The analysis of tuberculosis (TB) risk in each microregions of Brazil

Group 16

Introduction

Tuberculosis (TB) is a bacterial infection that spreads from person to person through the air. The germs will apply to the air when the infected person coughs or sneezes. Uninfected people need 2 or 3 breathing of these germs to become infected. TB commonly affects the lungs. According to the World Health Organization(WHO), TB is one of the significant causes of death, ranked the 13th leading cause of death and the second leading infectious killer after COVID-19 (above HIV/AIDS)[1]. In 2021, the number of people who died from TB was 1.6 million (187 thousand people with HIV included). Approximately 10.6 million people worldwide have been infected; six million males, 3.4 million females and 1.2 million youngsters. Brazil ranked 20th in the world most developed TB in 2017 (World Health Organization, 2017)[2]. To help the health authorities in Brazil allocated the resources for hospitals to cope with the TB. This report will analyze TB risk (the rate of TB cases per unit population) across Brazil from 2012-2014 in 557 administrative microregions to identify the riskiest area and suggest the most suitable place to allocate these resources. Explaining the a) spatial, b) temporal, and c) spatio-temporal structure of any systematic (structured) risk and indicating whether socio-economic covariates are related to the TB rate per unit population.

Preparing the data

In the first step of this report, we will start with the wrangling process by investigating the missing values in our data, which shows none of them. The TB rate column has been created using the number of TB cases divided by the people living in each area. We first use the summary function to look at the dataset to gain a brief overview of the data we are working with. The summary tells us that there are very large discrepancies or range of values in all variables. The discrepancies in values for the variables Poverty, Poor_Sanitation and Urbanisation especially caught our attention as high levels of Poverty, low levels of sanitation and low levels of urbanisation may indirectly cause increases in tuberculosis spread. We thought this because these variables are intercorrelated suggesting that high levels of Poverty lead to lower levels of sanitation due to low levels of urbanisation and as such, the spread of tuberculosis is not expected to be hindered and treatment for such disease may not be available. The histogram plotting(Figure 1) is used to observe the overall distribution of the response variable (TB cases per unit population) and any outliers that might appear distribution of the response variable; the results illustrated that between range 0.0001 - 0.0002 provided the highest frequency, while 0.0011 - 0.0012 gives the lowest frequency.

Furthermore, to expand the idea of variables being intercorrelated, we created a correlation matrix (figure 2) using ggpairs to examine the correlation between variables. This correlation matrix further

[1] https://www.who.int/news-room/fact-sheets/detail/tuberculosis

[2] https://books.google.co.uk/books?hl=th&lr=&id=1rQXDAAAQBAJ&oi=fnd&pg=PP1&ots=l93c_w1s0-&sig=J8rzItDTrou4eIgOuFNauNEUErQ&redir_esc=y#v=onepage&q&f=false

proves our idea of variables being intercorrelated as the values suggests the appropriate relationship. Figure 1 shows Unemployment, Timeliness and Urbanization have a solid positive correlation to the TB. The positive coefficients indicate that an increase in these variables is associated with an increase in the TB cases. While Poor Sanitation, Poverty and Illiteracy stand for a weak negative relationship between the infection rate and these three variables. The negative coefficient of them indicates that an increase is associated with a decrease in the TB.

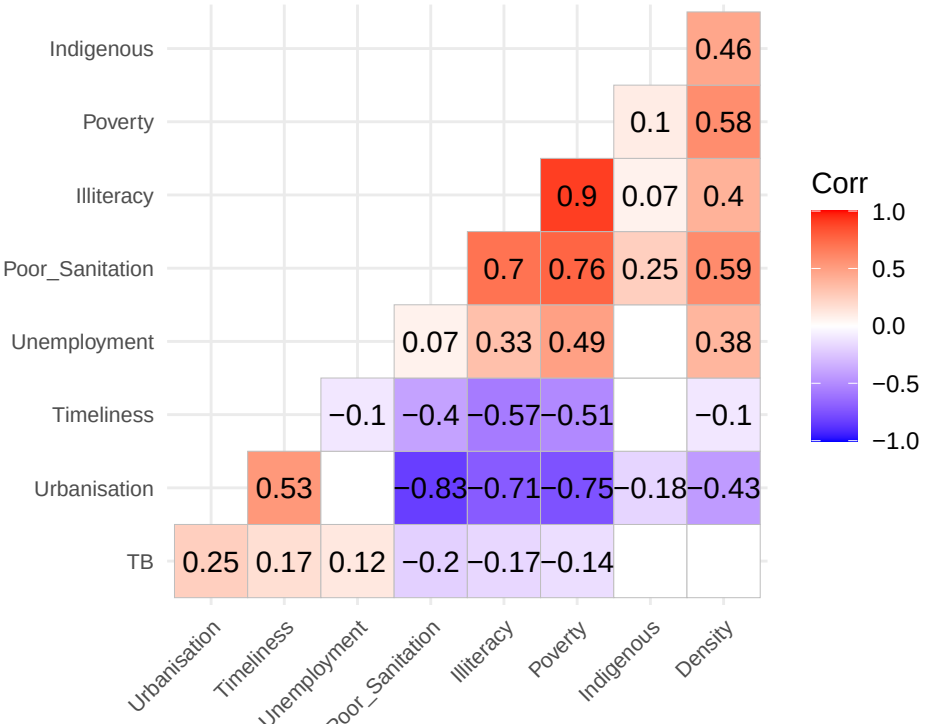

Figure 1: Correlation between variables

Data modelling

Next, we move onto fitting a model since this data is counted data so the Poisson distribution is best suit to model this data. But the data exploration did not show any clear linear patterns and to aviod an overdispersion so we need to move on to using a Generalized Additive Model (GAM). We experimented with a number of families for our dataset, including Negative Binomial, Poisson and Gaussian to see which family fit our data best. To test which family best captures our dataset, we used the Akaike Information Criterion (AIC) function for each family fit. Negative Binomial scored an AIC of 9308.667, Poisson scored an AIC of 14785.51 and Gaussian scored an AIC of 16515.51. Given how the Negative Binomial scored the lowest AIC score, this proved that the dataset would be best fit using a Negative Binomial family. The mathematical form of this model is:

$$
\begin{aligned}
TB\ Rating_i &\sim NB(\mu_i, k) \\
B(TB\ Rating_i) = \mu_i \qquad &and \qquad Var(TB\ Rating_i) = \mu_i + \frac{\mu_i^2}{k} \\
\eta_i &= log(\frac{\mu_i}{offset}) = log(\mu_i) + log(offset) \\
&= log(\mu_i) + f_i(x)
\end{aligned}
$$

While the $TB\ Rating_i$ is the TB rate in i areas of Brazil. η_i is the linked fuction and the $f_i(x)$ is the 'smooth function of the explanatory variable x. The NB is negative binomial distribution with mean μ_i and dispersion parameter k.

In the context of a Generalized Additive Model (GAM) or a Generalized Linear Model (GLM), the offset variable is used to account for a known relationship between the response variable and an independent variable. In this case, the offset($log(Population)$) variable is added to the model to account for the relationship between TB cases and the population size. Including an offset allows the model to focus on the relationship between the response variable (TB cases) and the other independent variables, while adjusting for the population size. The $log(Population)$ part of the offset variable represents the natural logarithm of the population size, which is a common transformation used when dealing with count data, such as the number of TB cases. By including the offset variable in the model, we are essentially modelling the rate of TB cases per unit population, rather than the absolute number of cases. This helps to control for differences in population size across microregions, allowing for a more meaningful comparison of TB risk between different areas. From our first attempt to fit a GAM model, we had some very promising results, most of our p-values were inline with our goal except for the variables Year and Poverty. To overcome this obstacle, we first removed the variable Year, and added a spine function with $k = 10$ to poverty due to its poor results in the $Pr(> |z|)$ values.

Table 1: Summary of GAM Model

term	edf	ref.df	statistic	p.value
s(Illiteracy)	2.656029	3.358818	5.385216	0.1752639
s(Indigenous)	1.004866	1.009575	7.222581	0.0073330
s(Urbanisation)	3.565047	4.474913	20.764592	0.0005748
s(Density)	2.905593	3.477470	27.818940	0.0000114
s(Poverty)	1.751090	2.208499	6.061193	0.0692455
s(Poor_Sanitation)	4.593172	5.495827	33.761653	0.0000055
s(Unemployment)	3.106777	3.850927	48.192741	0.0000000
s(Timeliness)	2.968971	3.728628	46.894455	0.0000000
s(lon,lat)	24.738501	27.885121	295.337367	0.0000000

Once the GAM has been adjusted, we produced an output of the summary of the model (as shown in Table 1). We set the response variable as TB cases and set the predictor variable as the offset variables represented by cubic regressions. From this output we can extract that all variables are of significance except for Illiteracy (0.175264) and Poverty (0.069245) due to their p-values being greater than 0.05, this suggests that adjustments made to our model were slightly inaccurate. Our summary also provides information regarding smooth variables, including the effective degrees of freedom (EDF), Chi-square values, and p-values. The EDF indicates flexibility of the smooth variable, where the higher EDF values the more flexible the variables are. From our output we determine that Poor_Sanitation deemed to be a flexible variable. Lastly, our adjusted R-squared value was 0.843, suggesting that the model explains a significant portion of the variance relative to the response variable. The deviance explained was 56.2%, suggesting that this model does not strongly fit the data relative to the null model.

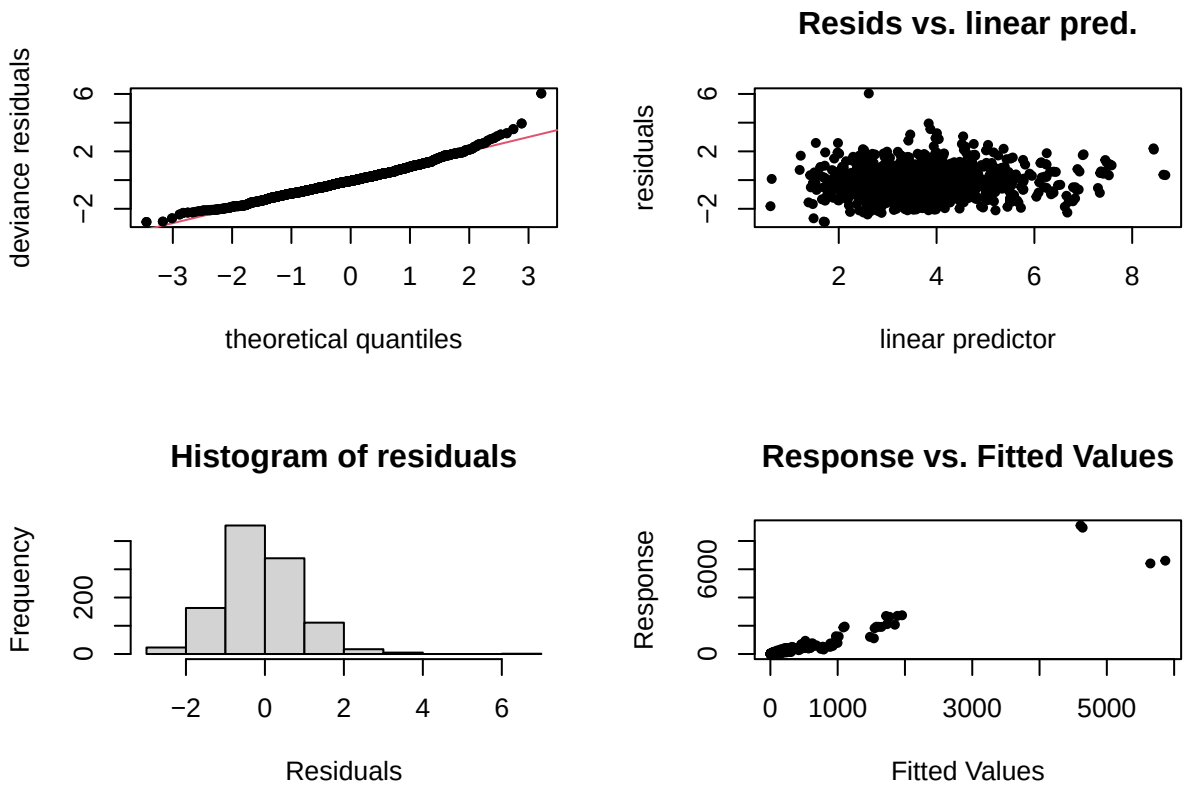

Figure 2: GAM Residuals

Now that we have created our model, we should objectively decide whether our model is flexible or not through residuals and a GAM check. We first start by examining the QQ plot (figure 2, top left), by definition if the model is a good fit against our data then we should notice that the residuals lie on a diagonal line. On the surface level, our QQ plot shows a relatively linear fit suggesting our model being a good fit against the data. However, if we inspect closer, we notice that as the theoretical quantities hit 3, the linearity of the fit starts to deviate, and we see an anomalous increase in deviance residuals. Moving on, in the following Residuals vs Fitted Values graph (figure 2, top right), if the model were to be an appropriate fit for the tuberculosis dataset, then by definition the graph should be evenly scattered, and no pattern should be seen. In the case of our figure, we do not see a particular pattern, suggesting an accurate fit to the data.

Furthermore, we also attempted to predict future values of TB risk based on 2012 and 2013 values of the dataset and we interpreted this prediction as 2014 values for TB risk, and since we have the actual values for 2014 this comparison was relatively simple to compare. Figure 4 shows a map of Brazil with our predicted values based on our model, interpreted as 2014 values, whilst figure 5 shows the actual values for the year 2014. We notice that the two figures were practically identical in Brazil's west coast whilst the east coast showed some differences. Since, our model can predict almost the same data to the original data so our model can works well. Then, we create the Figure 3 which is showing the predicted further TB rate in microregion from our model which could help allocated the resources for hospitals for Brazil.

Figure 3 : Predicted TB Risk (from the model)

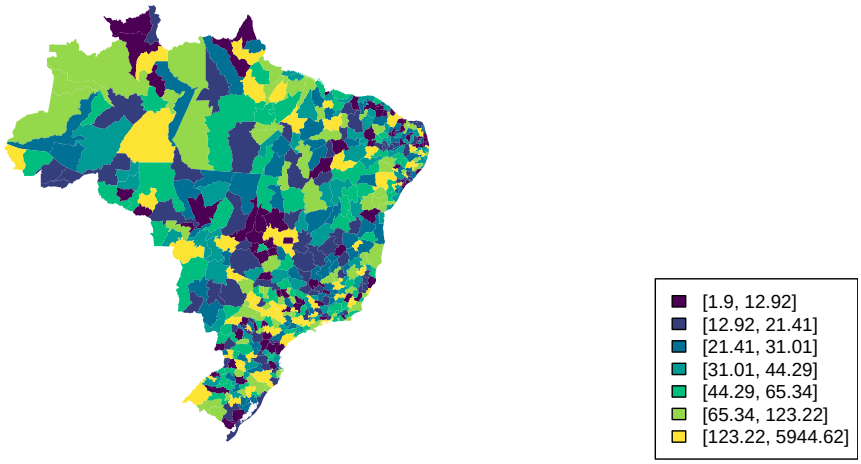

Figure 4 : Predicted TB Risk for 2014 (from trained data)

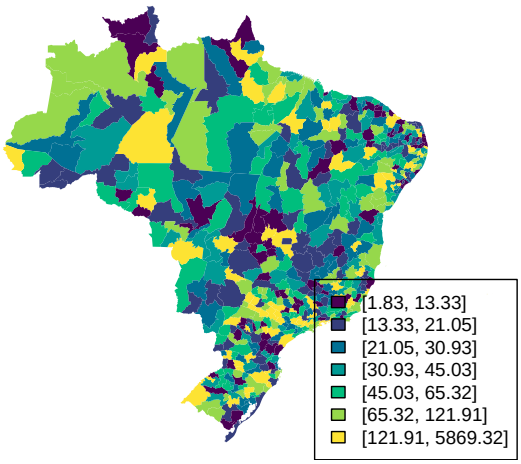

Figure 5 :TB counts for 2014 (from original data)

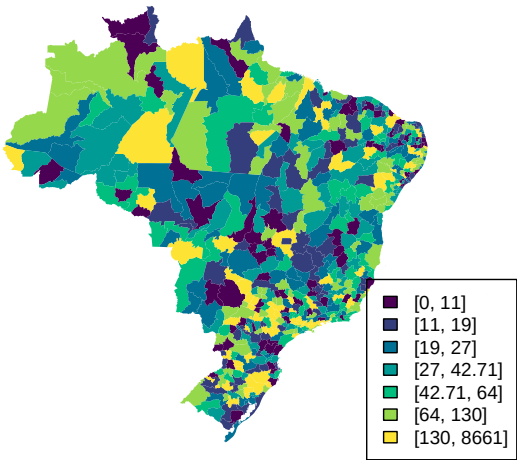

Table 2 : Top 33 highest TB rate

Density	Poor_Sanitation	Unemployment	Timeliness	TB	Population	Region
0.8593049	4.8939574	10.483849	75.53917	1829	2106861	13007
0.6581239	7.9454773	16.575136	62.04482	201	170735	26016
0.6766634	6.5183378	16.094200	60.93366	140	273084	26018
0.6097409	1.1668516	12.302265	46.57070	577	1165310	27011
0.6814175	0.2308221	8.511564	92.04545	746	1801601	35057
0.7283373	1.1006753	8.655322	88.62816	177	465698	35058
0.6990949	0.2235227	9.570883	95.89774	506	1373264	35059
0.6844803	0.6284011	8.546285	80.38721	394	1013390	35060
0.6979410	0.6394426	10.344167	87.22222	518	1344408	35062
0.5832238	0.1123016	9.403939	93.87713	1257	1494079	35063
0.6031208	3.4740495	6.536676	52.69670	735	868141	51017
0.8593049	4.8939574	10.483849	75.53917	1918	2242712	13007
0.6581239	7.9454773	16.575136	62.04482	187	177798	26016
0.6766634	6.5183378	16.094200	60.93366	142	284078	26018
0.6097409	1.1668516	12.302265	46.57070	609	1218197	27011
0.6814175	0.2308221	8.511564	92.04545	786	1876915	35057
0.7283373	1.1006753	8.655322	88.62816	180	488640	35058
0.6990949	0.2235227	9.570883	95.89774	573	1434359	35059
0.6844803	0.6284011	8.546285	80.38721	431	1063302	35060
0.6979410	0.6394426	10.344167	87.22222	456	1405959	35062
0.5832238	0.1123016	9.403939	93.87713	1240	1557125	35063
0.6031208	3.4740495	6.536676	52.69670	936	881902	51017
0.8593049	4.8939574	10.483849	75.53917	2032	2285623	13007
0.6581239	7.9454773	16.575136	62.04482	209	180423	26016
0.6766634	6.5183378	16.094200	60.93366	145	288043	26018
0.6097409	1.1668516	12.302265	46.57070	590	1229071	27011
0.6814175	0.2308221	8.511564	92.04545	729	1892513	35057
0.7283373	1.1006753	8.655322	88.62816	217	495465	35058
0.6990949	0.2235227	9.570883	95.89774	523	1449211	35059
0.6844803	0.6284011	8.546285	80.38721	416	1078149	35060
0.6979410	0.6394426	10.344167	87.22222	450	1421902	35062
0.5832238	0.1123016	9.403939	93.87713	1377	1570532	35063
0.6031208	3.4740495	6.536676	52.69670	847	890277	51017

Table 2 indicates the top highest 33 areas of TB rate at 98 quartiles. It shows that region 13007 has the most significant case, so Brazil's Heath Authorities must focus on this area.

Conclusion

Note that two of Brazil's largest indigenous population reside on the west of Brazil, the Amazonas and Mato Grosso, suggesting that the east coast, where dense cities such as São Paulo, Rio de Janeiro, etc reside, are expected to be more urbanized than the west coast. This could suggest that we have parametrized our model accordingly and accurately, however for more urbanized parts of the country there may be more external factors we are not considering. Perhaps a fixed value of the variables, such as illiteracy, poverty and sanitation levels for each microregion, hence in a more complex and urbanized society do not accurately depict the full story.

Appendix

```r
##Install package
# Package names
packages <- c('mgcv', 'fields', 'maps', 'sp', 'magrittr', 'ggcorrplot',
              'ggstatsplot', 'Boruta', 'dplyr', 'huxtable', 'AICcmodavg',
              'broom', 'corrr','knitr', 'kableExtra', 'readr','tidyr')

# Install packages not yet installed
installed_packages <- packages %in% rownames(installed.packages())
if (any(installed_packages == FALSE)) {
  install.packages(packages[!installed_packages])
}

# Packages loading
invisible(lapply(packages, library, character.only = TRUE))

load('/Users/chatchanokarsuwate/Downloads/datasets_project.RData')


#We can start by looking at the summary statistics of the data
summary(TBdata)

#By checking the data attribute, the data type is divided into integer and numeric,
#so there is no need to do data type conversion. Through the summary data,
#there is no obvious abnormal data and missing data.

#Pre-analysis and pre-processing
corr <- TBdata[, c(1:8)] %>%
  mutate(select(TBdata, TB))
corr1 = cor(corr[,c(1:9)])
p.mat <- cor_pmat(corr, use = "complete", method = 'pearson')
ggcorrplot(corr1,hc.order = TRUE, type = "lower", lab = TRUE, p.mat = p.mat, insig = "blank") +
  labs(title = 'Figure 1: Correlation between variables') +
  theme(plot.title = element_text(hjust = 0.5, vjust = -130, size = 10),
        axis.text.x = element_text(size = 8),
        axis.text.y = element_text(size = 8))
#Let's start by calculating the TB rate as the rate of TB cases per unit population:
library(ggpubr)
# risk is defined as the rate of TB cases per unit population
TBdata$TB_Rating <- TBdata$TB / TBdata$Population
# Histograms for each of the two variables so that their distribution can be observed
p1<-ggscatterstats(data = TBdata, x = "Indigenous", y = "TB_Rating")
p2<-ggscatterstats(data = TBdata, x = "Illiteracy", y = "TB_Rating")
p3<-ggscatterstats(data = TBdata, x = "Urbanisation", y = "TB_Rating")
p4<-ggscatterstats(data = TBdata, x = "Density", y = "TB_Rating")
p5<-ggscatterstats(data = TBdata, x = "Poverty", y = "TB_Rating")
```

```r
p6<-ggscatterstats(data = TBdata, x = "Poor_Sanitation", y = "TB_Rating")
p7<-ggscatterstats(data = TBdata, x = "Unemployment", y = "TB_Rating")
p8<-ggscatterstats(data = TBdata, x = "Timeliness", y = "TB_Rating")
ggarrange(p1, p2, p3, p4, p5, p6, p7, p8, nrow =3, ncol =3)

#confounding variables may not be considered, so we need further analysis to determine
# Create a train data
TBdata_train <- TBdata %>% filter(Year < 2014)

# Include the offset term
mod1 = gam(TB ~ offset(log(Population))+ s(Illiteracy, k = 10, bs = "cr") +
             s(Indigenous, k = 10, bs = "cr") + s(Urbanisation, k = 10, bs = "cr") +
             s(Density, k = 10, bs = "cr") + Poverty + s(Poor_Sanitation, k = 10, bs = "cr") +
             Unemployment + Timeliness + as.factor(Year) + Region + s(lon, lat), data = TBdata_
           family = nb(link = 'log'))
# Fit the GAM
mod2 = gam(TB ~ offset(log(Population))+ s(Illiteracy, k = 10, bs = "cr") +
             s(Indigenous, k = 10, bs = "cr") + s(Urbanisation, k = 10, bs = "cr") +
             s(Density, k = 10, bs = "cr") + Poverty + s(Poor_Sanitation, k = 10, bs = "cr") +
             Unemployment + Timeliness + as.factor(Year) + Region + s(lon, lat), data = TBdata_
           family = poisson)
# Fit the GAM
mod3 = gam(TB ~ offset(log(Population))+ s(Illiteracy, k = 10, bs = "cr") +
             s(Indigenous, k = 10, bs = "cr") + s(Urbanisation, k = 10, bs = "cr") +
             s(Density, k = 10, bs = "cr") + Poverty + s(Poor_Sanitation, k = 10, bs = "cr") +
             Unemployment + Timeliness + as.factor(Year) + Region + s(lon, lat), data = TBdata_
           family = gaussian(link = "identity"))

summary(mod1)
summary(mod2)
summary(mod3)

# 2x2 plot for the residuals
par(mfrow=c(2,2))
#  Runing gam.check on our original model
gam.check(mod1,pch=20)
gam.check(mod2,pch=20)
gam.check(mod3,pch=20)

#Check AIC
AIC(mod1)
AIC(mod2)
AIC(mod3)

#create a new offset model
mod4 = gam(TB ~ offset(log(Population))+ s(Illiteracy, k = 10, bs = "cr") +
             s(Indigenous, k = 10, bs = "cr") + s(Urbanisation, k = 10, bs = "cr") +
```

```
            s(Density, k = 10, bs = "cr")+ s(Poverty, k = 10, bs = "cr") +
            s(Poor_Sanitation, k = 10, bs = "cr") + s(Unemployment, k = 10, bs = "cr") +
            s(Timeliness, k = 10, bs = "cr") + Region + s(lon, lat),
          data = TBdata_train, family = nb(link = 'log'))
# Summarise the model
summary(mod4)
summary(mod1)

#create summary table
mod4 %>%
  tidy() %>%
  kable(caption = 'Table 1: Summary of GAM Model', position = 'center') %>%
  kable_styling(position = 'center', row_label_position = 'c',
                latex_options = 'HOLD_position')  %>%
  row_spec(0, bold = T, color = 'white', background = '#FA8072', align='c')

# Check GAM
# 2x2 plot for the residuals
par(mfrow=c(2,2))
#  Runing gam.check on our original model
gam.check(mod4,pch=20)
title('Figure 2: GAM Residuals', line = -27, outer = TRUE)
gam.check(mod1,pch=20)

#check AIC
AIC(mod4)
AIC(mod1)

#plot map using mod4 on traning dataset
predicted_TB_Risk <- predict(mod4, newdata = TBdata_train, type = "response")
plot.map(x = predicted_TB_Risk, main = "Predicted TB Risk", n.levels = 7, cex = 1)


# PLotting map of cases
plot.map(TBdata$TB[TBdata$Year==2014],n.levels=7,main="TB counts for 2014")

#final model
modf = gam(TB ~ offset(log(Population))+ s(Illiteracy, k = 10, bs = "cr") +
            s(Indigenous, k = 10, bs = "cr") + s(Urbanisation, k = 10, bs = "cr") +
            s(Density, k = 10, bs = "cr")+ s(Poverty, k = 10, bs = "cr") +
            s(Poor_Sanitation, k = 10, bs = "cr") + s(Unemployment, k = 10, bs = "cr") +
            s(Timeliness, k = 10, bs = "cr") + Region + s(lon, lat), data = TBdata,
          family = nb(link = 'log'))

layout(matrix(c(1,1,2,3), 2, 2, byrow = TRUE))
# PLotting map of cases
# PLotting map of cases
```

```r
predicted_TB_Risk <- predict(modf, newdata = TBdata, type = "response")
plot.map(x = predicted_TB_Risk, main = 'Figure 3 : Predicted TB Risk (from the model)', n.level

predicted_TB_Risk <- predict(mod4, newdata = TBdata_train, type = "response")
plot.map(x = predicted_TB_Risk,'Figure 4 : Predicted TB Risk for 2014 (from trained data)', n.l

# PLotting map of cases
plot.map(TBdata$TB[TBdata$Year==2014],'Figure 5 :TB counts for 2014 (from original data)',n.lev

# Create the summarize table
library(formattable)
formattable(selected_columns, list(
  TB = color_bar('#e9c46a'),
  Population = color_bar('#80ed99'),
  Density = color_bar('#f28482'),
  Poor_Sanitation = color_bar('#f28482'),
  Unemployment = color_bar('#f28482'),
  Timeliness = color_bar('#f28482')))
```