .

Now the test remain inconclusive if $d_L \leq d \leq d_U$. If the test 'd' falls in this region ($d_L \leq d \leq d_U$) then there is a need of correction of autocorrelation as such Durbin Watson test fails to arrive at final decision. If suppose 'd' falls in the inconclusive region then we recommend H test for autocorrelation as explained below.

3. **H test for autocorrelation:** Computing the H_0 and H_1 hypothesis,

$$H_0 : \rho = 0, H_1 : \rho > 0.$$

The test statistics is given below

$$h = r \frac{n}{1 - ns_{b1}^2}^{1/2}, \quad (20)$$

where r is the estimate of the first order autocorrelation, n is the sample size and s_{b1}^2 is estimated variance of the regression coefficient of the lagged dependent variable (y_{i-1}).

Decision Rule Reject H_0 if $h > Z_\alpha$, do not reject H_0 if $h \leq Z_\alpha$. It is important to make a note that 'h' from the equation 20 cannot be computed if $(1 - ns_{b1}^2) < 0$ [39]. Here, we compute the value of r by the equation 21 statistic as

$$r = 1 - \frac{d}{2}, \quad (21)$$

where d is the Durbin Watson test statistic (19).

4. Multicollinearity & related assessment in regression modelling: When explanatory variables (regressors) are correlated with one another, the problem of multicollinearity arising. Seriousness of the problem depends upon the degree of multicollinearity. We may have either serious multicollinearity that need to be assessed before final implementations of the model. Furthermore, we need to address two possible problems as mentioned below:

- The standard deviation of the regression coefficient are disproportionately large.
- There exist unstable estimates of the coefficients. Reliable estimates of the regression coefficient are difficult to obtain due to high standard errors.

(a) Methods to detect Multicollinearity:

- Pairwise Correlation Method
- Relatively larger value of F-statistic and smaller value of t-statistic indicate possible multicollinearity at disturbing state.
- Variance Inflation Factor (VIF's): Let $x_1, x_2, x_3, \dots, x_k$ be the k-explanatory variables in a regression. We perform the regression of x_j on the remaining (k-1) explanatory variables and call the coefficients of determination from the regression R_j^2 . The VIF for the variable x_j is given by the equation 22,

$$VIF_j = \frac{1}{1 - R_j^2}, \quad (j = 1, 2, \dots, k). \quad (22)_R$$

This is a measure of the strength of the relationship between each explanatory variables and all other explanatory variables in the regression. The value R_j^2 measures the strength of the relationship between x_j and the other k-1 explanatory variables. If $R_j^2 = 0$, there is no relationship among the regressors then $VIF_j = \frac{1}{1-0} = 1$. If R_j^2 increases, then VIF_j is also increases. Now, large value of VIF_j suggest serious problem of multicollinearity. A standard norm is if VIF value > 10 indicates high multicollinearity that will adversely affect the least square estimators and will be harmful to the model. Another popular criteria is if average of VIF_j is

$$\overline{VIF} = \frac{\sum_{j=1}^k VIF_j}{k} > 1,$$

then serious problem may exist. $\overline{VIF}$ indicates how many times larger the error sum of squares for the regression is due to multicollinearity than it is if the variables are uncorrelated. Freund & Wilson [40] represented that whenever VIF's are $< \frac{1}{1-R^2}$ where R^2 is the coefficient of determination for the model with all x variables included multicollinearity is not strong enough to affect the coefficient estimates.

- (b) **Corrections for Multicollinearity:** There exist multiple ideas, one of the simple way is to remove those variables from the model that are highly correlated with others, this will improve the situation. But such a method is not recommended in general. Several other statistical methods have been proposed as possible remedial measures with multicollinearity [41]. For example: principal components regression and ridge regression will be the preferred choice.

6. RESULT AND DISCUSSION

6.1. Exploratory Results:

The distribution of TB cases is not homogeneous across the country. Some districts witnessed higher density of occurrences of TB cases while other districts just having relatively fewer cases. Though, we have 806 districts in India

including the newly created districts in recent years. Proper, sufficient and authentic data of TB cases for many new districts are not available. As such we prefer to concentrate on data set retrieved from 624 districts of India that constitute our sample size(n) for computing the hotspots using saTScan. For purpose of exploratory and basic descriptive analysis, we have computed summary statistics presented in Table 2, where we have attempted to summarize the elementary statistics needed for subsequent analysis. It can be inferred from Table 2 that the mean number of cases during the yr.2020-23 varies from 2810 to 3167, with a dip value of 2810 for the yr.2021. The standard error of TB cases varies from 118 to 139 again with a dip value of 118 for the yr.2021. The median varies from lowest value of 2005 during the yr.2021 to the highest value of 2150 during the yr.2023. Though, mode is not preferred measure like mean and median but it plays vital role sometimes during the course of analysis. We observe that lowest mode value is 939 in 2021, highest mode value 3452 in 2022, surprisingly the yr.2020 & 2023 witnessed a moderate value of 1479 & 1567 respectively. It appears that TB cases are more prominent in the yr 2022 as compared to the yr.2020, 2021 & 2023. We found the big contrast between the yr.2021 & 2022. The mode value is almost (3.67) times of the cases in the yr.2021. The standard deviation is an indispensable measure of dispersion that contributes high for inferential statistics. The lowest SD value is observed in the year 2021 as 2951 whereas the highest SD value is 3481 in the year 2023, that exposes the vulnerability of TB cases lowest in 2021, moderate in the yr.2020 & 2022, but highest in the yr.2023.

Table 2: Exploratory and descriptive statistics for TB cases of the year 2020-23.

Descriptive Statistics \ Year	2020	2021	2022	2023
Mean	3083.05609	2810.80609	3193.091346	3166.700321
Standard Error	137.9916556	118.1591191	129.4622652	139.3625457
Median	2171	2005.5	2251	2149.5
Mode	1479	939	3452	1567
Standard Deviation	3447.030453	2951.613848	3233.966348	3481.275275
Sample Variance	11882018.95	8712024.307	10458538.34	12119277.54
Kurtosis	11.81630016	9.297082605	8.103540866	11.33323864
Skewness	2.918702776	2.596135506	2.414887593	2.843293604
Range	25829	21260	22771	26496
Minimum (Smallest)	10	12	9	5
Maximum (Largest)	25839	21272	22780	26501
Sum	1923827	1753943	1992489	1976021
Count	624	624	624	624
Confidence Level (95.0%)	270.9851274	232.0384068	254.2352888	273.6772526

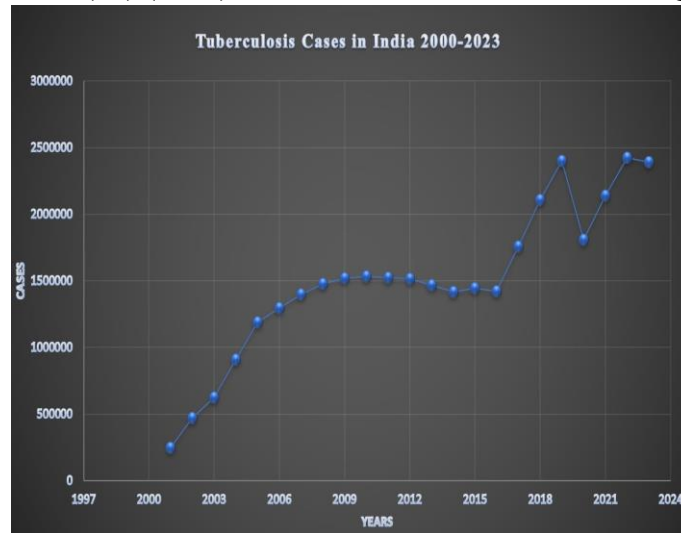


Figure 2: Trendline of Tuberculosis cases in India

We observe that the value of kurtosis is very high (>3) during the period of study which indicates leptokurtic curve, suggesting high level of peakedness of TB cases (clustering) distributed through the yr.2020-23 which further suggests proper investigation to find the reasons why such heterogeneity exists. This further suggests the probability of existence of hotspot and coldspot of TB cases in India. We are motivated to compute the hotspot using the method of scan statistics. The implication of skewness varies between 2 & 3 is positively skewed distribution of TB cases across the country. It enable us to search for suitable distribution as such we are motivated to apply negative binomial distribution to fit the model. For the purpose of hotpost identification & analysis the suitable model appears to be the Poisson model. The nature of our dataset is appears to be spatio-temporal. The least skewed value is 2.41 found for the yr.2022 as compared to remaining three years. The sample variance is varies from 87 lakh in the yr.2021 to 1.21 crore in the yr.2023 refelecting towards serious spread of TB cases in India during the yr.2020-23. Also, it is being observed that there are many districts where the cases are very low (5-12) to the other hand there are multiple districts where the highest count varies between 21,272 to 26,501 during the time period of study. Finally, from Table ?? we observe confidence level at LOS $\alpha=0.05$ varies from lowest value of 232 in the yr.2021 to the highest 273 in the yr.2023 reflects relatively higher concentration of diseases.

Figure 2 presents a trend line visualization plot during the yr.2000-23, enable us to have rough idea of growth pattern of TB cases.

6.2. Descriptive Results:

The decision making process for the dataset follows the hexagonal radial cycle which is shown in Figure 3. The role of description is the foremost requirement for any inferential approach as such we have computed the coefficient of variations and develop respective rankings to understand the role of each data elements at district level. The detailed computation has been presented in Table 3. We develop consistency ranking for all 624 districts out of which we concentrate on top 50 most consistent districts as shown in the Table 3. These districts provide us the basis for developing a model as well as give us an insight to make a comparison with hotspot districts identified and computed using saTScan statistics. For having an idea of outliers we have computed z-score for each districts and found year wise outliers districts as shown in Table 4.



Figure 3: Stepwise procedure in a decision process

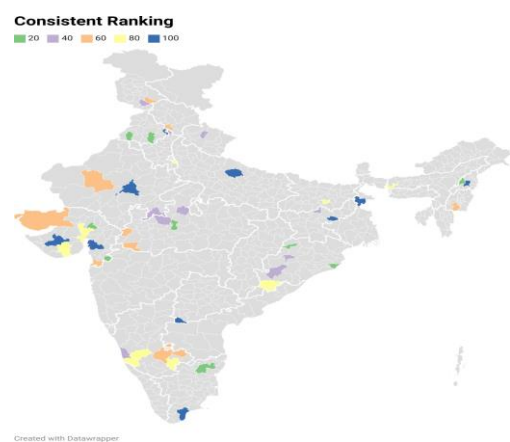


Figure 4: Geographical distributions of most consistent TB infected districts of India.

Table 3 represents the computations for consistency ranking and presents 50 most consistent districts, the corresponding visual maps has been shown in Figure 4.

Table 3: Consistency Ranking of top 50 TB infected districts during 2020-23.

Districts	2020	2021	2022	2023	Average	SD	CV	Consistency Rank
Gandhinagar	3424	3413	3401	3434	3418	14.21267	0.415818	1
Sangrur	1917	1948	1884	1928	1919.25	26.7753	1.395092	2
The Dangs	253	249	252	241	248.75	5.439056	2.186555	3
Tiruvannamalai	1821	1741	1807	1749	1779.5	40.37739	2.26903	4
Jagatsinghapur	370	354	378	363	366.25	10.21029	2.787792	5
Bhopal	11420	10808	10980	11472	11170	327.0311	2.927763	6
Muktsar	1254	1283	1305	1213	1263.75	39.76074	3.146251	7
Jharsuguda	650	656	661	697	666	21.15026	3.175714	8
Wokha	113	120	112	117	115.5	3.696846	3.200732	9
Malda	5780	5391	5703	5785	5664.75	186.3194	3.289102	10
Bharuch	2891	2886	3013	2780	2892.5	95.2558	3.2932	11
Jogulamba Gadwal	1415	1449	1414	1335	1403.25	48.32098	3.443504	12
Rajkot	4897	4601	4782	5010	4822.5	174.5556	3.619609	13
Ajmer	8626	8278	8382	7895	8295.25	304.0936	3.665876	14
Mohali	2206	2058	2186	2241	2172.75	79.80549	3.673018	15
Sitapur	12936	13355	13365	14129	13446.25	497.1354	3.697205	16
Dhanbad	3752	3557	3509	3439	3564.25	134.2172	3.76565	17
Thoothukudi	2085	1984	2185	2104	2089.5	82.62566	3.954327	18
Zunheboto	111	101	105	104	105.25	4.193249	3.984084	19
Dhubri	2061	2130	2248	2057	2124	89.20015	4.19963	20
Koraput	2707	2541	2529	2448	2556.25	108.6566	4.250624	21
Amreli	1579	1519	1625	1475	1549.5	65.95706	4.256667	22
Ahmadabad	4538	4219	4253	4106	4279	183.7444	4.294096	23
Lakhisarai	666	737	721	705	707.25	30.44531	4.304745	24
Dakshina Kannada	2967	2738	2969	2766	2860	125.2331	4.37878	25
Ramanagara	840	823	791	760	803.5	35.40716	4.406616	26
Faridabad	7448	7092	7890	7613	7510.75	333.454	4.43969	27
Bangalore	15104	13567	14689	14578	14484.5	652.2139	4.50284	28
Chikmagalur	748	740	761	685	733.5	33.47138	4.563242	29
Churachandpur	296	296	285	268	286.25	13.22561	4.620299	30
Jhabua	2443	2512	2645	2704	2576	119.6244	4.643805	31
Doda	907	970	1015	952	961	44.69899	4.6513	32
Tumkur	2575	2703	2885	2729	2723	127.2583	4.673458	33
Valsad	2850	2650	2972	2830	2825.5	132.7692	4.698964	34

Districts	2020	2021	2022	2023	Average	SD	CV	Rank
Solan	2173	2027	2096	1946	2060.5	96.86589	4.701086	35
Chikkaballapur	1374	1418	1537	1449	1444.5	68.91783	4.771051	36
Jodhpur	7665	7919	8574	7975	8033.25	384.909	4.791448	37
Kachchh	3376	3553	3452	3169	3387.5	162.7114	4.803289	38
Barwani	2393	2342	2550	2280	2391.25	115.4798	4.829265	39
Panchkula	1884	1938	2109	1993	1981	96.23929	4.858116	40
Kalahandi	1566	1571	1735	1646	1629.5	79.2822	4.865431	41
Sonapur	376	370	361	336	360.75	17.61391	4.882582	42
Kodarma	827	822	840	912	850.25	41.85989	4.923244	43
Ashoknagar	2489	2388	2627	2657	2540.25	125.1169	4.925379	44
Rudraprayag	309	299	320	284	303	15.29706	5.048534	45
Rajgarh	2028	2068	2251	2213	2140	108.594	5.074488	46
Udhampur	1019	1085	1151	1059	1078.5	55.43465	5.139976	47
Jhalawar	3172	3177	3441	3049	3209.75	165.1411	5.144983	48
Kolkata	1647	1542	1454	1520	1540.75	80.09734	5.198594	49
Udupi	1420	1325	1437	1291	1368.25	71.28055	5.209614	50

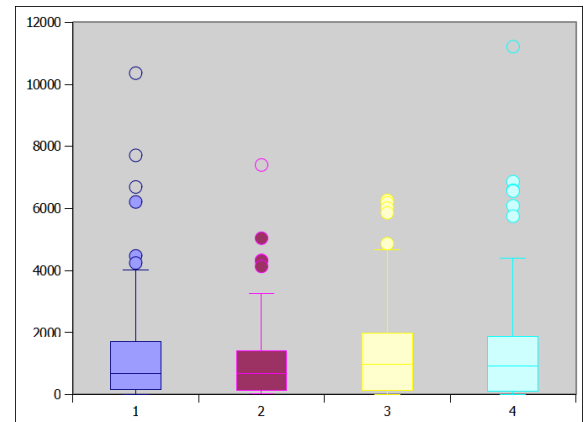
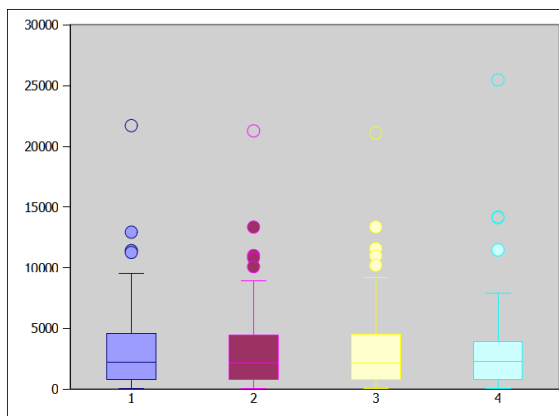


Figure 5: Boxplot for Hotspot and Coldspot of TB cases

The boxplot is the most powerful visual graph for representing five summary statistics in one go, namely Quartile 1 (Q_1), Quartile 2 (Q_2), Quartile 3 (Q_3), Inter Quartile Range (IQR), Range. Figure 5 presents the boxplot for understanding the outliers, the representation of mean, median and whiskers for hotspot as well as coldspot. Also, outliers of TB cases is being visualised clearly in Figure 5 one for hotspot and another for coldspot.

Table 4: Outlier districts for Tuberculosis cases in India

Districts\Year	2020	2021	2022	2023
Outlier Districts for Tuberculosis cases in India	Patna	Patna	Patna	Patna
	North West Delhi	North West Delhi	North West Delhi	Surat
	South Delhi	South Delhi	South Delhi	Bangalore
	Ahmadabad	Ahmadabad	Ahmadabad	Chhatarpur
	Surat	Surat	Surat	Jaipur
	Bangalore	Bangalore	Bangalore	Hyderabad
	Chhatarpur	Chhatarpur	Chhatarpur	Aligarh
	Ludhiana	Ludhiana	Bhopal	Ghaziabad
	Jaipur	Jaipur	Gwalior	Bareilly
	Hyderabad	Hyderabad	Ludhiana	Gorakhpur
	Aligarh	Agra	Jaipur	Kanpur Nagar
	Bareilly	Aligarh	Hyderabad	Lucknow
	Kanpur Nagar	Bareilly	Agra	Mathura
	Lucknow	Kanpur Nagar	Aligarh	Meerut
	Prayagraj	Lucknow	Bareilly	Sitapur
		Prayagraj	Ghaziabad	Varanasi
		Ghaziabad	Gorakhpur	
		Mathura	Kanpur Nagar	
		Meerut	Mathura	
		Moradabad	Meerut	
		Sitapur	Moradabad	
		Varanasi	Prayagraj	
			Saharanpur	
			Sitapur	
			Varanasi	

6.2.1. Scan Statistics and resulting Hotspot Analysis:

The present work made an attempt to find out the hotspot and coldspot based on scan statistics methodology. The computational equation 5 represents this is used for the purpose of hotspot calculation using the software saTScan. The software enabled primary hotspot of TB during the yr.2023 has shown in Figure 6, due to the space complexity we are restricted to present the visual figure for the yr.2023 only, the same we have computed for all the remaining years that could be shown further. Table 6 is the inferential statistics for hotspot characterisation of TB cases which presents the computed values of Relative Risk (RR), LogLikelihood Ratio (LLR), p-value for each of the three clusters namely primary hotspot (most likely cluster), secondary hotspot and tertiary hotpsot. We have deliberately discarded the corresponding statistical values for remaining clusters (quaternary, quinary etc.,) as we do not found them statistically significant.

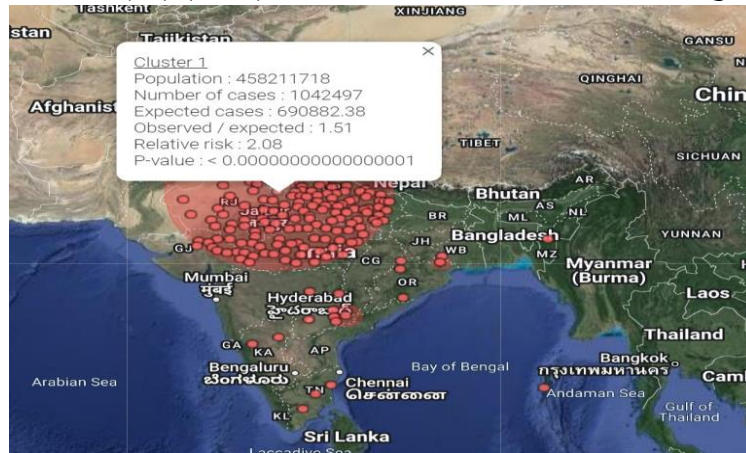


Figure 6: Primary Hotspot of Tuberculosis during 2023

Figure 7, 8, 9 presents the visualization of computed hotspots on various criteria namely most consistent district, hotspot based on count, and hotspot based on rates. All three computations are for 50 most consistent districts only. We declare the hotspot either static or dynamic after investigating the fact that whether the hotspot is changing every year or remains stagnant at some geographical location all through the period of years. Now, we are motivated to conduct the hypothesis testing whether we need to develop a separate strategy, policies and planning to tackle TB cases or should we go for single integrated approach. We have implemented the z-test, the computed z-value is 5.3708, critical z value is 1.6448 at $\alpha = 0.05$, where α is the level of significance. Since computed z value is greater than the critical value from the statistical table. Therefore we reject the null hypothesis, which implies that there is a significant difference between the two types of hotspot and subsequently we recommend a separate policy framework for each type of TB hotspot. Table 5 presents the detailed computed z-test for TB cases. Such a result is of great interest because without scientific evidences, the policy makers do not have any idea how to tackle the two types of TB cases effectively. Definitely, the present work will help us in optimization of resources, time, cost, technology, medicine distribution, etc and enable us to achieve the goal of TB eradication completely before the yr.2030.

Table 5: Output of z-test:Two Sample for Means

Two sample for means	Variable 1	Variable 2
Mean	6322.5	3541.928571
Known Variance	2744351	2685172
Observations	36	14
Hypothesized Mean Difference	0	
z	5.370841985	
P(Z<=z) one-tail	3.91849E-08	
z Critical one-tail	1.644853627	
P(Z<=z) two-tail	7.83699E-08	
z Critical two-tail	1.959963985	

Table 6: Satscan Output summary of Inferential statistics for Hotspot characterization of Tuberculosis cases during 2020-23.

Year / Clusters / Inferential Statistics		Relative Risk	Log Likelihood Ratio	P-value
2020	Primary	2.07	118128.1053	<0.000000000000000001
	Secondary	2.94	19346.31647	<0.000000000000000001
	Tertiary	3.34	12980.5997	<0.000000000000000001
2021	Primary	1.98	94967.70209	<0.000000000000000001
	Secondary	5.98	15607.57513	<0.000000000000000001
	Tertiary	2.64	13334.10312	<0.000000000000000001
2022	Primary	1.91	102306.0828	<0.000000000000000001
	Secondary	2.17	27124.7451	<0.000000000000000001
	Tertiary	1.91	16448.904	<0.000000000000000001
2023	Primary	2.08	130482.1453	<0.000000000000000001
	Secondary	2.47	39760.81017	<0.000000000000000001
	Tertiary	2.09	23735.32632	<0.000000000000000001

We may observe from Table 6 that for each of the year under study has the value of relative risk for hotspot districts (primary clusters) are 2.07, 1.98, 1.91 and 2.08 respectively which are smaller than the relative risk value of secondary clusters. From the same table we observe that the $LLR_P > LLR_S$ and exactly the same pattern is observed during the yr.2020-2023. Thus, with the help of Table 6 we summarize that for primary hotspots we have the least value of RR (relative risk) as compared with secondary and tertiary for each year of study, whereas we have relatively larger value of LLR (loglikelihood ratio) is formed in primary clusters as compared to secondary clusters. Figure 7, 8, 9 presents the visualization of hotspots of TB cases during the yr.2020-2023 based on three different criteria namely Hotspot by saTScan, Hotspot on the basis of cases, Hotspot on the basis of rates of TB cases respectively.

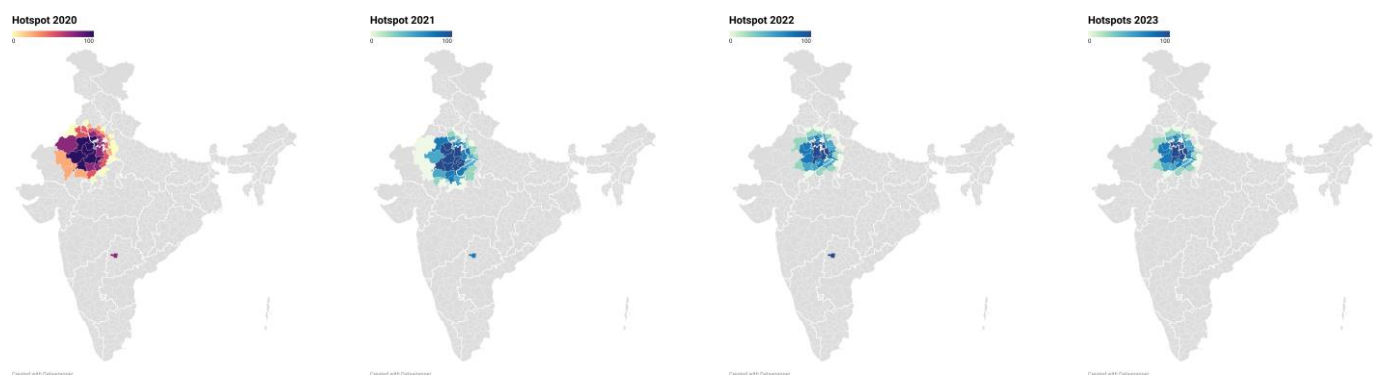


Figure 7: Hotspots of Tuberculosis cases during 2020-23

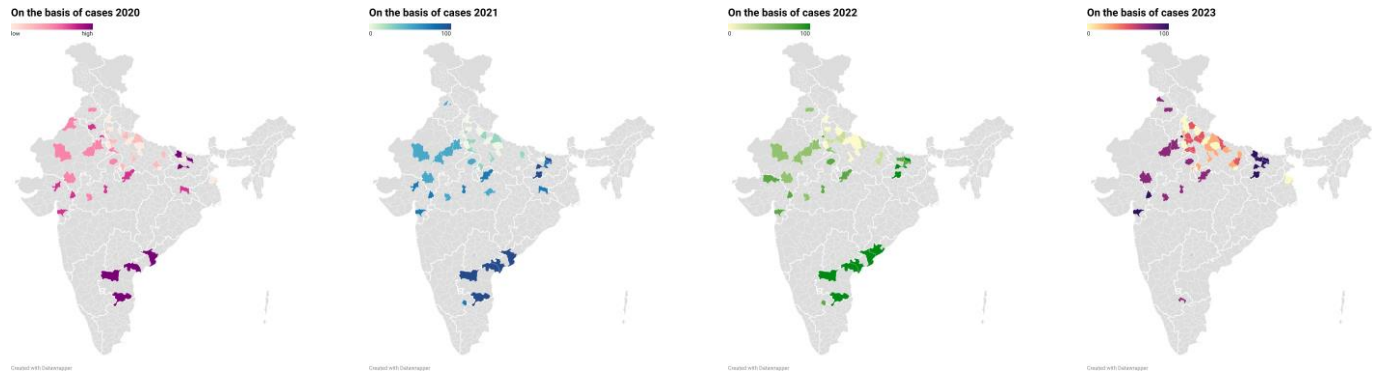


Figure 8: Hotspots on the basis of Tuberculosis cases during 2020-23

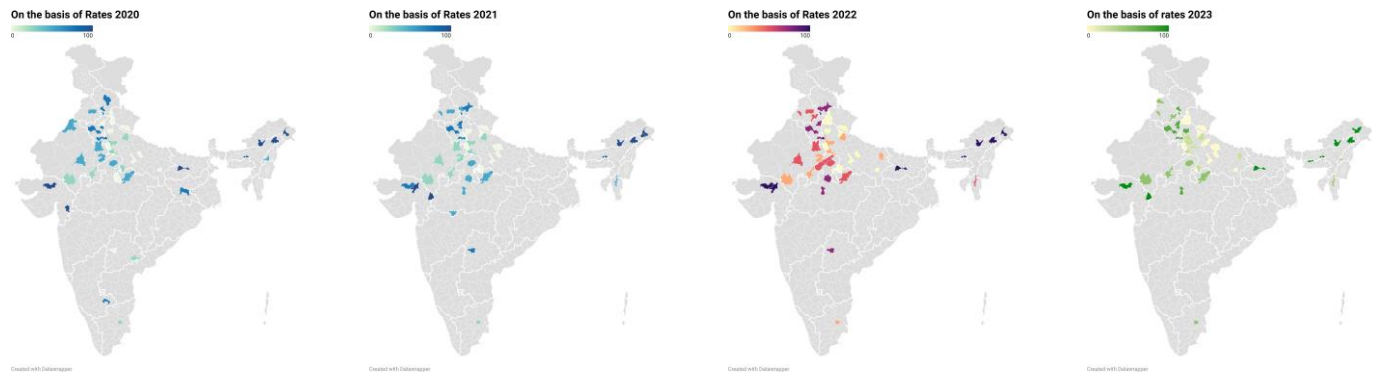


Figure 9: Hotspots on the basis of Rates of Tuberculosis cases during 2020-23

6.3. Statistical modelling and Inferential perspective:

Our goal is to develop and train a statistical model that will help our policy makers and programme implementors in predicting and planning the cases and control process of TB instances in India at district level. In order to fulfill our statistical needs, we used multiple linear regressions, and variable selection was done cautiously. Appropriate data processing has been carried out in order to obtain an understanding of outliers, missing values, and any other unnecessary informations. The fitted regression model is being represented by equation given below. In order to control the variability and stabilize the variance as well as to improve the performance of the model. We explored and implemented the two set of transformations namely $\log y \sim \log x$ and $\sqrt{y} \sim \sqrt{x}$ respectively on set of variables. One set consists of ten factors whereas other set has six factors. We deliberately wish to examine the model performance loaded with ten regressors and the same with six regressors.

The fitted regression for $\sqrt{y} \sim \sqrt{x}$ transformation of 10 variables is

$$y = 0.354x_1 + 0.600x_2 + (-0.158)x_3 + 0.110x_4 + (-0.577)x_5 + 0.190x_6 + (-0.009)x_7 + 0.136x_8 + 0.166x_9 + (-0.056)x_{10}, \quad (23)$$

The fitted regression for $\sqrt{y} \sim \sqrt{x}$ transformation of 6 variables is

$$y = 0.446x_1 + 0.594x_2 + (-0.156)x_3 + 0.086x_4 + (-0.475)x_5 + 0.198x_6, \quad (24)$$

The fitted regression for $\log y \sim \log x$ transformation of 10 variables is

$$y = 0.584 + 0.406x_1 + 2.507x_2 + (-1.945)x_3 + 0.622x_4 + (-1.340)x_5 + 0.124x_6 + (-0.073)x_7 + 0.529x_8 + 0.406x_9 + (-0.062)x_{10}, \quad (25)$$

The fitted regression for $\log y \sim \log x$ transformation of 6 variables is

$$y = 0.114 + 0.498x_1 + 2.781x_2 + (-1.685)x_3 + 0.398x_4 + (-1.066)x_5 + 0.206x_6, \quad (26)$$

The above equations 23, 24, 25, 26 can be used for short term as well as middle term forecast but definitely not recommend for long term forecast because of the possibilities of impacts of time series components and multiple uncontrolled dynamics across the country and may be even outside in any part of the globe. We observe from above fitted models (23-26) that four out of ten regressors namely x_3 , x_5 , x_7 and x_{10} impact y in reverse direction whereas model fitted with six regressors witnessed the reverse impact by x_3 and x_5 . This imply that if we attempt to increase these factors by one unit that are result into corresponding decrease in TB cases. The above equations may be used for forecasting TB cases with $100(1-\alpha)$ confidence level.

6.4. Model diagnosis and related violations:

We have attempted to investigate the validation of model using residual analysis, test of linearity, normality, homoscedasticity and independent of errors in the model using error analysis and error plots. Since normal probability plot gives a straight line pattern; hence normality seems to be preserved as shown in Figure 10. An overall view of the standardized error plots displayed by Figure 11 may be obtained by examining the standard residual versus fitted value plot. The absence of systematic patterns suggested the randomness of distribution of errors. Also, we do not found any cone shaped patterns. These facts support our models premises that the error terms are randomly distributed and preserve constant variance (homoscedastic property). Table 7 shows the computed values of Q-statistic, d statistic (Durbin Watson test statistic) and VIF (Variance Inflation factor) for all four transformed models. A hypothesis testing is performed for each models. Since computed Q values as shown in the column 2 of Table 7 for respective models are smaller than the critical value of Z at $\alpha=0.05$. Therefore, we fail to reject the null hypothesis in all four cases which means that there is no violation of assumption of homoscedasticity (constant variance) and absence of heteroscedasticity is confirmed. The respective values of Durbin Watson test statistic of four transformed models are 2.0974, 2.0726, 1.9871, 2.0378 shows absence of problem of autocorrelation in all the four models under the study. The VIF values are 2.1790, 6.2256, 1.8429, 5.9335 for $\log y \sim \log x$ (10 variables), $\sqrt{y} \sim \sqrt{x}$ (10 variables), $\log y \sim \log x$ (6 variables), $\sqrt{y} \sim \sqrt{x}$ (6 variables) respectively. Since in all four cases our VIF values are below 10, therefore we do not preserve serious multicollinearity in analysis of TB modelling.

Table 7: Test statistic and VIF value for transformed models

Transformations	Q statistic	(DW) d statistic	VIF
$\log y \sim \log x$ (10 variables)	-1.2015	2.0974	2.1790
$\sqrt{y} \sim \sqrt{x}$ (10 variables)	-2.2565	2.0726	6.2256
$\log y \sim \log x$ (6 variables)	-0.4061	1.9871	1.8429
$\sqrt{y} \sim \sqrt{x}$ (6 variables)	-2.0538	2.0378	5.9335

6.5. Goodness of fit and p-values assessment:

Table 8 shows the respective p-values for all the four models, which are very small and close to 0. Also, the respective standard errors shown from the table are statistically significant. Therefore, we reject the hypothesis by saying all the factors are statistically significant and models are valid.

Table 8: ANOVA							
Transformations	Source of variation	df	SS	MSS	F	P-value	SE
log y ~ log x (10 variables)	Regression	10	8.2134	0.8213	4.5981	0.00025	0.4226
	Residual	39	6.9663	0.1786			
	Total	49	15.1798				
log y ~ log x (6 variables)	Regression	6	6.9431	1.1571	6.0412	0.00012	0.4376
	Residual	43	8.2365	0.1915			
	Total	49	15.1797				
$\sqrt{y} \sim \sqrt{x}$ (10 variables)	Regression	10	150.69	15.0690	20.9024	1.09E-12	0.8490
	Residual	40	28.8369	0.7209			
	Total	50	179.527				
$\sqrt{y} \sim \sqrt{x}$ (6 variables)	Regression	6	149.2708	24.8784	36.1795	2.93E-15	0.8292
	Residual	44	30.2561	0.6876			
	Total	50	179.527				

6.6. Model performance criteria and discussion:

Table 9 presents the computed values of R^2 and Adjusted R^2 , PRESS statistic and R^2 for predicted of various transformation model. From investigating the models we may prefer model with $\sqrt{y} \sim \sqrt{x}$ transformation for 10 variables, followed by $\sqrt{y} \sim \sqrt{x}$ for 6 variables, log y ~ log x for 10 variables and log y ~ log x for 6 variables respectively. Table 10 presents the model selection values of standard statistical criteria namely AIC (Akaike information criterion), SBIC (Schwarz Bayesian information criterion), C_p (Mallow's Cp), MSD (Mean Square Deviation), MAD (Mean Absolute Deviation), MAPE (Mean Absolute Percentage Error). The value of spearman's rank correlation between two attributes such as Goodness of fit and Model performance is 0.8 lies between 0 & 1, that shows the statistically significant relationship between the goodness of fit of the model and performance of the model for TB cases, keeping in the view of above discussion though any of the transformed model will do a good job but we prefer ($\sqrt{y} \sim \sqrt{x}$).

Table 9: Computed values of statistics for transformed models

Transformations	R^2	R^2_{adj}	PRESS	R^2_{PRESS}	Ranking (best fit model)
log y ~ log x (10 variables)	0.5410	0.4234	14.5633	0.0406	3
$\sqrt{y} \sim \sqrt{x}$ (10 variables)	0.8393	0.7782	91.1221	-1.0564	1
log y ~ log x (6 variables)	0.4573	0.3816	14.8354	0.0226	4
$\sqrt{y} \sim \sqrt{x}$ (6 variables)	0.8314	0.7895	81.4619	-0.8384	2

Table 10: Model selection criteria

Transformations	AIC	SBIC	C_p	MSD/MSE	MAD	MAPE	Ranking (best model performance)
$\log y \sim \log x$ (10 variables)	-76.5468	-55.5745	13	0.1393	0.2848	0.7148	1
$\sqrt[3]{y} \sim \sqrt[3]{x}$ (10 variables)	-5.5184	15.5138	14	0.5767	0.5432	0.4522	3
$\log y \sim \log x$ (6 variables)	-76.1719	-62.7877	9	0.1647	0.3103	0.7328	2
$\sqrt[3]{y} \sim \sqrt[3]{x}$ (6 variables)	-11.1162	2.2679	10	0.6051	0.5588	0.4847	4

Table 11 presents the lower and upper confidence interval for coefficients associated with the independent regressors for each type of transformed models and corresponding ranks has been assigned. Also, we have attempted to find out the ranking of each of the factors through four transformed models such a ranking is useful to understand factor wise significant under each of the four categories of transformed model. The non-pathogenic factors to be held responsible for TB cases and their growth in India could be PM(2.5), CO, relative humidity, NO_2 , SO_2 . Therefore, our analysis provide strong evidence to support the theory that the role of air pollutants and relative humidity followed by population density, increasing ozone level (O_3), temperature and malnutrition.

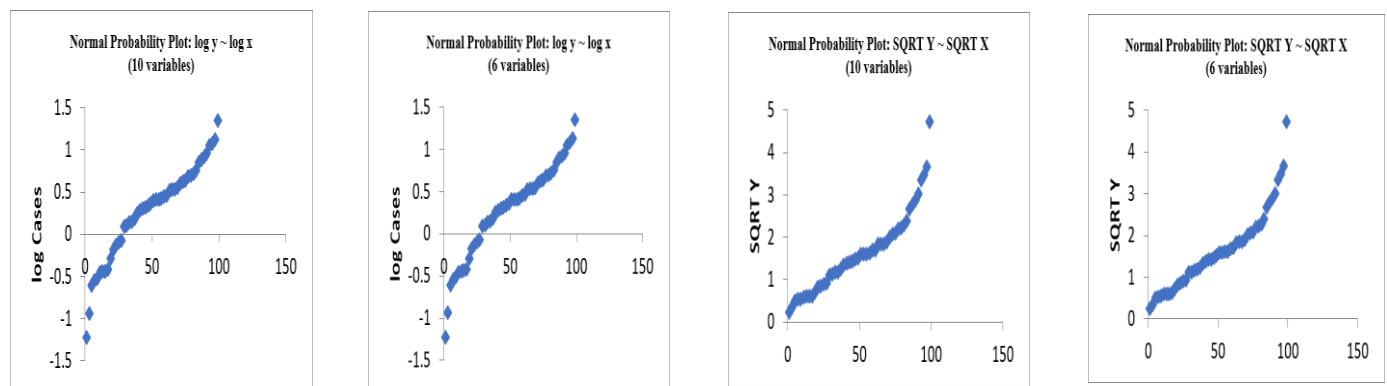


Figure 10: Normal Probability Plot for all transformed models.

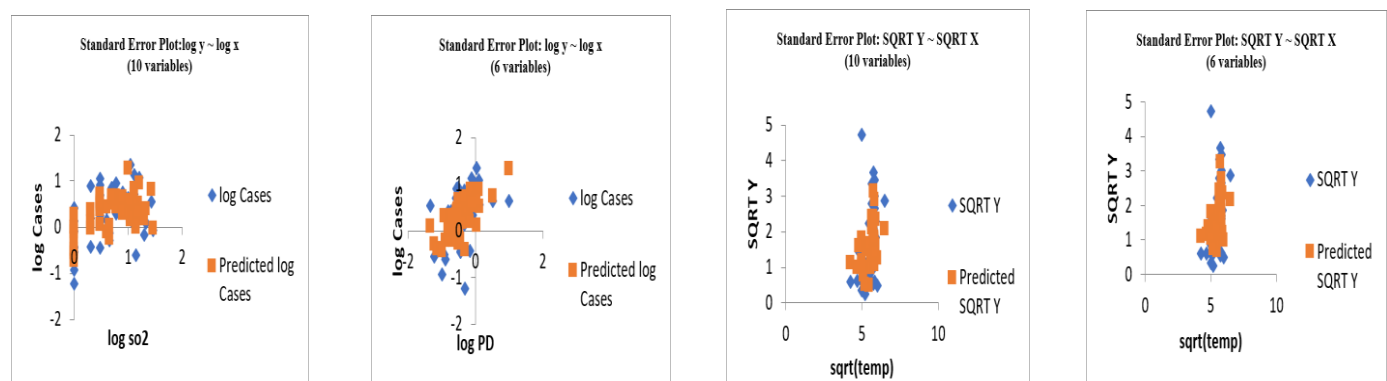


Figure 11: Standard Error Plot.

Table 11: Confidence Interval and Confidence length for the estimated coefficients

Name of Variables	Modelling*	Standard Error	Lower 95.0%	Upper 95.0%	Confidence Length	FR*	OFR*
log PD	1	0.17472	0.052907716	0.759718359	0.706810644	2	9
log temperature	1	1.200002	0.080271094	4.934738287	4.854467193	10	32
log humidity	1	0.814832	-3.59347733	-0.297169388	3.296307942	9	30
log PCI	1	0.327178	-0.038848967	1.284710843	1.323559809	4	21
log malnutrition	1	0.593216	-2.540078865	-0.140294425	2.39978444	8	28
log NO ₂	1	0.41977	-0.724395009	0.973734109	1.698129118	7	25
log PM(2.5)	1	0.207778	-0.493592723	0.346948036	0.84054076	3	11
log O ₃	1	0.399492	-0.278396777	1.337701628	1.616098405	6	23
log SO ₂	1	0.167959	0.067147818	0.746606783	0.679458964	1	8
log CO	1	0.381943	-0.83514614	0.709960043	1.545106183	5	22
sqrt (PD)	2	0.310834	-0.273920643	0.982515398	1.256436042	9	20
sqrt (temperature)	2	0.252288	0.090734636	1.11051989	1.019785253	7	14
sqrt (humidity)	2	0.157824	-0.477826609	0.160122158	0.637948768	4	6
sqrt (PCI)	2	0.487647	-0.874741405	1.096401699	1.971143104	10	26
sqrt (malnutrition)	2	0.293625	-1.1710242	0.015854189	1.18687839	8	18
sqrt (NO ₂)	2	0.212748	-0.239350325	0.620608394	0.859958719	6	12
sqrt PM(2.5)	2	0.085031	-0.181257791	0.162450057	0.343707848	1	1
sqrt (O ₃)	2	0.182041	-0.231059685	0.504779039	0.735838723	5	10
sqrt (SO ₂)	2	0.140598	-0.117550306	0.450768304	0.56831861	3	5
sqrt (CO)	2	0.105692	-0.270323684	0.156899511	0.427223195	2	2
log PD	3	0.167521	0.160563838	0.836240631	0.675676792	1	7
log temperature	3	1.167827	0.426392191	5.136687151	4.71029496	6	31
log humidity	3	0.781918	-3.26216468	-0.108390611	3.153774069	5	29
log PCI	3	0.309331	-0.2252188	1.022432188	1.247650988	3	19
log malnutrition	3	0.585391	-2.246956815	0.114149116	2.36110593	4	27
log NO ₂	3	0.268126	-0.333903379	0.74755159	1.081454969	2	15
sqrt (PD)	4	0.283781	-0.125751211	1.018093012	1.143844223	5	17
sqrt (temperature)	4	0.225728	0.139780881	1.049630198	0.909849317	3	13
sqrt (humidity)	4	0.128987	-0.416638382	0.10327577	0.519914151	1	3
sqrt (PCI)	4	0.412731	-0.745212167	0.918396471	1.663608638	6	24
sqrt (malnutrition)	4	0.275633	-1.030538883	0.080465172	1.111004054	4	16
sqrt (NO ₂)	4	0.130195	-0.063452666	0.46132714	0.524779806	2	4

Modelling* 1 is $\log y \sim \log x$ (10 variables), 2 is $\log y \sim \log x$ (6 variables), 3 is $\sqrt{y} \sim \sqrt{x}$ (10 variables), 4 is $\sqrt{y} \sim \sqrt{x}$ (6 variables),

FR* is factor ranking for different models, OFR* is overall factor ranking.

7. CONCLUSION

To conclude with, we may say that there are almost fifty districts of India that is the hotspot of TB cases, geographically located mostly in north-western states of the country as shown in the Figure 7. As such, we have pinpointed the locations witnessing highest concentration of TB cases during the period of study yr.2020-23; the resultant hotspot districts needs thorough investigations on urgent priority to achieve optimal as well as feasible solutions of the problem at the moment. We found ten factors namely population density, temperature, relative humidity, per capita income, malnutrition, NO_2 , $PM(2.5)$, O_3 , SO_2 , CO out of which six factors viz population density, temperature, relative humidity, per capita income, malnutrition and NO_2 are the pertinent one. It is being evidenced through our modelling process that with sincere efforts we may control these factors and able to apply the break on TB infections and its growth. We become successful in obtaining the best fit model as shown in the equations (23-26). Model performance has been explained satisfactorily and presented in the Table 9. Though, the current work is quite comprehensive in nature but have a complete clarity in regards to factors identification and hotspot determination. The obtained results suggest that we should tackle the primary clusters (MLC) districts first and then concentrate on districts under secondary and tertiary clusters too. We learned that there is a need of launching an effective TB awareness programme through out the country. Also, we should focus on implementation of resource utilisation unbiasedly, efficiently, consistently and sufficiently. From the result we may strongly recommend the following 16 districts namely Patna, Surat, Bangalore, Chhatarpur, Jaipur, Hyderabad, Aligarh, Ghaziabad, Bareilly, Gorakhpur, Kanpur Nagar, Lucknow, Mathura, Meerut, Sitapur, Varanasi characterised as the most prominent hotspot as well as the most consistent districts during the period under the study. We must plan to control and eliminate the factors as mentioned above are responsible for TB infection and growth on urgent basis. With higher probability confidence level, we may foresee the fact and findings of our work can be generalised in any part or atleast in some part of the world. In conclusion, we may recommend strongly that TB cases in India is exclusively due to man-made factors and not by chance, hence can be eliminated effectively with highest possible accuracy and reliability. This will boost our confidence and morals for achieving the complete success towards the set goal. We may anticipate that the hotspot districts of TB cases, if tackled effectively then it would be of great help towards achieving our goal of the complete eradication of TB cases in India by the yr.2030.

7.1. Future Scope

The present work provide ample opportunity to pursue further in-depth research at micro levels (sub-divisions, blocks, panchayats and villages). The people from expert domain (bio-medical professionals) and health care specialist will derive, frame, and develop an effective decision making tools to combat ever alarming TB cases from India.

8. FUNDING

There was no particular grant awarded for this research by governmental, private, or non-profit funding organizations.

9. DATA AVAILABILITY

The datasets used and/or analyzed in this study are available from the corresponding author on reasonable request.

10. CONFLICTS OF INTEREST

The authors declare no conflict of interest.

11. REFERENCE

- [1] Kulldorff, M. S. (1998). v9. 0: software for the spatial and space-time statistics. National Cancer Institute.
- [2] India, T. B. (2021). Report 2022: Central TB Division. Ministry of Health and Family Welfare Available from: <https://tbcindia.gov.in/2020> Last accessed on.
- [3] Ni-Kshay, Government of India: Tuberculosis Report 2020,2021,2022,2023.
- [4] NFHS-5, (2020) National family health survey. Government of India: Ministry of Health and Family Welfare.
- [5] Anjali, B., & Kumar, J. (2021). Spatio-Temporal Aspect of Suicide and Suicidal Ideation: An Application of SaTScan to Detect Hotspots in Four Major Cities of Tamil Nadu. *Journal of Scientific Research*, 65(9).
- [6] Bhatia, U., Tshering, Kumar, J., & Choubey, D. K. (2022). Drowsiness image detection using computer vision. In *Soft Computing: Theories and Applications: Proceedings of SoCTA 2020, Volume 1* (pp. 667-683). Springer Singapore.
- [7] Manish, G. V., Simran, Kumar, J., & Choubey, D. K. (2021, December). Identification of Hotspot of Rape Cases in NCT of Delhi: A Data Science Perspective. In *International Conference on Information Systems and Management Science* (pp. 485-496). Cham: Springer International Publishing.
- [8] Khushboo Sukhija et al., (2017) Spatial visualization approach for detecting criminal hotspots: An analysis of total cognizable crimes in the state of Haryana, 2nd IEEE International Conference on Recent Trends in Electronics Information and Communication Technology (RTEICT).
- [9] Reddy, G. V. M., Iswarya, J. K., & Choubey, D. K. (2023) Time Series Analysis of National Stock Exchange: A Multivariate Data Science Approach.
- [10] Garrison, F. H., & BLOCKER Jr, T. M. G. (1976). An introduction to the history of medicine. *Plastic and Reconstructive Surgery*, 57(4), 510.
- [11] Dye, C., & Williams, B. G. (2008). Eliminating human tuberculosis in the twenty-first century. *Journal of the Royal Society Interface*, 5(23), 653-662.
- [12] Yang, L. J., Li, T., & Chen, W. (2020). Study on spatial clustering characteristics of tuberculosis in China, 2013-2018. *Zhonghua Liu Xing Bing xue za zhi= Zhonghua Liuxingbingxue Zazhi*, 41(11), 1843-1847.
- [13] Zulu, L. C., Kalipeni, E., & Johannes, E. (2014). Analyzing spatial clustering and the spatiotemporal nature and trends of HIV/AIDS prevalence using GIS: the case of Malawi, 1994-2010. *BMC infectious diseases*, 14(1), 1-21.
- [14] Xia, J., Cai, S., Zhang, H., Lin, W., Fan, Y., Qiu, J., ... & Nie, S. (2015). Spatial, temporal, and spatiotemporal analysis of malaria in Hubei Province, China from 2004-2011. *Malaria journal*, 14(1), 1-10.
- [15] Kulldorff, M., & Nagarwalla, N. (1995). Spatial disease clusters: detection and inference. *Statistics in medicine*, 14(8), 799-810.
- [16] Huang, L., Kulldorff, M., & Gregorio, D. (2007). A spatial scan statistic for survival data. *Biometrics*, 63(1), 109-118.
- [17] Tango, T., & Takahashi, K. (2012). A flexible spatial scan statistic with a restricted likelihood ratio for detecting disease clusters. *Statistics in medicine*, 31(30), 4207-4218.

- [18] Li, X. X., Wang, L. X., Zhang, H., Du, X., Jiang, S. W., Shen, T., ... & Zeng, G. (2013). Seasonal variations in notification of active tuberculosis cases in China, 2005-2012. *PloS one*, 8(7), e68102.
- [19] Selmane, S., & L'hadj, M. (2021). Spatiotemporal analysis and seasonality of tuberculosis in Algeria. *Int J Mycobacteriol*, 10(3), 234-42.
- [20] Mesquita, C. R., Concei, E. C., Monteiro, L. H. M. T., da Silva, O. M., Lima, L. N. G. C., de Oliveira, R. A. C., ... & Lima, K. V. B. (2021). A clinical-epidemiological and geospatial study of tuberculosis in a neglected area in the Amazonian region highlights the urgent need for control measures. *International Journal of Environmental Research and Public Health*, 18(3), 1335.
- [21] Tanjung, R., Mahyuni, E. L., Tanjung, N., Simarmata, O. S., Sinaga, J., & Nolia, H. R. (2021). The spatial distribution of pulmonary tuberculosis in Kabanjahe District, Karo regency, Indonesia. *Open Access Macedonian Journal of Medical Sciences (OAMJMS)*, 9(E), 817-822.
- [22] Yu, Y., Wu, B., Wu, C., Wang, Q., Hu, D., & Chen, W. (2020). Spatial-temporal analysis of tuberculosis in Chongqing, China 2011-2018. *BMC infectious diseases*, 20, 1-12.
- [23] Chen, J., Qiu, Y., Yang, R., Li, L., Hou, J., Lu, K., & Xu, L. (2019). The characteristics of spatial-temporal distribution and cluster of tuberculosis in Yunnan Province, China, 2005-2018. *BMC Public Health*, 19(1), 1-13.
- [24] Khazaei, S., Nematollahi, S., Doosti-Irani, A., Zahiri, A., Mofarrah-Zat, A., Ayubi, E., ... & Saatchi, M. (2019). Spatio-temporal analysis of tuberculosis in Hamadan Province, west of Iran, from 1992 to 2013. *Tanaffos*, 18(1), 18.
- [25] Andrade, H. L. P. D., Ramos, A. C. V., Crispim, J. D. A., Santos Neto, M., Arroyo, L. H., & Arcio, R. A. (2021). Spatial analysis of risk areas for the development of tuberculosis and treatment outcomes. *Revista Brasileira de Enfermagem*, 74.
- [26] Shojaei, S. R. H., Waghei, Y., & Mohammadzadeh, M. (2018). Geostatistical analysis of disease data: a case study of tuberculosis incidence in Iran. *Journal of Applied Statistics*, 45(8), 1476-1483.
- [27] Gemechu, L. L., & Debusho, L. K. (2022). Spatial co-clustering of tuberculosis and hiv in Ethiopia. *Diseases*, 10(4), 106.
- [28] Mollalo, A., Mao, L., Rashidi, P., & Glass, G. E. (2019). A GIS-based artificial neural network model for spatial distribution of tuberculosis across the continental United States. *International journal of environmental research and public health*, 16(1), 157.
- [29] Jones, S. G., & Kulldorff, M. (2012). Influence of spatial resolution on space-time disease cluster detection. *PLoS One*, 7(10), e48036.
- [30] Yeboah-Manu, D., Asare, P., Asante-Poku, A., Otchere, I. D., Osei-Wusu, S., Danso, E., ... & Gagneux, S. (2016). Spatio-temporal distribution of *Mycobacterium tuberculosis* complex strains in Ghana. *PloS one*, 11(8), e0161892.
- [31] Huang, L., Michael Abe, E., Li, X. X., Bergquist, R., Xu, L., Xue, J. B., ... & Li, S. Z. (2018). Space-time clustering and associated risk factors of pulmonary tuberculosis in southwest China. *Infectious diseases of poverty*, 7(04), 63-73.

- [32] Tuntun, M., Aminah, S., & Yusrizal, C. H. (2023). Distribution pattern and spatial analysis of factors for tuberculosis (TB) cases in Bandar Lampung City in 2022. *Bali Medical Journal*, 12(1), 50-58.
- [33] Liu, M. Y., Li, Q. H., Zhang, Y. J., Ma, Y., Liu, Y., Feng, W., ... & Guo, X. H. (2018). Spatial and temporal clustering analysis of tuberculosis in the mainland of China at the prefecture level, 2005-2015. *Infectious diseases of poverty*, 7(1), 1-10.
- [34] Rao, H., Shi, X., & Zhang, X. (2017). Using the Kulldorff's scan statistical analysis to detect spatio-temporal clusters of tuberculosis in Qinghai Province, China, 2009-2016. *BMC infectious diseases*, 17, 1-11.
- [35] Frnt, C., Ahmed, M. S., Dabo-Niang, S., & Genin, M. (2023). Investigating spatial scan statistics for multivariate functional data. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 72(2), 450-475.
- [36] Kaindal, S., Venkataramana, B., & Kumar, J. (2023, March). Cancer Hotspot Identification and Analysis: A Scan Statistics Approach. In *International Conference on Information Technology* (pp. 13-28). Singapore: Springer Nature Singapore.
- [37] Devore, J. (2007). *Applied regression analysis: A second course in business and economic statistics*.
- [38] Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2005). *Applied linear statistical models*. McGraw-hill.
- [39] Judge, G. G., Griffiths, W. E., Hill, R. C., Ltkepohl, H., & Lee, T. C. (1991). *The theory and practice of econometrics* (Vol. 49). John Wiley & sons.
- [40] Freund, R. J., Wilson, W. J., & Sa, P. (2006). *Regression analysis*. Elsevier.
- [41] Cornell, J. A. (1987). *Classical and modern regression with applications*.
- [42] Kumar, R. (2024). Exploring cause-specific strategies for suicide prevention in India: A multivariate VARMA approach. *Asian journal of psychiatry*, 92, 103871.
- [43] Daniel TM. Hermann Brehmer (2011) The origins of tuberculosis sanatoria. *Int J Tuberculosis Lung Disease*.