IS590PD Practical Health Data Analytics Project

Final Report

Acute Lower Respiratory Diseases in Americas

Vinu Prasad Bhambore – vpb2

Dhruman Jayesh Shah - djshah5

Srijith Srinath – ssrina2

A report submitted in part fulfilment of the degree of

MS in Information Management (MSIM)

Instructor: Dr. Ian Brooks

School of Information Sciences

University of Illinois at Urbana-Champaign

May 2020

Table of Contents

Introduction

The primary purpose of this research work is to build a time series model that can be applied to predict and forecast any type of disease's death rate. However, in this particular report we provide a walk-through of building a time series model to predict two separate lower respiratory diseases' death rate namely Bronchitis and Tuberculosis. To achieve a time series model, we will try to work around three research questions. Firstly, we try to see how the indicators affect the death rate count. Second, we will see if the given dataset can be used to build a time series model. Lastly, we will try to include the indicators to build a machine learning model that can learn and predict the time series in interest.

The paper will focus on the different steps that were taken in order to get to the time series model. Here we will illustrate how the indicators from World Bank data have an influence on death rate which was collected from PAHO Mortality data. We will also show how adding indicators we were able to increase the predictive capability of the time series model. In the end of this research we will have steps that can be used to predict any type of disease given we have the required data sets. For most of our analysis, we have provided evaluation metrics that were used in order to achieve an efficient model for time series prediction.

Background

Before the analysis, it's important to understand what Bronchitis and Tuberculosis are and the symptoms of these diseases. Firstly, Bronchitis is an inflammation of the lining of your bronchial tubes, which carry air to and from your lungs. People tend to have thickened mucus, which can end up being discolored. At the same time, a few people tend to have cold or other respiratory infection with acute bronchitis. However, in chronic bronchitis people have severe inflammation of the lining of bronchial tubes with constant irritation.

Tuberculosis is another type of infectious disease that mainly affects the lungs. This disease is also known to spread to other parts of the body and also cause permanent damage to them. Some of the symptoms include cough, fever, weight loss or night sweats, this can lead to more transmission of the bacteria to others. Tuberculosis usually spreads from one person to another through air when the infected person coughs or sneezes.

The paper uses these two diseases to conduct the analysis because they are easily communicable, and the chances of community spread is really high. By building a model on such diseases that will look at country level statistics, the chances of identifying potential spread can be improved. With a good and efficient time series model we can try to prevent these diseases or improve the indicators to ensure better health care service in general.

Research Questions

The paper will try to answer three research questions which are briefly discussed in this section. Firstly, the paper tries to illustrate how the indicators are influencing the death rate count for each of the countries we have selected. Here we will try to plot a few graphs and conduct exploratory data analysis to identify the extent of influence of all the indicators we collected.

Secondly, the paper will explore whether a time series model is viable. As the central limit theorem dictates that for a sample to represent a population dataset, the sample must contain at least 30 observations. Since the data collected was from 1996 to 2017, this means that a time series model might not have enough data for building a time series prediction model.

In the end, the paper will demonstrate how using indicators a time series model can be built. Here, we will use all the indicator data such as GDP, population, health expenditure and number of physicians per 1000 people in order to predict the death rate for each of the countries.

Datasets

There are primarily two datasets we have used for the analysis. Each of these datasets are explained in detail below. One thing to note is in the images there might be some extra columns that we haven't used, the description includes only those columns that are of our interest.

PAHO Mortality Dataset

This dataset was collected from the WHO website, which basically has information about health situations and trends from the region of Americas. The dataset consists of records or observations of deaths with regard to a particular disease which can be identified by the ICD10 column.

	CountryName	MortalityYear	Gender	AgeGroupCode	ICD10	Deaths
0	Brazil	2017	Male	21	I479	1
1	Brazil	2017	Male	21	C925	1
2	Brazil	2017	Male	21	I451	1
3	Brazil	2017	Male	21	D292	1
4	Brazil	2017	Male	21	L519	1

Fig 1: PAHO Mortality Dataset

The columns we will use in the analysis are:

1. CountryName - the name of the country
2. MortalityYear - the year when the observation was recorded for
3. ICD10 - the column that will help in identifying the type of disease

4. Deaths - the total number of deaths for that particular disease in that particular year

World Bank Dataset

This dataset was collected from the Data World bank website, which basically has information about the different indicators we will be using to add more information to the already existing PAHO Mortality dataset. Each observation in this dataset consists of data with regards to a country and its associated statistics.

	CountryName	MortalityYear	Class	Deaths	GDP	Health_Expenditure	Number_of_Physicians_per1000_people	Population
0	Argentina	1997	Bronchitis	282	8543.028534	NaN	NaN	35657429.0
1	Argentina	1997	Tuberculosis	144	8543.028534	NaN	NaN	35657429.0
2	Argentina	1998	Bronchitis	282	8772.063210	NaN	3.0021	36063459.0
3	Argentina	1998	Tuberculosis	127	8772.063210	NaN	3.0021	36063459.0
4	Argentina	1999	Bronchitis	314	8381.253998	NaN	NaN	36467218.0

Fig 2: World Bank Dataset

The columns we will use in the analysis are:

1. CountryName - the name of the country
2. MortalityYear - the year when the observation was recorded for
3. Class - whether it is Bronchitis or Tuberculosis
4. Deaths - total number of deaths observed
5. GDP - Gross Domestic product value for a country
6. Health_Expenditure - the amount of money spent on health on a average
7. Number_of_Physicians_per1000_people - the total number of physicians per 1000 people observed in the country
8. Population - the population of the country

Data Problems

The dataset that was collected had multiple problems with respect to whether we could use it for a time series model or not. In both the datasets some of the problems that we observed were missing data, limited data points and inconsistency in naming. Each of these problems is discussed in detail below, along with these are also provided steps which were executed to clean up the dataset.

Data Inconsistency

There was inconsistency of how the countries were named in PAHO Mortality and World Bank datasets. There was no other simpler method than to list all the unique names of the countries between the datasets and perform set operation to identify the differences. Once we had the inconsistent names, we manually named the countries with the same spelling and capitalization in both the datasets. The reason this step was important for data cleaning was because of the merge that needed to be done once the datasets were cleaned.

Limited Data Points

In the PAHO Mortality dataset it was observed that there were countries that had no data points for a few ICD codes. This could be either because the countries did not report any deaths for these ICDs or there were no deaths observed at all. This is clearly illustrated in the below two graphs.

This first graph is for Bronchitis while the second is for Tuberculosis. The graphs show the death counts for each of the ICDs over time. Here we observe that the legend shows different colors for the ICDs, however in the graph we can see 2-3 predominant values for the ICDs. This helps in understanding about the limited data points for either of the two diseases when looked at from the ICD level. The reason this could end up being a problem is that a time series modeling algorithm does not function efficiently with missing or limited data points.

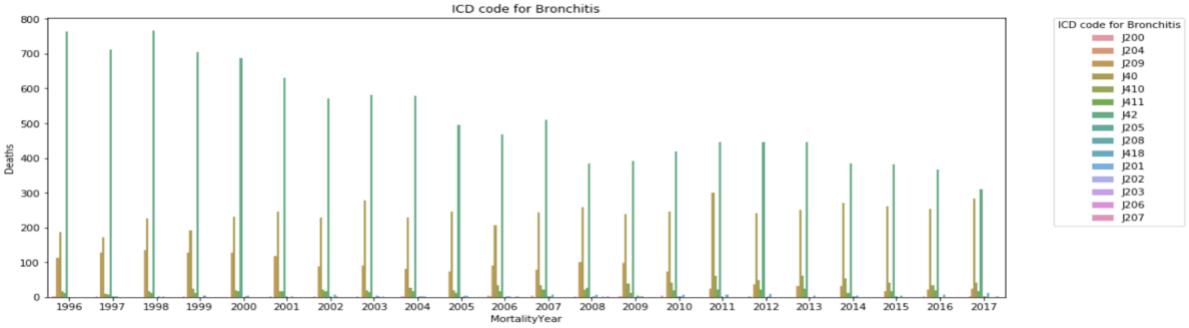


Fig 3: ICD code distribution for Bronchitis

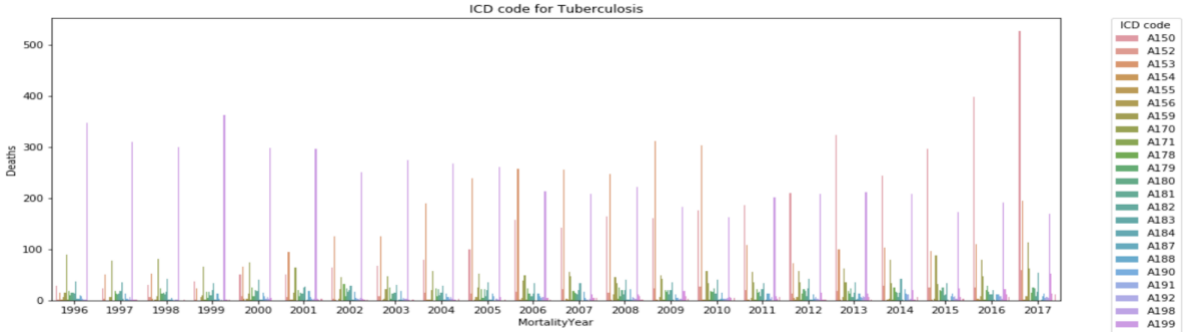


Fig 4: ICD code distribution for Tuberculosis

However, there is a way we can fix this by actually taking a look at the data from a higher level of abstraction. What this means is, the observations are classified as whether they are Bronchitis or Tuberculosis. Even though this might cause a loss of granularity, we observe that the limited data points for a time series is avoidable by aggregation at a higher level. Below is a graph with aggregated data for Bronchitis and Tuberculosis.

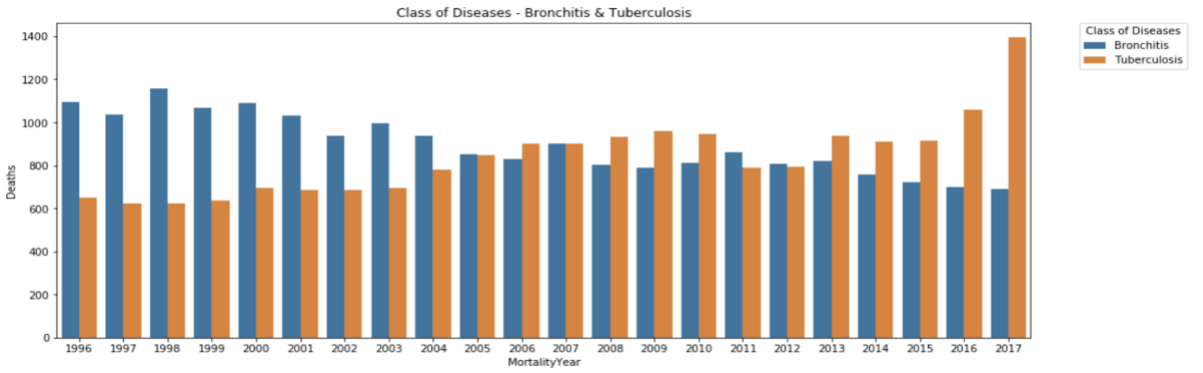


Fig 5: Distribution once the ICD codes are merged

It was observed that with a provided layer of abstraction of just Tuberculosis and Bronchitis few countries were still having either limited or missing data points. This further led to filtering countries where the data was observed for the entire time frame. The countries that the time series modeling was executed for are - United States, Canada, Brazil, Colombia, Argentina, Ecuador, Peru, Mexico, Cuba and Puerto Rico.

Missing Data Points

For the World Bank dataset, we observed that there were few missing values or Nan values. This would become a problem when we are training a time series model. For this we simply used a neural network method for data imputation called the MICE method. This type of imputation works by filling the missing data multiple times. Multiple Imputations are much better than a single imputation as it measures the uncertainty of the missing values in a better way. The MICE method is a more intelligent alternative as supposed to simply filling the missing data with mean, backward or forward fill.

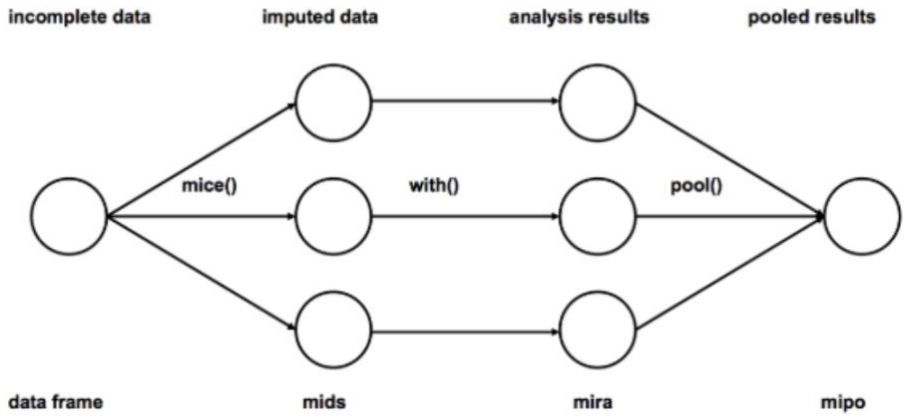


Fig 6: MICE method

Methods

The methods section briefly discusses the various technologies and steps that were executed in order to answer each of the above research questions. For the first research question, the extent of influence of the indicators on death rate was done using simple python libraries such as plotly and bokeh. Using these libraries bar charts, scatter plots and line charts were illustrated in order to check and assess the influence of indicator data on the death rate count. All these graphs were displayed in a jupyter notebook which helped in quick feedback and reusable code.

For the second research question, validity of building a time series model for the provided dataset. Here we used ARIMA and 'auto_arima' models available in the 'statsmodel' and 'pmdarima' libraries in python. The ARIMA model or also known as Autoregressive Integrated Moving Average best works with time series data that exhibit no seasonality. This type of model captures a suite of different standard temporal structures in a time series. Which thereby ends up providing us with a strong prediction tool which can forecast how the time series would look like in the future. The model takes into account all the patterns which ends up being a tool to even flag anomalies. To assess the performance of the model residual mean square error along with AIC value were used as evaluation metrics. Plotting the residual mean square error is also another way to assess if the model is predicting efficiently. At the same time auto_arima model, a stepwise model that checked if the parameters we provided to the function that is the p, q and d values were tuned to be the best performant. This function helps in reinforcing the ARIMA model that was built.

The last research question, to build a machine learning model that can include the indicators for predicting the time series nature of the death rate count. Here we used a VAR model also known as Vector Autoregressive model in order to train the time series. At its core the VAR model performs univariate regression time series modelling. It tries to predict multiple time series variables using the same model. This model will be used to not only train and test the time series data that is available but will also try to forecast the data into the future. The metrics to evaluate VAR model will be Mean absolute percentage error along with AIC and root mean square error. If these values are low, that means that the model is performing with good accuracy.

Vector Auto-Regression Model

We further wanted to build a model that can predict the death count due to the two diseases mentioned in each of the ten countries under consideration. We used a Vector Auto-Regression model for that. The 'statsmodels' package in python was used for the implementation.

The dataset was divided according to country and the class of disease and a model was created for each of these divisions. Totally, there were ten models which predicted the death count from Bronchitis in all the ten countries and there were ten more models that predicted the death count from Tuberculosis. While building the model, three columns 'Country Name', 'Class' and 'Zone' were dropped as there was a model for each country. One-step ahead forecasting was conducted to check how the model performs. This helps us to evaluate the accuracy of the model as we already

know the values but are still predicting. For each country, the data was present for twenty years (1997 to 2017). So, in order to evaluate the model, the first seventeen years (1997 to 2013) were considered as a training set and the model was trained. The death count for the years 2014 to 2017 was forecasted and compared with the actual values. The country Argentina was chosen as the sample to plot the intermediate results. The values and plots mentioned further is in regard to the model for Argentina for Bronchitis.

The final dataset obtained after processing and cleaning was a Non-Stationary dataset and hence the order of differencing was calculated. Augmented Dickey-Fuller Test (ADF Test) was performed to calculate the order of differencing. For each country-case model, the function would check if the time-series is stationary or not and if it is not, it dynamically calculates the order using the ADF test. Once the order of differencing is calculated, the time-series dataset is differenced using the pandas inbuilt method df.diff().

Further to check the closeness of the variables, a Cointegration Test was conducted. Cointegration test helps to establish the presence of a statistically significant connection between two or more time series variables. It was observed that the population is not required in the prediction of the death count.

```
Name   ::  Test Stat > C(95%)    =>   Signif
 ----------------------------------------
Deaths ::  325.81    > 60.0627   =>   True
GDP    ::  157.33    > 40.1749   =>   True
Health_Expenditure ::  51.4      > 24.2761   =>   True
Number_of_Physicians_per1000_people ::  12.86     > 12.3212   =>   True
Population ::  0.16      > 4.1296    =>   False
```

Fig 7: Results of Cointegration Test

Once the model was fit with data, Forecast Error Variance Decomposition (FEVD) was plotted. It indicates the amount of information each variable contributes to the other variables in the autoregression.

The model was used to make predictions and the results were generated, but this result is right now in second differenced form. In order to invert the differencing operation, a function ‘invert_transformation’ was created. This function takes the differenced dataframe and inverts the results to get the forecast to the original scale.

The actual values for the years 1997 to 2017 and the predicted values for the years 2013 to 2017 were plotted using bokeh package. The results are as shown below.

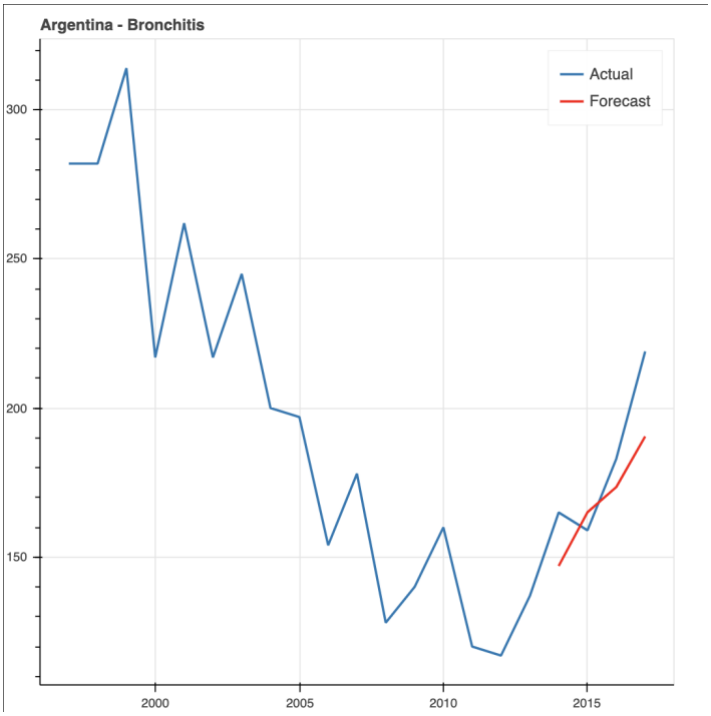


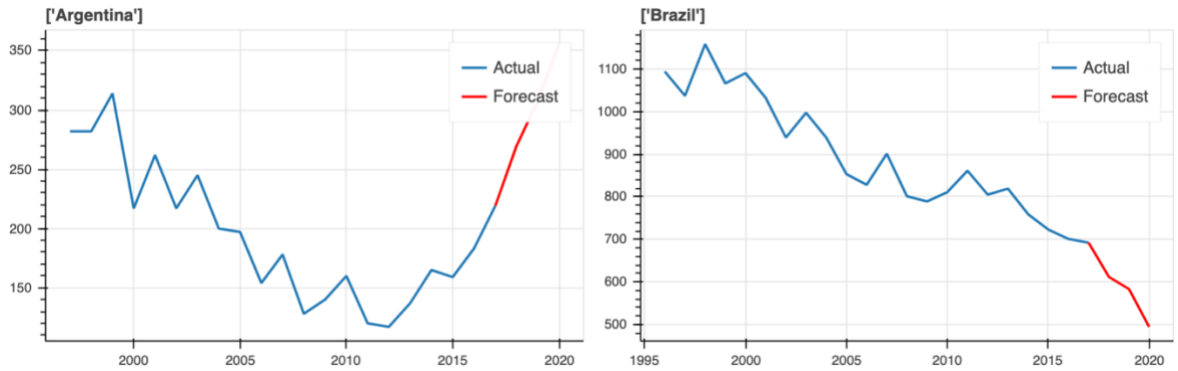
Fig 8: Death count prediction for Argentina (for Bronchitis)

Results

A function was created to measure different evaluation metrics and the forecast accuracy of the model for 'death count' was calculated. The Mean Absolute Percentage Error (MAPE) of the forecast was 8.22%. The Root Mean Squared Error (RMSE) was 17.7671 and the Akaike information criterion (AIC) of the model was 50.07

Using the same functions, country specific models were created for both the diseases and plotted on an interactive graph. Here, the model was trained for the years 1997 to 2017 and the death count was predicted until the year 2020. The death count forecasts are as shown in the below two images.

Bronchitis prediction results for all countries:



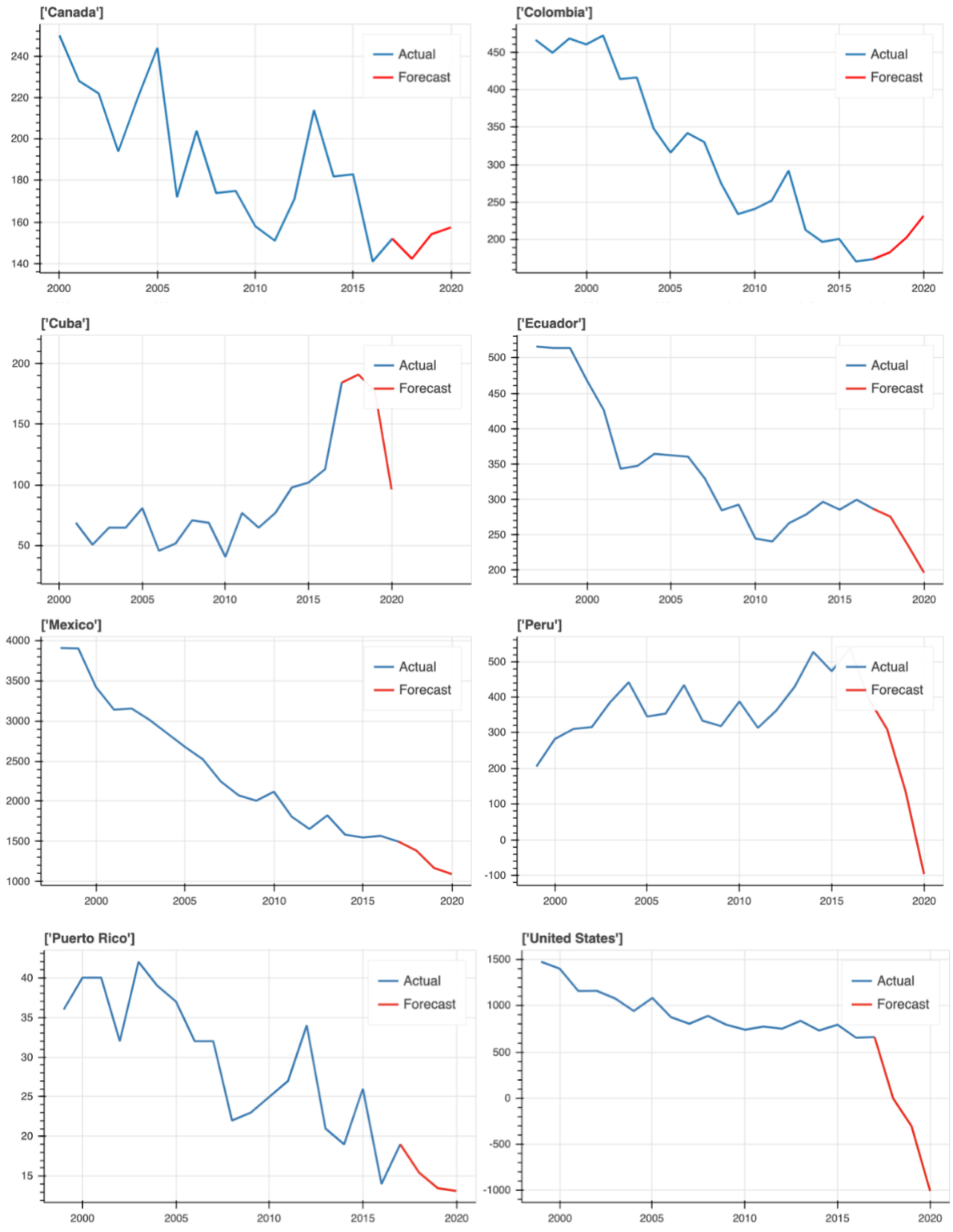


Fig 9: Bronchitis prediction results

Tuberculosis prediction results for all countries:

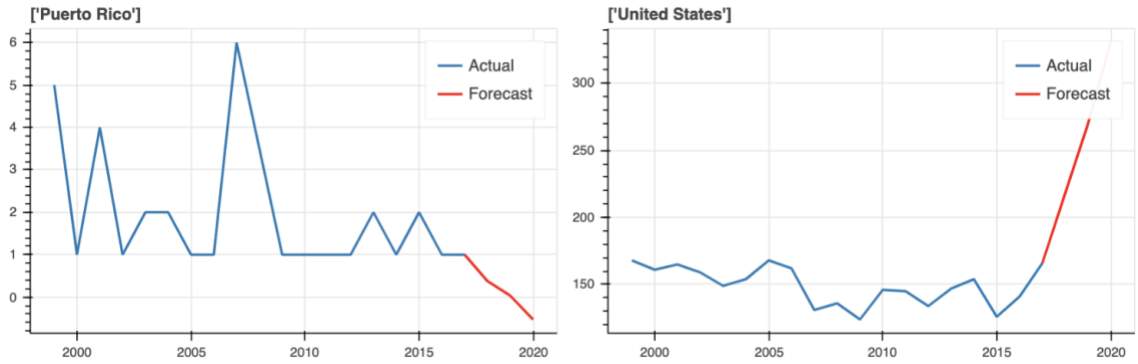


Fig 10: Tuberculosis prediction results

Discussion

The primary goal of the research project was to investigate the death rate in the PAHO mortality data, focusing on acute lower respiratory diseases: Bronchitis and Tuberculosis. We built a time series model to predict the death rate of these two diseases in the Americas region. In order to get a better prediction result, we included several indicators to see their effect on the death rates for the two diseases.

The results of the time series prediction model for Bronchitis showed that there is a decline in the death count in countries like the United States of America and Puerto Rico from the years 2018 to 2020. On the other hand, in countries like Argentina and Colombia there is an increase in the death count from the years 2018 to 2020. This observation can be attributed to the higher GDP and health care expenditure in the countries where a decline in the number of deaths have been observed.

The trends observed in the prediction results for the time series prediction model for Tuberculosis were contrasting to the ones observed for Bronchitis. The death count for countries like the United States of America and Ecuador showed a steep incline from the years 2018 to 2020 as compared to the declining death count observed in the countries like Argentina and Cuba from 2018 to 2020. We believe this is a rather striking result and it could be investigated further by looking at the other indicators as well as looking at the history of Tuberculosis in such countries before 1997.

Our findings and research have a scope limited to only Tuberculosis and Bronchitis with a limited number of indicators used to build the time series prediction model. One way to carry this research further would be to look into other indicators like the number of hospitals in the country and insurance data to get a deeper look at the history of the diseases. This could also be broadened to include other diseases as well.

Conclusion

The main goal of our research focused on building a time series prediction model to forecast the death rate of Bronchitis and Tuberculosis in the Americas. We managed to predict the death rate up to the year 2020 and observed a few contrasting results between the death rate for the two diseases.

The lays the foundation for future research where a more comprehensive data, for example, monthly data for the deaths caused by these two diseases worldwide would be better to build a more accurate and robust predictive model. The focus area of this research can be expanded to include all the diseases in the PAHO mortality dataset and build a similar time series model to predict the death count for all the diseases. It can also be carried out on a global basis to include as many countries affected by deaths caused due to these two diseases.

References

Kang, E. (2017, August 26). Time Series: ARIMA Model. Retrieved from https://medium.com/@kangeugine/time-series-arima-model-11140bc08c6

Prabhakaran, S. (2020, April 28). ARIMA Model – Complete Guide to Time Series Forecasting in Python. Retrieved from https://www.machinelearningplus.com/time-series/arima-model-time-series-forecasting-python/

2020 ICD-10-CM Codes. (n.d.). Retrieved from https://www.icd10data.com/ICD10CM/Codes

Bronchitis Symptoms & Treatment. (n.d.). Retrieved from https://my.clevelandclinic.org/health/diseases/3993-bronchitis

Tuberculosis (TB). (n.d.). Retrieved March 24, 2020, from https://www.who.int/news-room/fact-sheets/detail/tuberculosis

Seabold, Skipper, and Josef Perktold. "Statsmodels: Econometric and Statistical Modeling with Python." Proceedings of the 9th Python in Science Conference. 2010.

Prabhakaran, S. (2020, April 27). Vector Autoregression (VAR) – Comprehensive Guide with Examples in Python. Retrieved from https://www.machinelearningplus.com/time-series/vector-autoregression-examples-python/

GDP per capita (constant 2010 US$). (n.d.). Retrieved from https://data.worldbank.org/indicator/NY.GDP.PCAP.KD?end=2018&start=1960

Current health expenditure per capita (current US$). (n.d.). Retrieved from https://data.worldbank.org/indicator/SH.XPD.CHEX.PC.CD

Physicians (per 1,000 people). (n.d.). Retrieved from https://data.worldbank.org/indicator/SH.MED.PHYS.ZS

Appendix

The jupyter notebooks used for coding and all the data used in this project can be found in the following GitHub link: https://github.com/srijith-srinath/PracticalHealthData