CLUSTERING AND REGRESSION BASED APPROACH FOR STUDYING THE CORRELATION BETWEEN TB DYNAMICS AND GHG EMISSIONS

Samarjeet Kaur and Sudhanshu Sharma

Student, ECE Dept.; CSE Dept.; Bharati Vidyapeeth’s College of Engineering, India
Email: sammarjeetkaur@gmail.com, sudhanshu041@gmail.com

Abstract

Tuberculosis (TB) is a very contagious and lethal disease. Its treatment ranges from at least a few months to a few years in severe cases, and is not guaranteed to work every time. Since Tuberculosis is transmitted by air and directly affects the lungs, the possibility of existence of some correlation between greenhouse gas (GHG) emissions, which weaken the respiratory organs over time, and Tuberculosis cases, has been explored in this study. This manuscript specifically explores the correlation between greenhouse gas emissions and the occurrence and mortality rates of Tuberculosis cases across the world. Initial data visualization indicated that there was a positive correlation between the emission of certain gases and the occurrence of TB cases, which was further confirmed by the use of regression and K-Means clustering algorithm. A regression R² score of 0.64 was obtained for predicting mortality rates using GHG emissions as input. The analysis indicated that greenhouse gases do have some impact on the extent and fatality of TB cases.

Keywords – Tuberculosis, Greenhouse gases, K-Means clustering, Linear regression, Data Visualization.

1. Introduction

One of the deadliest diseases plaguing the modern world, Tuberculosis is a lower respiratory infection, claiming 1.6 million lives in 2016 alone [1]. Being an airborne disease, it puts those high at risk who live in crowded places, as well as the younger populace. Its high infection time, ease of transmission, as well as the long recovery period [2], contribute to it being an impeding factor in development.

Greenhouse gas is an umbrella term for water vapor, carbon dioxide, methane, nitrous oxide, ozone, and artificially synthesized compounds, such as Chlorofluorocarbons (CFC’s). Of the vast number of gases, Carbon Dioxide is the most prominent of all, followed by methane. Together, they constitute more than 90% of the gases, other than water vapor, released in the atmosphere. Together, they are responsible for raising the temperature on a planet than the temperature of the planet in the absence of the cloud, on earth. Their lifetime in the atmosphere ranges from a few years (methane) [4], to thousands of years (fluorinated gases), while other gases move to different parts of the ecosystem (carbon dioxide etc.) due to their ability to interact with other elements of the atmosphere, as well as land and sea systems. Carbon dioxide has long been known to be retained by lungs and cause irreparable damage to already infected ones [7], vehicles and factories being major polluters of it. Nitrogen dioxide, formed by the nitrogen and oxygen present in the atmosphere, is another gas which contributes to Tuberculosis related deaths. Greenhouse gas emissions have been rising steadily over the past few centuries, with some gases exceeding more than 100 times from their previous levels. This has led to an increase in trapping of heat reflected from the surface of the earth, which consequently increases the temperature of the earth. An increase of 1.4°F (0.8°C) has been observed in the global average temperature in the last 100 years, with 2016 being the hottest year since temperatures began recording in 1895, with the previous 15 years raising the bar every year since 2001. This effect is detrimental [3] to the earth, as well as resulting in extreme climatic changes on different parts of the planet. Heat waves have become hotter, while sea levels are rising alarmingly fast, both being a direct effect of global warming. It causes disruption in the ecosystem of Earth, which cannot be mended in a few years, and which will be harmful for the generations to come. But recording data of their concentration in the atmosphere, as well as the type of greenhouse gas released, is a relatively new practice.

Since the emission of greenhouse gases directly affects the atmosphere, their impact on the health has been researched [8], leading to proof that the meteorological factors do impact the number of Tuberculosis cases in a region.

Besides this, efforts have been put in calculating the success rate of a treatment [9], with the geographic location...
of the patient being a key variable in correctly predicting the outcome of the treatment. Advancements in computers has contributed towards the development of models which can process data and then forecast future values based on previous values, with people using it in weather predictions, to spam detection. However, their scope is not just limited to these applications, with supervised machine learning models being used to detect Tuberculosis [10], with success rate north of 99.8% in most cases. Despite the availability of data and modern computational resources, an extensive analysis to explore the possible correlation between greenhouse gas emissions (while treating each gas as an individual contributor) and TB cases has not be carried out before. This is probably because the explicit influence of GHGs on TB cases is not very intuitive. Nonetheless, the fact that GHGs impact the quality of air and have an effect on the functioning of the human respiratory system, even the remote possibility of there existing a relation between the two cannot be dismissed.

This study being an explorative one, aims to reveal the correlation between TB cases and GHG emissions across the world and establish a quantitative basis for the same with the help of Linear Regression and K-Means clustering.

2. Methodology

2.1 Dataset

The objective of this work was to find the correlation between frequency of Tuberculosis cases and volume Greenhouse Gas Emissions (GHGs) across the world. Therefore, two kinds of data were needed. The GHG data was obtained from Kaggle [11], officially hosted by the United Nations (UN) website [12]. This covers the period from 1990 to 2014, and has the most recent submission of the information from every source. It is made up of the name of the country, followed by the year and the amount of anthropogenic emission of greenhouse gases, namely Carbon Dioxide (CO2), Methane (CH4), Nitrous Oxide (N2O), Hydrofluorocarbons (HFCs), Perfluorocarbons (PFCs), their unspecified mixes, and Sulphur Hexafluoride (SF6), unspecified mix of HFCs and PFCs, GHG cumulative emissions, along with Nitrogen Trifluoride (NF3). The emission values were recorded in kiloton CO2 equivalent. It was released by the UN in 2016.

The Tuberculosis data was segregated into different tables containing a wide variety of parameters. WHO website had an overwhelming number of parameters (more than 500), most of which weren’t directly related to the transmission of the disease, but rather the coping mechanism of the nation. Exploring the effects of each of those features was beyond the scope of one study. Factors such as budget allocated to the health centers, number of management units, were considered as out of scope in the current study, and thus had to be discarded in favor of more influencing variables. Therefore, relevant parameters were filtered out based on subjective factors and only WHO TB burden estimates and country-wise TB notifications data was used for the analysis. More specifically, total incidents per 100,000 people, mortality per 100,000 cases and absolute number of relapse cases were the final parameters taken from the data.

2.2 Data Preprocessing

Since the data inventory curated by UN and WHO relied on period submissions by respective countries, it was subject to a lot of heterogeneity in terms of the parameters and time frame across which it was recorded. Therefore, as the first step, a common time interval was determined over which both the GHG and TB data records were recorded. It was seen that starting from 2000 up till 2014, records of the chosen parameters in both the datasets were present. Further, the countries present in GHG dataset were not the same as those present in the TB dataset. Hence, country filtering was performed and only those countries were retained which were present in both the datasets. This resulted in a total of 39 countries. The next step was to filter out relevant parameters. Mortality and incident rates per 100,000 of the population were chosen from the TB dataset. It is to be noted the records for number relapse cases were majorly missing in the dataset. For some countries, this value had while certain other countries had no records of relapse cases at all. Interpolating missing values was not possible due to the sparsity of records and hence this parameter was not included in the analysis.

For the GHG data, the cumulative values of GHG emissions were dropped because the aim was to emphasize on the individual gas contributions. It was also found that records corresponding to PFCs, HFCs, NF3 and unspecified mix of HFCs and PFCs were absent for many countries. Therefore, the values for these categories of gases were not considered. Whereas, the emission values corresponding to CO2, CH4, N2O and SF6 were considered for the experimental analysis.

2.3 Data Visualization

In order to assess and understand the data at hand and choose a suitable machine learning model for predicting target variables, it was essential to explore the correlation between different parameters of the dataset using visualization methods. Therefore, for every greenhouse gas in the GHG dataset, the maximum correlation value with every parameter in the TB dataset was calculated. Three kinds of correlations
namely Kendall, Spearman and Pearson were calculated to eliminate methodology bias. Fig 2. Depicts the data visualization pipeline.

Fig 2. Data Visualization pipeline

Fig 3. Number of countries with frequency of TB incidents maximally and positively correlating with CO2, CH4, N2O and SF6

Fig 4. Number of countries with mortality per 100,000 cases maximally and positively correlating with CO2, CH4, N2O and SF6

The graphs depicted above were generated for each of the three parameters of the TB dataset. Three kinds of correlation values were calculated for each TB parameter and GHG gas pair and the maximum values were retained. These bar graphs show the frequency of countries which had the maximum correlation values corresponding to the greenhouse gases for every TB parameter. It can be observed that for each TB parameter, CH4 (methane) has the highest maximal correlation frequency. This hints at the possibility of methane negatively impacting TB occurrence and fatality of cases. On the other hand, the contributions of CO2, N2O and SF6 are varying across parameters.

2.4 K-Means Clustering

Dividing a large number of data into groups which have some sort of distinction between them is known as clustering. Clustering is an unsupervised learning algorithm used for gaining useful insights about the structure of the data. The data points in a particular group are related to each other by a common property, and there are many algorithms which help us to achieve this segregation.

Considering the heterogenous nature of the parameters, K-Means [13] clustering was used to find out the underlying subgroups within the data for a better understanding of the correlation that exists between greenhouse gases and TB cases. In order to perform clustering on the given data, 6 input features were chosen namely the mortality rate, incident rate and emission values of CO2, NH4, N2O and SF6 in kilotonne CO2 equivalent.

For cluster determination, it was important to remove or fill in missing data field wherever needed. Figures for relapse cases were majorly absent across many years for multiple countries. Therefore, elimination was preferred over interpolation in this case since there wasn’t enough data for reasonable estimation. Hence, only the mortality rate and frequency of cases were the chosen features from the TB dataset. As for the GHG dataset, all four gas emission values were considered. The data was standardized before feeding it as an input by removing the mean and dividing it by unit variance.

The K-Means algorithm requires that the number of clusters be specified beforehand. This step may introduce manual bias therefore elbow point [14] determination was performed to determine the optimal number of clusters. For this, the clustering algorithm was executed 39 times (corresponding to the total number of countries) and the sum of the squared differences (SSE) score was recorded. The SSE curve was generated by plotting the SSE score against the number of clusters and the elbow point was determined using Kneed Python library.

The Kneed package was used for locating the knee point on the curve which turned out to be 6. Therefore, the optimal K value was found to be 6.
Since the data at hand was in the form of a time series data, it was crucial to fix the time variable so as to not consider it as an input feature. Hence, clustering was performed separately for each year from 2000 to 2014. It was seen that the algorithm yielded the same results for every year. For illustration purposes, cluster plot for a single year is depicted below.

The PCA scatter plot depicted above shows the 39 datapoints with different colors symbolizing their clusters. Points having the same color belong to the same cluster.

The output of the clustering algorithm has been presented in Table 1 which shows the country to cluster number mapping. It can be seen that cluster 1 forms the largest cluster while clusters 2 and 3 contain single countries.

United States of America and Russian Federation are the largest greenhouse gas emitters as per the data provided. This could be one of the reasons why they were mapped to separated clusters. In depth analysis of the clusters was beyond the scope of this study because the size of the dataset was not large enough to derive widely applicable conclusions.

### Table 1: Cluster numbers and corresponding countries

| Cluster Number | Countries                        |
|----------------|----------------------------------|
| 0              | Romania, Lithuania, Belarus      |
| 1              | Luxembourg, France, Iceland, Croatia, Belgium, Ireland, Monaco, Turkey, Netherlands, Denmark, Finland, Hungary, Slovenia, Canada, Spain, Norway, Italy, Greece, Australia, Cyprus, Slovakia, Austria, Switzerland, Sweden, Poland, New Zealand, Malta |
| 2              | United States of America         |
| 3              | Russian Federation               |
| 4              | Japan, Germany                   |
| 5              | Latvia, Estonia, Portugal, Bulgaria |
| 6              | Ukraine                          |

### 2.5 Multiple Linear regression

Regression is a statistical method [15] employed in a variety of fields in which the trait/value of one variable is predicted using the value(s) of other variable(s). There are different regressions to solve a host of problems, with linear, multiple, and logistic being the most popular ones.

In this study, multiple linear regression was chosen considering the number of features and the size of the dataset. Regression was performed separately for both the TB parameters. The incident and mortality rates were chosen as the target parameters while the four gas emission values were the input features to the model. It is to be noted that the record for every year was treated as an individual data point. There were 585 data points in the original set out of which 3 had null values. Hence, before feeding the data as input the null values were removed resulting in 582 data points. This data was split into train and test set in 85:15 ratio. The scores were obtained on the test set and the scatter plot was plotted using the test set and the predicted labels respectively.
Fig. 7 Scatter plot of true labels and predicted labels obtained from regression after predicting incidents per 100,000 people.

Fig. 8 Scatter plot of true labels and predicted labels obtained from regression after predicting mortality per 100,000 people.

Table 2: Evaluation metric values

| Parameter     | R2 Score | Root mean squared error (RMSE) |
|---------------|----------|--------------------------------|
| Mortality per 100k | 0.6419   | 0.3582                         |
| Incidents per 100k | 0.3242   | 0.6758                         |

The performance of the regression model was evaluated using the R squared [16] metric also known as the coefficient of determination and the Root Mean Squared Error (RMSE) [17]. R2 score denotes the proportion in variance of the dependent variable which the independent variable can help predict. RMSE is essentially a measure of how far the predicted output is from the ground truth. The square root of the mean difference predicted and ground truth results in the RMSE value.

Fig 7 and 8 show the scatter plot representation of predictions in comparison with the ground truth for incident and mortality rates respectively. The predictions are seemingly more accurate in the mortality plot as compared to incidents plot. This is validated by the R2 and RMSE scores given in Table 2. It can be observed that the R2 score for incidents is two times that of mortality predictions. This implies that there’s a greater correlation between mortality cases and GHG emissions. The obtained scores indicate that the model is performing moderately well for the mortality predictions and it can be improved further by training it on a larger dataset.

3. Conclusion

This research work explored the correlation between greenhouse gas emissions and the occurrence and mortality rates of Tuberculosis cases across the world. Initial data visualization indicated that there was a positive correlation between the emission of certain gases and the occurrence of TB cases, and was further confirmed by the use of regression and K-Means clustering algorithm. The analysis indicated that greenhouse gases indeed do have some impact on the extent and fatality of TB cases.

In the future, we plan to increase the scope of this research by using more methods of validation and testing, which will expand this study to include wider number of parameters to derive further useful insights. Also, an in depth exploration of the impact of CH4 on TB cases can be conducted.

References

[1] Goosby, Eric, Dean Jamison, Soumya Swaminathan, Michael Reid, and Elizabeth Zuccala. "The Lancet Commission on Tuberculosis: building a Tuberculosis-free world." Lancet (London, England) 391, no. 10126 (2018): 1132.
[2] Churchyard, Gavin, Peter Kim, N. Sarita Shah, Roxana Rustomjee, Neel Gandhi, Barun Mathema, David Dowdy, Anne Kasmar, and Vicky Cardenas. "What we know about tuberculosis transmission: an overview." The Journal of infectious diseases 216, no. suppl_6 (2017): S629-S635.
[3] Talaat, Adel M., Sarah K. Ward, Chia-Wei Wu, Elizabeth Rondon, Christine Tavano, John P. Bannantine, Rick Lyons, and Stephen A. Johnston. "Mycobacterial bacilli are metabolically active during chronic Tuberculosis in murine lungs: insights from genome-wide transcriptional profiling." Journal of bacteriology 189, no. 11 (2007): 4265-4274.
[4] Jasmer, Robert M., Christopher B. Seaman, Leah C. Gonzalez, L. Masae Kawamura, Dennis H. Osmond, and Charles L. Daley. "Tuberculosis treatment outcomes: directly observed therapy compared with self-administered therapy." American journal of respiratory and critical care medicine 170, no. 5 (2004): 561-566.
[5] Lambert, Marie-Laurence, Eppo Hasker, Armand Van Deun, Dominique Roberfroid, Marleen Boelaert, and Patrick Van der Stuyft. "Recurrence in Tuberculosis: relapse or reinfection?." The Lancet infectious diseases 3, no. 5 (2003): 282-287.

[6] Fernandes, Fernanda Monteiro de Castro, Eder de Souza Martins, Daniella Melo Arnaud Sampaio Pedrosa, and Maria do Socorro Nantua Evangelista. "Relationship between climatic factors and air quality with Tuberculosis in the Federal District, Brazil, 2003-2012." Brazilian Journal of Infectious Diseases 21, no. 4 (2017): 369-375.

[7] West, John B. "Causes of carbon dioxide retention in lung disease." New England Journal of Medicine 284, no. 22 (1971): 1232-1236.

[8] Li, Zhongqi and Pan, Hongqiu and Liu, Qiao and Wang, Jianming. Exploring the Role of Meteorological Factors in Predicting Incident Pumonary Tuberculosis: A Time-Series Study in Eastern China (8/3/2020). Available at SSRN: https://ssrn.com/abstract=3667848 or http://dx.doi.org/10.2139/ssrn.3667848

[9] Baussano, Iacopo, E. Pivetta, L. Vizzini, Francesca Abbona, and M. Bugiani. "Predicting Tuberculosis treatment outcome in a low-incidence area." The International journal of Tuberculosis and lung disease 12, no. 12 (2008): 1441-1448.

[10] Alile, Solomon Osarumwense. "A Novel Approach For Detecting Tuberculosis Based On Observed Manifestations Using Supervised Machine Learning."

[11] U. Nations, “International Greenhouse Gas Emissions,” Kaggle, 16-Nov-2017. [Online]. Available: https://www.kaggle.com/unitednations/international-greenhouse-gas-emissions. [Accessed: 08-Dec-2020].

[12] "UNdata explorer," United Nations. [Online]. Available: http://data.un.org/Explorer.aspx. [Accessed: 05-Nov-2020].

[13] Capó, Marco, Aritz Pérez, and Jose A. Lozano. "An efficient approximation to the K-means clustering for massive data." Knowledge-Based Systems 117 (2017): 56-69.

[14] Antunes, Mário, Joana Ribeiro, Diogo Gomes, and Rui L. Aguiar. "Knee/elbow point estimation through thresholding." In 2018 IEEE 6th International Conference on Future Internet of Things and Cloud (FiCloud), pp. 413-419. IEEE, 2018.

[15] Santiago, Celine B., Jing-Yao Guo, and Matthew S. Sigman. "Predictive and mechanistic multivariate linear regression models for reaction development." Chemical science 9, no. 9 (2018): 2398-2412.

[16] Plonsky, Luke, and Hessameddin Ghanbar. "Multiple regression in L2 research: A methodological synthesis and guide to interpreting R2 values." The Modern Language Journal 102, no. 4 (2018): 713-731.

[17] Kaneko, Hiromasa. "A new measure of regression model accuracy that considers applicability domains." Chemometrics and Intelligent Laboratory Systems 171 (2017): 1-8.