# BIOS 611 Project 1

Michael Jetsupphasuk

October 07, 2020

Note: work in progress. I left things like writing and model-building out and focused on following the prescribed workflow since it seems like that's what you wanted us to get down moreso than the other stuff. Also, I know some of the code and variable names need to be cleaned up. Formatting of plots and tables also need work.

## Introduction

Tuberculosis (TB) is a deadly infectious disease that has persisted for centuries. Treatments created in the mid to late 20th century, along with public health campaigns, lowered disease prevalence and mortality, especially in wealthy countries like the United States where the disease has been essentially eradicated. A vaccine for TB has been available for nearly a century but it does not have 100 percent efficacy so the disease persists despite the vaccine being widely administered. Complicating the public health picture was the emergence of multi-drug resistant TB (MDR-TB) in the 1970s and 80s. Tuberculosis continues to kill many people today.

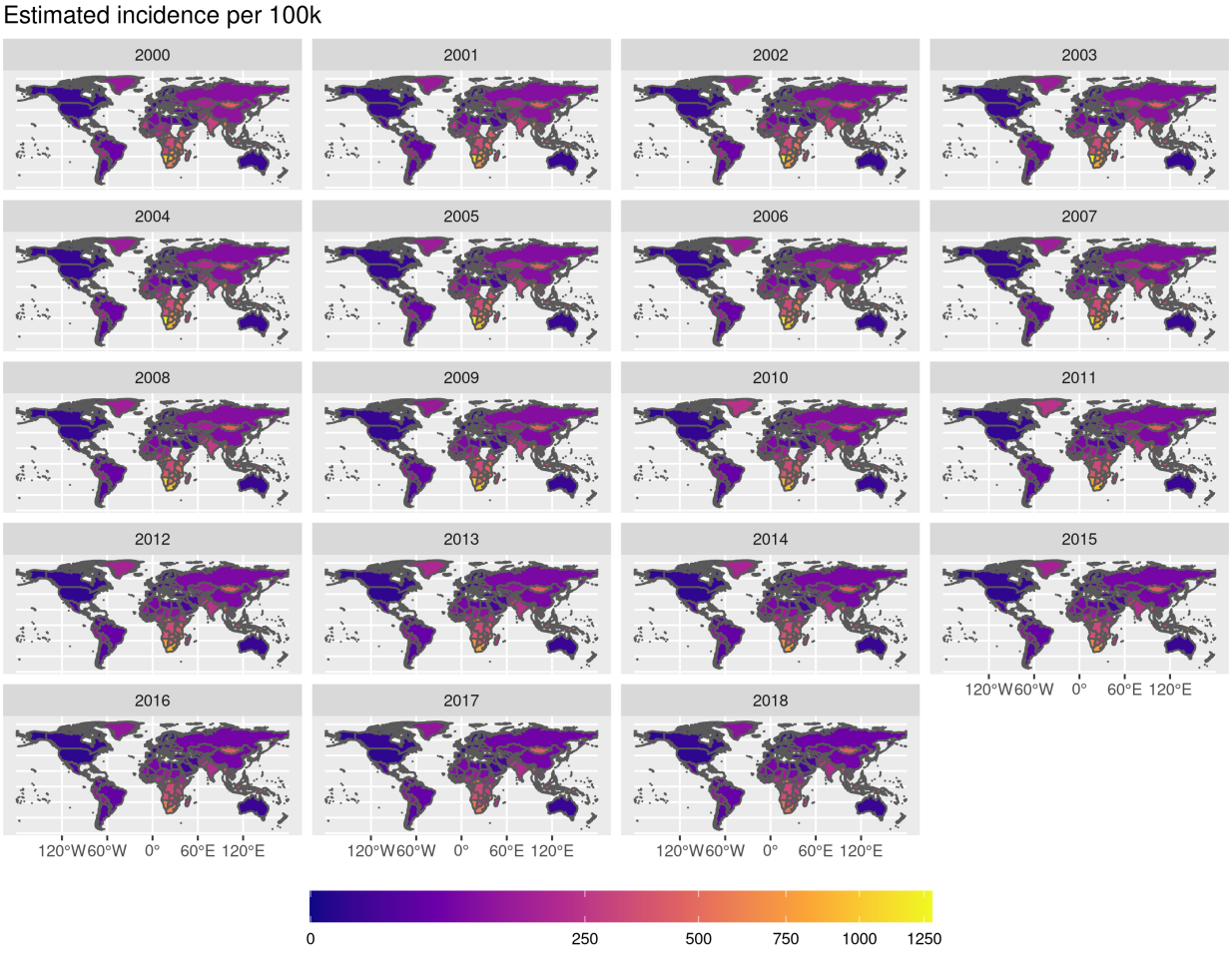
The purpose of this report is to explore recent TB data and get a better sense of incidence and how it varies by country and type of country. I am also interested to see how characteristics of TB vary within the subset of countries with high incidence rates. Are these countries similar in respect to mortality? Cost of treatments? MDR-TB rates?

## Data

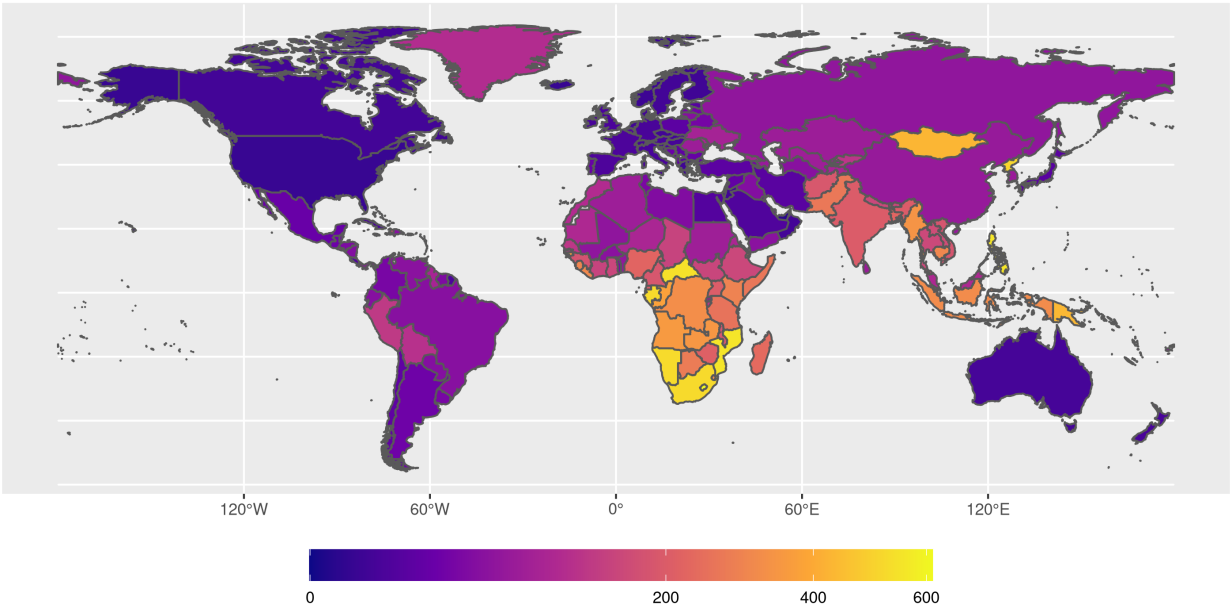
I use TB data from the World Health Organization (WHO). Some of the data is cross-sectional and some is longitudinal by year from 2000-2018.

I also use a variety of country-level indicators from the World Bank. I chose indicators that have little to no missingness and that may reasonably relate to TB incidence, but the decision was somewhat arbitrary. A task for later may be to include all indicators and do feature selection to filter out the un-important ones.

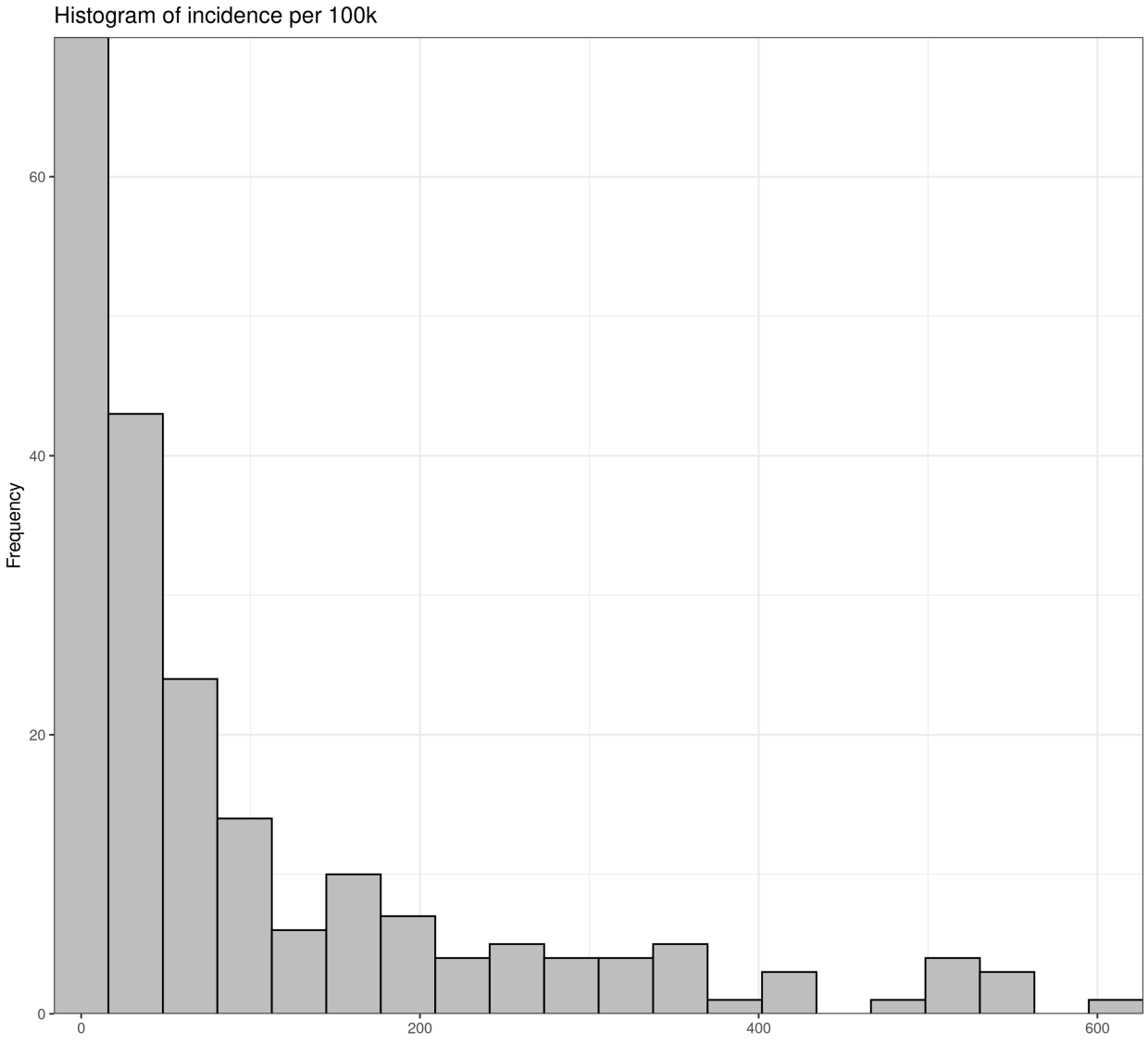
## Mapping TB incidence

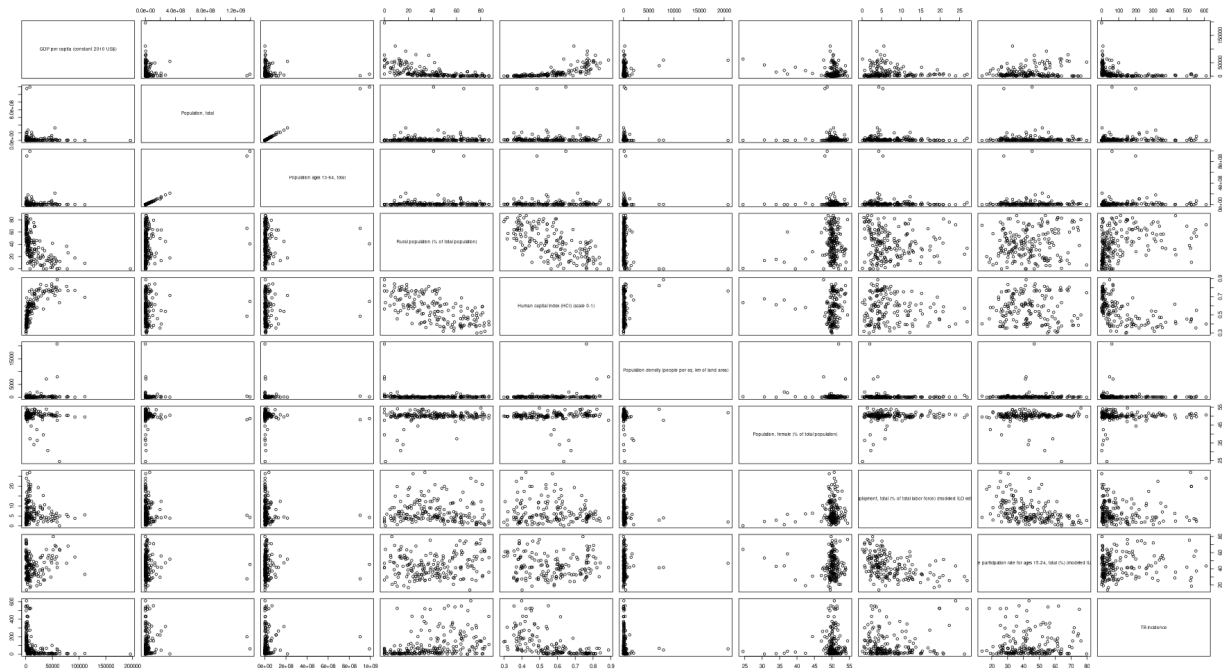


Estimated incidence per 100k, 2018



## Predicting TB incidence with country-level development data





**Feature selection**

Ended up with a lot more missingness than I anticipated. So I had to throw away a lot of the variables I was interested in and was only left with the basic ones, and even then I still had to throw away many countries because of missingness. The result was too few variables to do any meaningful feature selection so I am putting a pin in this part for now. There are some things I could do with imputing the missing data but I'm not sure if it's worth the effort for the purposes of this class. See below for variables used.

| Series Name | Series Code |
|---|---|
| GDP per capita (constant 2010 US$) | NY.GDP.PCAP.KD |
| Population, total | SP.POP.TOTL |
| Population ages 15-64, total | SP.POP.1564.TO |
| Rural population (% of total population) | SP.RUR.TOTL.ZS |
| Human capital index (HCI) (scale 0-1) | HD.HCI.OVRL |
| Population density (people per sq. km of land area) | EN.POP.DNST |
| Population, female (% of total population) | SP.POP.TOTL.FE.ZS |
| Unemployment, total (% of total labor force) (modeled ILO estimate) | SL.UEM.TOTL.ZS |
| Labor force participation rate for ages 15-24, total (%) (modeled ILO estimate) | SL.TLF.ACTI.1524.ZS |

**Model selection**

I didn't hold out a test set here to compare different models (yet). But I did do validation for each model separately. I used linear regression and random forests both with and without the HCI variable. The errors for linear regression were from 5-fold cross-validation and the errors for the random forests were the out-of-bag errors.

| model | include_hci | cv_error |
|---|---|---|
| Linear regression | Yes | 17664.87 |
| Linear regression | No | 18698.60 |
| Random forest | Yes | 11895.96 |
| Random forest | No | 14196.45 |

Coefficients for the linear regression model with HCI:

Table 1:

| | *Dependent variable:* |
|---|---|
| | e_inc_100k |
| NY.GDP.PCAP.KD | 0.0002 |
| | (0.001) |
| SP.POP.TOTL | 0.00000 |
| | (0.00000) |
| SP.POP.1564.TO | −0.00000 |
| | (0.00000) |
| SP.RUR.TOTL.ZS | 0.434 |
| | (0.599) |
| HD.HCI.OVRL | −479.962*** |
| | (111.170) |
| EN.POP.DNST | 0.005 |
| | (0.005) |
| SP.POP.TOTL.FE.ZS | 2.873 |
| | (3.001) |
| SL.UEM.TOTL.ZS | 2.640 |
| | (2.011) |
| SL.TLF.ACTI.1524.ZS | 0.213 |
| | (0.814) |
| Constant | 180.744 |
| | (146.605) |
| Observations | 153 |
| $R^2$ | 0.326 |
| Adjusted $R^2$ | 0.284 |
| Residual Std. Error | 118.101 (df = 143) |
| F Statistic | 7.687*** (df = 9; 143) |
| *Note:* | *p<0.1; **p<0.05; ***p<0.01 |

Coefficients for the linear regression model without HCI:

Table 2:

| | *Dependent variable:* |
|---|---|
| | e_inc_100k |
| NY.GDP.PCAP.KD | −0.002** |
| | (0.001) |
| SP.POP.TOTL | 0.00000** |
| | (0.00000) |
| SP.POP.1564.TO | −0.00000* |
| | (0.00000) |
| SP.RUR.TOTL.ZS | 1.399** |
| | (0.589) |
| EN.POP.DNST | 0.006 |
| | (0.006) |
| SP.POP.TOTL.FE.ZS | −1.284 |
| | (3.011) |
| SL.UEM.TOTL.ZS | 4.217** |
| | (2.095) |
| SL.TLF.ACTI.1524.ZS | 0.914 |
| | (0.845) |
| Constant | 58.933 |
| | (152.421) |
| Observations | 153 |
| $R^2$ | 0.238 |
| Adjusted $R^2$ | 0.196 |
| Residual Std. Error | 125.125 (df = 144) |
| F Statistic | 5.628*** (df = 8; 144) |
| *Note:* | *p<0.1; **p<0.05; ***p<0.01 |

## Characterizing TB outcomes

Still to do.