JOURNAL OF
EMERGING INVESTIGATORS

Article

Impact of population density and elevation on tuberculosis spread and transmission in Maharashtra, India

Melosa Rao[1], Ann Johnson[2]

[1]Podar International School, CAIE, Powai, Mumbai, India
[2]Yale University, New Haven, Connecticut

SUMMARY

The causative agent of Tuberculosis (TB) in humans is *Mycobacterium tuberculosis*. When lung TB patients cough, sneeze, or spit, TB spreads through the air. India accounts for over 2.4 million recorded cases of TB, about 26% of the world's TB cases. This research ascertained the bearing of both the population density and the average elevation above mean sea level (MSL) on the number of cases of TB recorded by the districts in Maharashtra, India. Using multiple regression analysis, we demonstrated that about 75% of the variance in the reported cases of TB per thousand people can be attributed to the average elevation above MSL and the population density. We found a strong positive correlation between the number of TB cases per thousand people and the population density and a strong negative correlation between the number of TB cases per thousand people and the average elevation above MSL.

INTRODUCTION

Mycobacterium tuberculosis is the most prevalent causative agent of tuberculosis (TB) in humans (1). *Mycobacterium bovis* is a causative agent of TB in cattle, known as bovine TB (2). *M. bovis* is known to jump the species barrier and cause TB infection in humans and other mammals (2). When *M. bovis* or *M. tuberculosis* infects the lungs, TB is spread through the air when patients cough, sneeze or spit (2). Every year, ten million people are infected with TB (3). Although TB is preventable and curable, 1.5 million people die from TB each year (3). India has the highest incidence of TB, with 26% of the world's TB cases and over 2.4 million recorded cases of TB in 2020 (3).

With a timely diagnosis and treatment with antibiotics, most people who develop TB can be cured and future transmission curtailed (4). However, drug-resistant TB continues to be a public health threat (4). Worldwide in 2019, close to half a million people developed rifampicin-resistant TB (RR-TB) of which 78% had multidrug-resistant TB (MDR-TB) (4). MDR-TB is defined as resistance to rifampicin and isoniazid, the two most effective first-line anti-TB agents (4). Treatment requires a course of second-line drugs for at least 9 months and up to 20 months, supported by counseling and monitoring for adverse events (4). An article by Mathema *et al* lists four main approaches used to measure TB transmission and identify its drivers (5). The approach used to identify high TB risk countries and regions is case notification rates (5).

There are 28 states and 8 union territories in India. States comprise of a number of administrative districts. Maharashtra has 36 administrative districts spread over 300,000 km^2 and a projected population in 2020 of 125 million. Climatic factors in Maharashtra, such as temperature and rainfall, vary due to topological factors including elevation and geographical position. A previous study in India quantified the climatic variations in the smear-positive cases of TB by measuring the amplitude, or peak to the trough distance across seasons. The maximum amplitude of Himachal Pradesh, a state in North India, was 40.01% and the minimum for Maharashtra was 3.87% (6). For all the above reasons, this study evaluates the impact of average elevation above the mean sea level (MSL) of districts on TB cases per thousand people by the districts in Maharashtra.

There are few quantitative studies that establish the association between population density and the incidence of TB. A research study from Canada shows a significant association between TB cases and housing density, isolation from health services and income levels. Overcrowded housing has the potential to increase exposure of susceptible individuals to infectious TB cases, and isolation from health services may increase the likelihood of TB (7). Moreover, a research study in Mexico found that altitude had a strong inverse relationship with TB cases (8). This research investigates the impact of the population density and average elevation above MSL on the number of TB cases per thousand people by the districts in Maharashtra.

Further, we aim to elucidate whether the average elevation above MSL and the population density impact the transmission and spread of TB. By performing a multivariable regression analysis, we identify a linear correlation between TB per thousand people and various factors including average elevation above MSL and population density.

RESULTS

It was observed that the number of TB cases per thousand people and population density in the Mumbai Metropolis were extremely high (**Figure 1**). Also, the neighboring districts of Thane, Raigad and Palghar have significantly higher cases of TB per thousand people as compared to other districts (**Figure 1**). Moreover, cities like Pune, Nagpur, Kolhapur and Aurangabad have higher TB cases per thousand people than other districts. This can be attributed to the travel and migration between urban centers and neighbouring districts. Also, Maharashtra has three major topological regions, the Konkan Region or the coastal belt with low elevation is shown in a lighter shade of green, the tall mountainous region or the Western Ghats is displayed in dark green and the Deccan Plateau that slopes down towards the northeast (**Figure 2**).

To determine the impact of the average elevation above MSL on TB cases per thousand people a two-tailed student's *t*-test for equal variances was performed on two groups, one of districts with a low average elevation above MSL and the other of districts with a higher average elevation above MSL. We found that the mean values of TB per thousand people was significantly greater in areas of low elevation than that of high elevation (p = 1.36E-05, t critical=2.04, t statistic 5.19, CI: 95%). In addition, the mean value of TB cases per thousand people of the group of districts with low average elevation(1.06 0.17) was higher than the mean value of TB cases per thousand people of the group of districts with high average elevation (0.685 0.04).

The dataset was tested for four assumptions of linear regression. The first assumption was that there was a linear relationship between the outcome and the predictor variables. The scatter graph between the population density and the TB cases per thousand people indicate a strong positive linear correlation (Coefficient of Correlation, R = 0.7651) and the Scatter graph etween the average elevation and the TB cases per thousand people show a strong negative linear correlation (Coefficient of Correlation, R = 0.7112). Additionally, the second assumption was met as the dot-plot of standardized residuals was observed to be almost normally distributed and the normal probability plot approximately followed a linear pattern. The third assumption that the independent variables are not highly correlated was satisfied as we found the value of R, the correlation coefficient of the Pearson's bivariate correlation between the population density and the average elevation above MSL to be 0.4389, which is less than 0.80. The Final assumption was accepted as the variance of the error terms was similar across the values of the independent variables and showed no clear pattern in distribution suggesting that the residuals are homoscedastic and approximately rectangular (9).

To evaluate the impact of the population density and the average elevation above MSL on the transmission and spread of TB, we performed multiple variable regression analysis on the dataset.

To ascertain the goodness of fit of the regression model and identify the associations between TB cases per thousand people and variables such as average elevation above MSL and population density, the values adjusted R^2 and multiple R, were calculated.We identified a strong correlation between reported cases of TB per thousand people and variables such as average elevation above MSL and population density (Adjusted R^2 = 0.7449, Multiple R = 0.8717). Adjusted R^2 indicates the goodness of fit of the regression model. Moreover, We found a linear correlation between TB cases per thousand people and variables such as average elevation above MSL and population density (ANOVA, p-average elevation: 3.32E-05, p-population density: 1.83E-06, F-statistic: 1.22E-10, CI: 95%). Furthermore, the coefficient value of population density was significant and positive, indicating a positive linear correlation between TB cases per thousand people and population density. The coefficient value of the average elevation above MSL was significant and negative, demonstrating a negative linear correlation between the TB cases per thousand people and the average elevation above

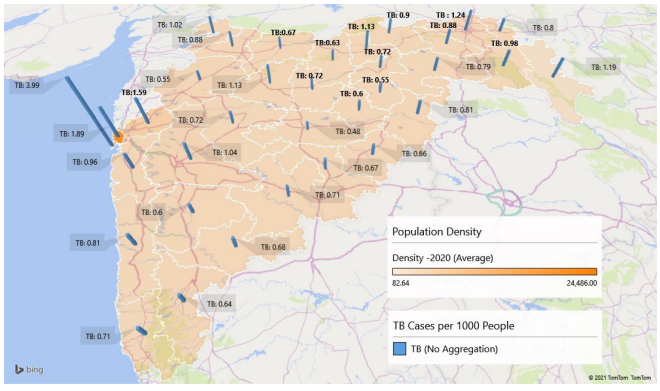

Figure 1. Data visualization of population density and TB cases per thousand people by districts of Maharashtra in 2020. The height of the bars indicate the number of TB Cases per thousand people and the intensity of orange color represents the variation in the population density.

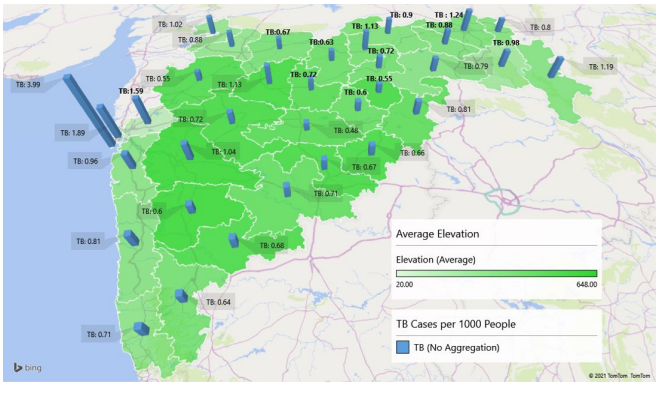

Figure 2. Data visualization of average elevation above MSL and TB cases per thousand people by districts of Maharashtra in 2020. The height of the bars indicate the number of TB Cases per thousand people and the intensity of orange color represents the variation in the Average Elevation above MSL.

MSL. The equation of the regression line was written as:

$Y = 0.000051 * X_1 - 0.00105 * X_2 - 1.22961$
TB Cases per Thousand People = 0.000051 * Population Density – 0.00105 *Average Elevation above MSL – 1.22961

The coefficient of determination, R^2 of the bivariate regression between population density and TB cases per thousand people is 0.5853 (**Figure 3**). Also, the R^2 values of the bivariate regression between average elevation and TB cases per thousand people shown are 0.5058 (**Figure 4**). The adjusted R^2 of the multiple variable analysis is 0.7449. The adjusted R^2 signifies that adding average elevation as an independent variable to the bivariate regression between population density and TB cases per thousand people improves the model fit. Also, adding population density to the bivariate regression between average elevation and TB cases per thousand people significantly improves the model's goodness of fit.

From the above analysis of the regression statistics of the multiple variable regression analysis between reported cases of TB per thousand people and variables such as average elevation above MSL and population density, it is evident that there is a strong positive linear correlation between population density and reported cases of TB per thousand people. Additionally, there is a strong negative correlation between average elevation above MSL and reported cases of TB per thousand people.

DISCUSSION

Using a regression model, we identified a strong positive correlation between population density and TB cases per thousand people. Our results agree with the findings by Clark *et al* that population density, in this case quantified by housing density, has a positive association with the TB cases (7). While migration of labor impacts population density of urban centers, it will interesting to understand the bearing of road, rail, water and air connectivity between urban centers and neighbouring districts on increase in population density on the spread and transmission of TB.

Similarly, we found a strong negative correlation between average elevation above MSL and TB cases per thousand people. Altitude is negatively associated with both pressure and temperature (8). M. *tuberculosis* grows and spreads more effectively at higher temperatures (8). Our results agree with a research study in Mexico which concluded that altitude had a strong inverse relationship with TB cases (8).

We highlighted the gaps in the current literature in our study. To slow the spread of TB transmission, additional research needs to be conducted using Census 2021 data at the sub-district level. Future research should be undertaken to determine the impact of the average elevation, efforts of healthcare workers, availability of treatment,and behavioral factors such as smoking and drinking on the spread and transmission of TB.

Our study assumed that the population densities and the reported TB cases in a given year were evenly spread across the sub-districts. However, that is not the case; the contact rate, spread, and transmission of TB vary across the district. The average elevation above MSL for districts with Western Ghats or hills may not accurately represent the areas with maximum and minimum elevation in the district, especially when the difference between these elevations is significant for a district. To address this discrepancy and further improve the goodness of fit, the regression analysis can be performed on data of sub-districts instead of districts. Census 2021 data will provide a considerable opportunity to conduct such research studies.

This research study establishes a strong positive correlation between the number of reported cases of TB per thousand people and the population density. Likewise, there is a strong negative correlation between the number of recorded cases of TB per thousand people and the average elevation above MSL.

MATERIALS AND METHODS

Data sources

The Nikshay system is the most accurate database of all TB patients across the country. Data of reported cases of TB

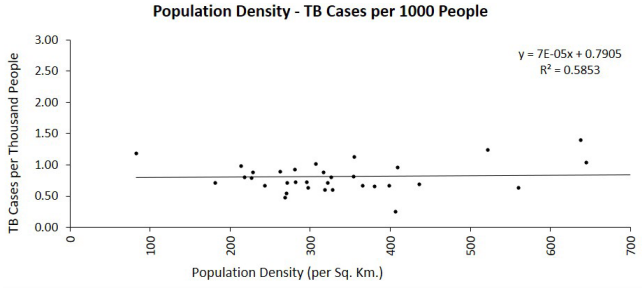

Figure 3. Scatter plot between population density and TB cases per thousand (1000) people by districts of Maharashtra in 2020. There is a strong positive linear correlation between population density and TB cases per thousand people (Coefficient of Correlation, R = 0.7651).

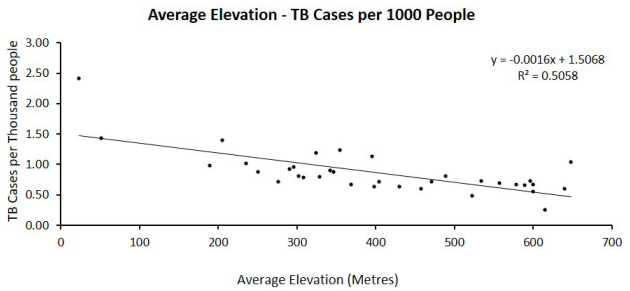

Figure 4. Scatter plot between average elevation above MSL and TB cases per 1000 people by districts of Maharashtra in 2020. There is a strong negative linear correlation between average elevation above MSL and TB cases per thousand people (Coefficient of Correlation, R = 0.7651).

by data points in the districts of Maharashtra was first extracted and downloaded from the Nikshay Portal and then sorted and consolidated by districts into datasets (10). Data downloaded from the Nikshay portal is not entirely by districts. Data was consolidated by districts before calculating the recorded cases of TB by the districts of Maharashtra.

Mumbai Suburban and Mumbai City districts are part of the same megacity. TB patients travel across both these districts every day, for work, treatment and care. For this research study, the districts of Mumbai Suburban and Mumbai City are merged as Mumbai Metropolis.

Aadhaar Saturation data for the population from UIDAI portal accounts for the number of Aadhaar card residents of the district and the migrant population. District portals and survey reports display the Aadhaar Saturation data, and it is considered the most accurate population data of the district at the specified time (11). The area of the districts in square kilometers was collated from district portals and survey reports (12). The average elevation above MSL by districts of Maharashtra were collected from topographic-map.com (13) and validated with the data obtained from the district portals and survey reports. District-level data from these portals of relevant governmental and non-governmental agencies were downloaded, extracted, validated, and compiled into a dataset for 2020.

Data visualization

To visualize the association between the reported cases of TB and the population density, **Figures 1** and **2** were constructed using the 3D Maps tool in MS Excel 2019.

Statistical analysis

To determine the impact of the average elevation above MSL on the TB cases per thousand people, the dataset was sorted in the ascending order of the average elevation above MSL. This was then divided into two groups, the first of districts with a low average elevation above MSL and the second of districts with a higher average elevation above MSL. A boxplot of the TB cases per thousand people was performed and the outliers were eliminated. As the ratio of the variances of both the groups was below 4, they were then considered as equal and a two-tailed student's *t*-test for equal variances was performed.

Before conducting the multiple variable regression model, the dataset was tested for four assumptions of linearity. The first assumption was satisfied when the scatterplot in **Figure 3** showed a linear relationship between TB cases per thousand people and population density. Moreover, each point on the scatter plot represents (population density, TB cases per thousand people) of a particular district of Maharashtra. Also, the scatterplot in **Figure 4** displays a linear relationship between the TB cases per thousand people and the average elevation above MSL. Moreover, each point on the scatter plot (x, y) represents (average elevation above MSL, TB cases per thousand people) for districts of Maharashtra.The second assumption was that the regression residuals were normally distributed. We checked this by looking at the dot plot for normal distribution or the normal probability plot for an approximately linear pattern. The third assumption was that the independent variables were not highly correlated with each other, which is tested using a correlation matrix. When computing a matrix of Pearson's bivariate correlations among population density and average elevation, the magnitude of the correlation coefficient, R, should be less than 0.80. The final assumption was that the residuals are homoscedastic and approximately rectangular-shaped. Residual (e) is the error that isn't explained by the regression line. The e is the difference between the predicted value and the observed value (14). Homoscedasticity in this assumption signifies that the variance of the error terms are similar across the values of the independent variables (15).

Multiple variable regression analysis was performed on the dataset using the data analysis tools in MS Excel. The regression statistics can be calculated by checking the value of the coefficient of determination, R^2, as it denotes the measure of the goodness of fit of our regression model. The adjusted $R2$ represents $R2$ that has been adjusted for the number of predictors in the regression model and indicates the extent to which the variance in the TB cases per thousand people can be attributed to the population density and average elevation. The coefficient of correlation, Multiple R, indicates the extent of correlation between TB cases per thousand people and predictors, such as population density and average elevation. If the p-population density, p-average elevation and F-statistic values in the ANOVA were less than 0.05, as the data was tested at 95% confidence interval, then the null hypothesis that there is no correlation between TB cases per thousand people and independent variables, population density and average elevation, can be rejected. The equation of regression line can be written as:

> TB cases per thousand people = coefficient of population density * population density + coefficient of average elevation above MSL * average elevation above MSL + Y-Intercept value.

The sign and value of the coefficients of the independent variable signify the nature of the correlation between TB cases per thousand people and independent variables, population density and average elevation.

ACKNOWLEDGEMENTS

I would like to thank the JEI editorial team and the scientific reviewers from Stony Brook University, Washington University School of Medicine in Saint Louis, and Harvard Medical School for the detailed feedback. Their change requests and guidance helped me to improve the quality of my research paper. Finally, I thank the entire team of Lumiere Education and Mr. Vikrant Choursiya, for the guidance and support.

Received: March 28, 2021
Accepted: October 9, 2021
Published: November 7, 2021

REFERENCES

1. "Basic Tb Facts." Centers for Disease Control and Prevention. Centers for Disease Control and Prevention, 20 Mar. 2016. Web. 25 Feb. 2021.
2. "Fact Sheets." Centers for Disease Control and Prevention. Centers for Disease Control and Prevention, 01 Sept. 2012. Web. 25 Feb. 2021.
3. "Tuberculosis (TB)." *Who.int*. N.p., 2021. Web. 25 Feb. 2021.
4. "Global Tuberculosis Report 2020." World Health Organization, World Health Organization, 2021, www.who.int/publications/i/item/9789240013131.
5. Mathema, Barun, *et al*. "Drivers of Tuberculosis Transmission." *The Journal of Infectious Diseases*, vol. 216, no. suppl_6, 2017, pp. S644–53. *Crossref*, doi:10.1093/infdis/jix354.
6. Narula, Pankaj, *et al*. "Analyzing Seasonality of Tuberculosis across Indian States and Union Territories." *Journal of Epidemiology and Global Health*, vol. 5, no. 4, 2015, p. 337. Crossref, doi:10.1016/j.jegh.2015.02.004.
7. Clark, Michael, *et al*. "The Association of Housing Density, Isolation and Tuberculosis in Canadian First Nations Communities." *International Journal of Epidemiology*, vol. 31, no. 5, Oct. 2002, pp. 940–45, doi:10.1093/ije/31.5.940.
8. Vargas, Furuya, and Guzman. "Effect of Altitude on the Frequency of Pulmonary Tuberculosis." n. pag. Print.
9. Conrad George Carlberg. Regression Analysis: Microsoft Excel. Indianapolis, Indiana, Que, 2016.
10. "Dashboard::Nikshay Reports." Reports.nikshay.in, reports.nikshay.in/Reports/TBNotification.
11. "Aadhaar Public Data Portal." Gov.in. N.p., Jan. 2021. Web. 25 Feb. 2021.
12. "District of India." Home | DISTRICTS OF INDIA. 7 July 2020. Web. 05 Feb. 2021.
13. "Maharashtra." *Topographic-map.com*. N.p., Feb. 2021. Web. 25 Feb. 2021.
14. Stephanie. "Residual Values (Residuals) in Regression Analysis." *Statistics How To*. 08 June 2021. Web. 22 June 2021.
15. "Homoscedasticity." *Statistics Solutions*. 03 Aug. 2021. Web. 20 Aug. 2021.