MTHM506 Statistical Data Modelling Group Project

Analysis of 2012–2014 Brazil Tuberculosis Data

Group 2

March 2023

1 Introduction

1.1 Problem Statement

This report contains an analysis of tuberculosis (TB) data originating from Brazil using Generalized Additive Models (GAMs). Brazil is divided into 557 administrative microregions, and the available data contains counts of TB cases in each microregion for each of the years from 2012 to 2014.

1.2 Exploratory Analysis of Data and the Problem

The TB data from Brazil includes 1,671 entries or samples with 14 columns of numeric data types that specify the characteristics of each sample. The columns are: Indigenous, Illiteracy, Urbanisation, Density, Poverty, Poor Sanitation, Unemployment, Timeliness, Year, TB, Population, Region, lon, and lat. The columns 'lon' and 'lat' stand for longitude and latitude. The dataset has no missing values. The 'region' is stored as a continuous variable despite being a factor variable. Nonetheless, changing it depends on the task at hand. The collection includes coordinates that describe the precise geospatial locations of the microregions listed. The next section gives a detailed exploration of the data.

1.3 Data Exploration

Values quoted are in % unless stated otherwise. An in-depth analysis of the datasets reveals that the mean and median values for the indigenous population are low, but the maximum value is 50, which suggests that there are individual areas where the indigenous population is concentrated. It is important to explore these areas to see if the poverty, sanitation and TB incidence differ significantly from the baseline. The mean and median illiteracy rates are only 14 and 11, respectively. However, the maximum value of 41 suggests that there are specific areas with significant parts of the population lacking access to education. This may suggest that the area has poor sanitation, but it needs further examination. There are still some areas that are less urbanised, where the TB occurrence may differ from the more urbanised areas. The median, and minimum values for urbanisation are 72 and 22, respectively, suggesting that most areas are highly urbanised.

Based on the population density data, most locations show that one person can have their individual room. But the highest value of 1.6 highlights some places with very high population densities, which sharply increase the rate of TB transmission. The distribution of the poverty data suggests that each district has different poverty levels, with only a limited number of districts where poverty is not a significant issue. Although the mean value for poor sanitation is around 13, the maximum score of 58 indicates that some districts have poor sanitation and substantial disease risk. Although average unemployment rates are low, a maximum value of 20 indicates that some isolated regions may experience severe economic hardship and potentially significant morbidity rates. With a minimum value of 0, notification timeliness data has a fairly wide range.

Timeliness, Unemployment, and Urbanization approximately follow a Normal Distribution with a few outliers, while the remainder consists of multimodal distributions. The above suggests that employing semi-parametric or non-parametric models to demonstrate the relationship between the target and predictors would be useful. The target variable in this study is a risk, defined as $\frac{TB}{Population}$, whereas the remaining variables are possible predictors. As demonstrated in Figure 5, most features in the dataset exhibit some correlaton. It is vital to note that some characteristics are anticipated to have positive correlations (tuberculosis versus population, density versus poverty) and vice-versa. Specifically, population density, poverty, health conditions, unemployment and notification timeliness are likely to be high due to the high population density, low economic share per capita, high

poverty rate, and high unemployment rate. See Figure 5 for a correlogram of the 8 socio-economic covariates.

2 Model

We want to model the count of cases TB_i by actually modelling ρ_i (the *rate* of TB incidence per unit of population) using

$$TB_i \sim Pois(\lambda_i = z_i\rho_i)\ TB_i \text{ indep.}$$
$$log(\lambda_i) = log(z_i) + log(\rho_i)$$

where TB_i is the count of TB cases. We are using the canonical link function - log. z_i is the total population, which is taken as an offset. Model $log(\rho_i)$ as

$$log(\rho_i) = \sum_{j=1}^{8} f_j(x_{i,j})$$
$$f(x_i) = \sum_{k=1}^{q} \beta_k b_k(x_i)$$

where $x_{i,j}$ is the jth covariate (out of 8 socio-economic covariates) for the ith instance/datum in the dataset, $f(\cdot)$ is a smooth function of said covariate and $b_k(\cdot)$ is a basis function with k knots. Hence, the model boils down to

$$TB_i \sim Pois(\lambda_i = z_i\rho_i)\ TB_i \text{ indep.}$$
$$log(\lambda_i) = log(z_i) + \sum_{j=1}^{8}\sum_{k=1}^{q} \beta_{j,k} b_{j,k}(x_{i,j})$$

Looking at the distribution of the residuals of the model, we can see that the data is clearly far too overdispersed to be modelled by a Poisson distribution, which has a fixed dispersion parameter (see Figure 2). Even with 60 knots per smooth term the model does not seem to have enough flexibility which may be another indicator that a Poisson model is unsuitable for the data. The residuals vs fitted plot fans out, indicating that the model does not have enough flexibility to fit well. The edf is also close to the maximum degrees of freedom, and increasing the number of knots does not resolve the problem - see results for `gam.check(model_poisson)` in Appendix. We propose the conventional alternative to the Poisson - the Negative Binomial model. Doing so, leads to a drop in the AIC. See Table 1 Appendix for a showcase of different model configurations and their associated AIC.

1

$$TB_i \sim NB(\lambda_i, \sigma_i^2)\ TB_i \text{ indep.}$$
$$\lambda_i = z_i\rho_i;\ \sigma_i^2 = \lambda_i + \frac{\lambda_i^2}{\phi}$$
$$log(\lambda_i) = log(z_i) + \sum_{j=1}^{8}\sum_{k=1}^{q} \beta_{j,k} b_{j,k}(x_{i,j})$$

where ϕ is a dispersion parameter, later estimated by the `gam` function in R.

When having a look at the relationship between the squared residuals and the fitted values, one sees that the relation is not exactly quadratic, but rather close to 0. This would reflect the relation between model variance and the expected value in a Gaussian Distribution Model (additional evidence is provided by the Residuals vs. Fitted plot in Figure 2). However, fitting a Gaussian model leads to very skewed residuals, indicating that the data is apparently not Gaussian. So, the model distribution is changed to Negative Binomial with the same parameterisation except for the feature that the count of TB cases is now Negative Binomial distributed with mean λ_i as described above.

Given this base model, we investigate whether all given socio-economic covariates are needed to explain the response or whether there exists a model with fewer parameters. The p-value for the smooth term of Illiteracy points towards it not being statistically significant. Poverty, although not statistically insignificant, has the second largest p-value. These terms are sequentially dropped and the resulting model is checked against the original model via a Likelihood Ratio Test conducted

using the `anova` function in R (see Table 2). We find that leaving out Illiteracy does not alter the model at a 5% level of significance, whereas taking out both Poverty and Illiteracy does. So, in the following, we use a model with all of the socio-economic covariates except Illiteracy. Note that this converts our linear predictor to

$$log(\lambda_i) = log(z_i) + \sum_{j=1}^{7} \sum_{k=1}^{q} \beta_{j,k} b_{j,k}(x_{i,j})$$

This leaves us with a model with AIC = 14,391.19 and 43.9% of deviance explained. Running `gam.check()` lets us analyse the residual plots (see Figure 3) and examine the basis functions for the model. The QQ plot tells us that the model fails to predict well on the upper and lower ends of the response variable. Increasing the knots to 20 per covariate leads to marginal improvement with 44.9% deviance explained and hence, it is discarded. More efficient extensions can be to add 1) spatial, 2) temporal and 3) spatio-temporal covariates.

First, we will try adding spatial terms. The spatial model adds a smoothed term which is a function of the longitude and the latitude. A bivariate function is used as it makes sense to assume that there are more cases at certain locations (defined by the interaction between latitude and longitude) than others, in comparison to there being more cases at locations with a certain longitude for any latitude, or vice-versa. Hence, our linear predictor is now

$$log(\lambda_i) = log(z_i) + \sum_{j=1}^{7} \sum_{k=1}^{q} \beta_{j,k} b_{j,k}(x_{i,j}) + \sum_{k=1}^{q} \beta_k b_k(lon_i, lat_i)$$

Using this model with the regular `s` smoother function from the `mgcv` package leads to a model that can explain 56.4% of the deviance and has a slightly lower AIC of 14,013.13. The QQ plot still points to the upper and lower tails being incorrectly predicted. At the cost of significantly more computation, using a tensor product smooth `te` on the bivariate spatial term with 20 knots allows us to make a decent improvement. This gets us to 69.9% deviance explained. The QQ plot looks considerably better with only a few problematic instances at the top and bottom quantiles (see Figure 4).

We contest this with an extension on the model with only socio-economic covariates, but instead of adding spatial terms, we add the temporal dimension `Year`. The linear predictor becomes

$$log(\lambda_i) = log(z_i) + \sum_{j=1}^{7} f_{2012,j}(x_{i,j}) \times x_{2012} + \sum_{j=1}^{7} f_{2013,j}(x_{i,j}) \times x_{2013} + \sum_{j=1}^{7} f_{2014,j}(x_{i,j}) \times x_{2014}$$

where the new terms x_{2012}, x_{2013}, x_{2014} are indicator variables equating to 1 if `Year` is respectively 2012, 2013, 2014 and 0 otherwise. Exercising some shorthand, it can be expressed as

$$log(\lambda_i) = log(z_i) + \sum_{t=2012}^{2014} \sum_{j=1}^{7} f_{t,j}(x_{i,j}) \times x_t$$

where x_t is now the indicator variable for `Year`. A slightly separate approach can be tested with `Year` as a covariate instead of a grouping variable. In that case, the linear predictor would be

$$log(\lambda_i) = log(z_i) + \sum_{j=1}^{7} f_{t,j}(x_{i,j}) + \sum_{t=2012}^{2014} \beta_t x_t$$

Neither of the temporal formulations show much increase in deviance explained. Their QQ plots are also much worse than the spatial model, showing gross deviations on high as well as low quantiles. Finally, we create a spatio-temporal model, including both `Year` and `lon` and `lat`. Its linear predictor is formulated as

$$log(\lambda_i) = log(z_i) + \sum_{t=2012}^{2014} \left(\sum_{j=1}^{7} \sum_{k=1}^{q} \beta_{t,j,k} b_{t,j,k}(x_{t,i,j}) + \sum_{k=1}^{q} \beta_{t,k} b_{t,k}(lon_{i,t}, lat_{i,t}) \right) \times x_t$$

This is a model which includes the term for the location, and estimates a functional relation for each year and each explaining variable. The AIC of this model does not drop much when compared

to the spatial model (see Table 1). Reasons for this may be that having 3 levels of factors is not enough granularity to discern any effect from the temporal dimension. Naive estimates would point towards the spread being higher in winter months as TB spreads through inhaling tiny droplets from coughs or sneezes. Having more granular data at the season or month level could bring highlight any temporal patterns, if they do exist.

So, the spatial model (given that it is simpler) is the model we choose to best explain the ratio of TB cases per capita. To recall, it is formulated as

$$log(\lambda_i) = log(z_i) + \sum_{j=1}^{7}\sum_{k=1}^{q} \beta_{j,k} b_{j,k}(x_{i,j}) + \sum_{k=1}^{q} \beta_k b_k(lon_i, lat_i)$$

Considering the spatial model, it fits well even though the largest residuals are higher than expected from the model distribution. For districts that have a high number of cases, the predictor does not seem as accurate. But the highest residuals do not arise when the ratio of TB cases per capita is extraordinarily high, but rather when the absolute number of TB cases is high (see residuals vs response). The variance of the model still seems too low for extreme values. There are some predicted values in that high segment of response values (absolute number of TB cases) where the prediction for the response is lower than the actual value. Conversely, in the low segment of response values, there are some instances with a higher prediction than the actual value. Using this model, we predict the rate of TB per 100,000 inhabitants. See Figure 1 for a plotted comparison of predicted and actual TB rates.

3 Critical Review

As demonstrated in Figure 4, the QQ plots show deviations on the farthest quantiles, which demonstrate that the predictions do not cover the full range of the data. The independence assumption of the `Year` variable can be questioned wih notions of temporal and spatial correlation, thus providing justification for its exclusion. The data points of a certain region in 2012, 2013 and 2014 may not be fully independent of one another as it is likely that the conditions in that region have not substantially changed. The model seems unnecessarily complex if we add additional smooth terms for each covariate grouped on `Year` as it hardly brings any additional explanatory power. Parametric coefficients of the `Year` factor in `spatio.temporal.model.2` barely differ from each other at around -8.4 for `Year` $= 2012$, and varying by 0.004 and -0.038 respectively for 2013 and 2014. These respectively correspond to multipliers of 1.004 and 0.963 on the response scale (See `summary(spatio.temporal.model.2)` in Appendix). Another possible violation of the independence assumption arises from spatial correlation. The regions which are located closely to one another are mutually dependent on one another in terms of the number of TB cases as well as the socio-economic determinants of the spread of infectious diseases.

4 Conclusions

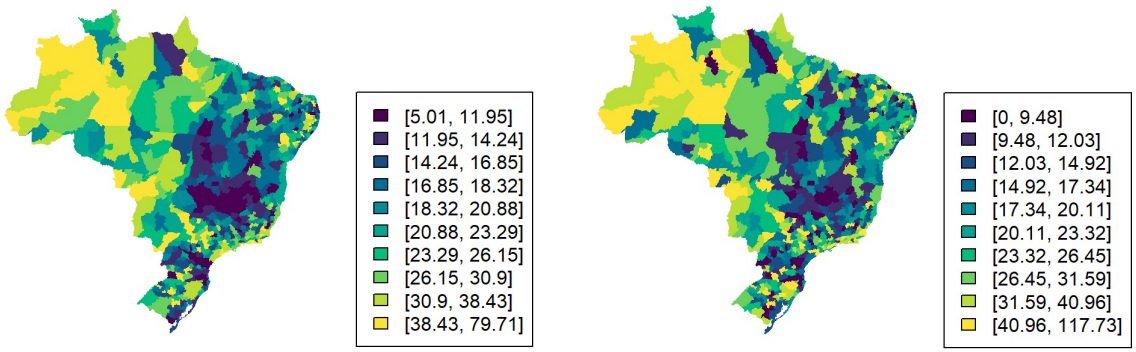
Based on the correlogram, we can conclude that no single socio-economic covariate has much linear correlation with TB incidence, but illiteracy, urbanisation, poverty, sanitation, unemployment and timeliness of notification are all weakly correlated with TB incidence. There are stronger correlations between these socio-economic covariates. Increase in illiteracy, poverty, unemployment and poor sanitation will simultaneously correlate to decreases in urbanisation and timeliness of notification, all contributing to increases in TB incidence. Unsurprisingly, poverty is strongly correlated with several socio-economic covariates. Sanitation and urbanisation are also strongly correlated, implying that investments can be directed twoards improvement of urban infrastructure and health resources have a greater impact on reducing TB incidence. According to our predicted TB incidence map, Brazil's central region has a lower incidence overall. This presents an opportunity to learn from what has worked in this region and how improvements and resources can be further directed to Brazil's north-west state of Amazonas along with localised areas in the south and east, particularly in the southern and south-eastern states of Sao Paolo and Rio de Janeiro. It is worth noting that these areas of Sao Paolo and Rio de Janiero could benefit from assistance from neighbouring regions with fewer cases of TB (see Figure 6).

5 Reference

Wood, S. N. (2017). Generalized Additive Models: An Introduction with R (2nd ed.). CRC Press.

6 Appendix

6.1 Figures



(a) Predicted (b) True

Figure 1: Predicted (a) and True (b) rates of TB per 100,000 inhabitants. North-western parts of the country as well as parts near the south exhibit high TB incidence per capita. These would roughly correspond to the states (*estados*) of Amazonas in the north-west, Sao Paolo in the south and Rio de Janeiro on the south-east coast. Refer to state map in Figure 6.

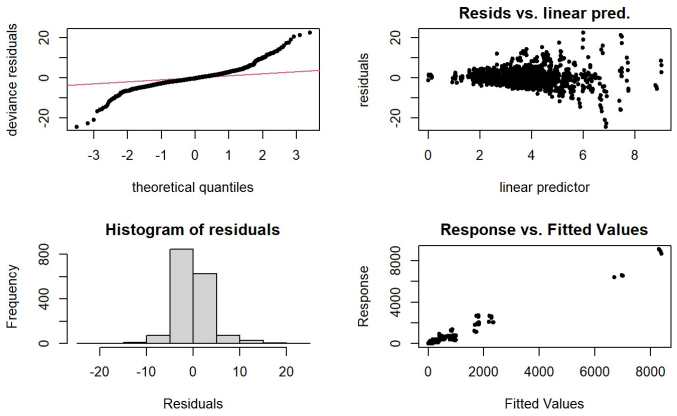


Figure 2: `gam.check( )` results for `model_poisson`. The data is overdispersed as can be seen from both the QQ plot and the fanning of the Resids vs linear pred. plot. The histogram of residuals is also quite different from a gaussian distribution.

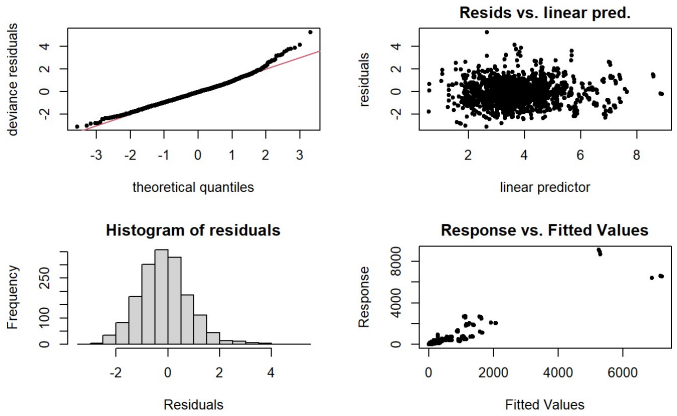


Figure 3: `gam.check( )` results for `model_nb_2`. Although there is an improvement on the Poisson model, the model has difficulty in correctly predicting values on the upper and lower quantiles.

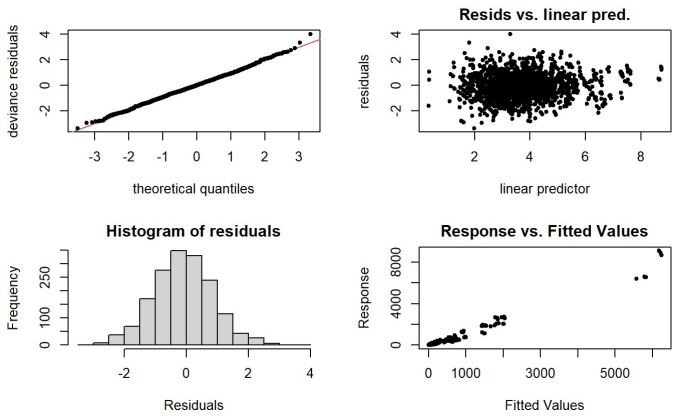


Figure 4: `gam.check( )` results for `spatial.model.2`. This is the model that has finally been used as temporal additions on top of this provide very little additional explanatory power.

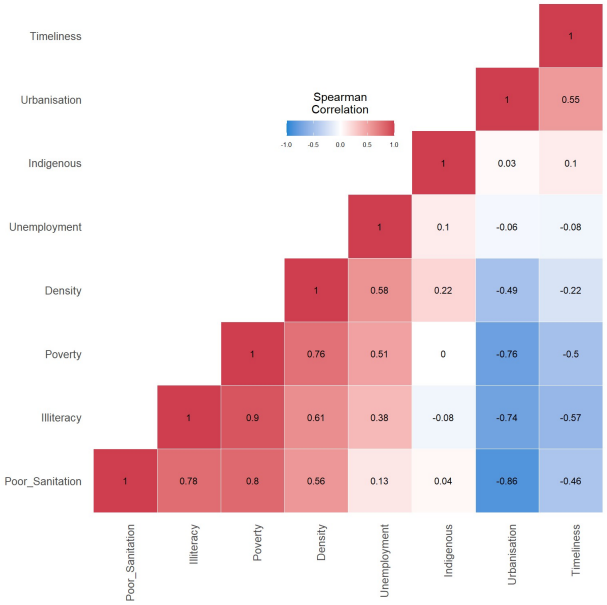


Figure 5: Correlogram shows socio-economic covariates with highest positive and negative correlations.

Figure 6: State map of Brazil.

6.2 Tables

Model	AIC	Deviance Explained
model_poisson	34047.36	66.9%
model_nb	14391.19	43.9%
model_nb_2	14389.56	43.9%
model_nb_3	14405.00	43.0%
model_nb_time	14389.70	44.0%
temporal.model	14390.43	44.0%
spatial.model	14013.13	56.4%
spatial.model.2	13650.8	69.9%
spatial.model.temporal	14227.12	52.2%
spatial.model.temporal.2	13647.93	70.1%

Table 1: Comparison of models.

Model 1	Model 2	p-value	Conclusion
model_nb	model_nb_2	0.5556	Illiteracy dropped as a covariate
model_nb_2	model_nb_3	0.001861	Poverty is retained
spatial.model.2	model_nb_2	< 2.2e-16	Spatial covariates added into the model

Table 2: Pairwise ANOVA tests for model comparison.

6.3 Code

```
# Load Data
load("datasets_project.RData")
# Import Libraries
library(mgcv) # required for GAM

#fit poisson model with socio-economic variables
model_poisson <- gam(formula = TB ~ offset(log(Population)) + s(Indigenous)
+ s(Illiteracy) + s(Urbanisation) + s(Density) + s(Poverty) + s(Poor_Sanitation)
+ s(Unemployment) + s(Timeliness),
data = TBdata,
family = poisson(link = 'log')
)

#add flexibility
model_poisson.more.knots <- gam(formula = TB ~ offset(log(Population)) + s(Indigenous, k = 60)
+ s(Illiteracy, k = 60) +  s(Urbanisation, k = 60) + s(Density, k = 60)
+ s(Poverty, k = 60) + s(Poor_Sanitation, k = 60) + s(Unemployment, k = 60)
+ s(Timeliness, k = 60),
data = TBdata,
family = poisson(link = 'log')
)

#fit negative binomial model with  socioeconomic
model_nb <- gam(formula = TB ~ offset(log(Population)) + s(Indigenous)
+  s(Illiteracy) + s(Urbanisation) + s(Density) + s(Poverty) + s(Poor_Sanitation)
+ s(Unemployment) + s(Timeliness),
data = TBdata, family = nb(link = 'log')
)

#fit a linear relation between squared residuals and prediction to see whether another model describes
#the variance-fitted values relation better
summary(lm(log(model_nb$residuals^2) ~ log(predict(model_nb, type = 'response'))))

#drop Illiteracy
```

```r
model_nb_2 <- gam(formula = TB ~ offset(log(Population)) + s(Indigenous) +  s(Urbanisation) + s(Density)
+ s(Poverty) + s(Poor_Sanitation) + s(Unemployment) + s(Timeliness),
data = TBdata,
family = nb(link = 'log')
)

# Likelihood ratio test
anova.gam(model_nb_2, model_nb, test = 'F') # p-value is over 0.05
# The models are statistically indistinguishable

model_nb_3 <- gam(formula = TB ~ offset(log(Population)) + s(Indigenous) +  s(Urbanisation) + s(Density)
+ s(Poor_Sanitation) + s(Unemployment) + s(Timeliness),
data = TBdata,
family = nb(link = 'log')
)
# Likelihood ratio test
anova.gam(model_nb_3, model_nb_2, test = 'F')# p-value is less than 0.05
# The models are statistically different. Poverty should not be excluded.

### Model chosen (with socio-economic covariates) is the negative binomial without Illiteracy
### as the effect of illiteracy cannot be reliably stated to be non-zero

#### Introducing temporality as a grouping variable
#Temporal model
model_nb_time <- gam(formula = TB ~ offset(log(Population)) + s(Indigenous, by = Year)
+ s(Urbanisation, by = Year) + s(Density, by = Year) + s(Poverty, by = Year) + s(Poor_Sanitation, by = Year)
+ s(Unemployment, by = Year) + s(Timeliness, by = Year),
data = TBdata,
family = nb(link = 'log')
)

#### Temporality as a covariate
TBdata$Year.asFactor <- factor(TBdata$Year)

temporal.model <- gam(formula = TB ~ offset(log(Population)) + s(Indigenous)
+ s(Urbanisation) + s(Density) + s(Poor_Sanitation) + s(Unemployment) + s(Poverty)
+ s(Timeliness) + Year.asFactor,
data = TBdata ,
family = nb(link = 'log')
)

### Adding spatial covariates
spatial.model <- gam(formula = TB ~ offset(log(Population)) + s(Indigenous)
+ s(Urbanisation) + s(Density) + s(Poor_Sanitation) + s(Unemployment) +s(Poverty)
+ s(Timeliness) + s(lon , lat),
data = TBdata ,
family = nb(link = 'log')
)

### Using separate smoothers
spatial.model.2 <- gam(formula = TB ~ offset(log(Population)) + s(Indigenous)
+ s(Urbanisation) + s(Density) + s(Poor_Sanitation) + s(Unemployment) + s(Poverty)
+ s(Timeliness) + te(lon , lat , k = 20),
data = TBdata ,
family = nb(link = 'log')
)

# Check if this model is significantly different from one with only socio-economic covariates
anova.gam(spatial.model.2, model_nb_2, test = 'LRT')

#### Spatio-temporal model
spatio.temporal.model <- gam(formula = TB ~ offset(log(Population)) + s(Urbanisation, by = Year.asFactor)
+ s(Density, by = Year.asFactor) + s(Poverty, by = Year.asFactor)
```

```
+ s(Poor_Sanitation, by = Year.asFactor) + s(Timeliness, by = Year.asFactor)
+ s(Unemployment, by = Year.asFactor) + te(lon,lat, by = Year.asFactor), data = TBdata, family = nb(link = 'log'))

### Spatio-temporal model (with Year as parametric covariate)
spatio.temporal.model.2 <- gam(formula = TB ~ offset(log(Population)) + s(Indigenous)
+ s(Urbanisation) + s(Density) + s(Poor_Sanitation) + s(Unemployment) + s(Poverty)
+ s(Timeliness) + te(lon , lat , k = 20) + Year.asFactor,
data = TBdata ,
family = nb(link = 'log')
)

### Prediction
preds <- round(predict(spatial.model.2 , newdata = TBdata , type = 'response'),4)
preds_rate <- preds/TBdata$Population*100000

### Plot
plot.map(preds_rate , n.levels = 10)
```

6.4 Model Summaries and Residual Checks

```
# check summary
summary(model_poisson)

Family: poisson
Link function: log

Formula:
TB ~ offset(log(Population)) + s(Indigenous) + s(Illiteracy) +
    s(Urbanisation) + s(Density) + s(Poverty) + s(Poor_Sanitation) +
    s(Unemployment) + s(Timeliness)

Parametric coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -8.449827   0.004199   -2012   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
                    edf Ref.df Chi.sq p-value
s(Indigenous)     8.961  8.999  569.4  <2e-16 ***
s(Illiteracy)     8.989  9.000 2704.0  <2e-16 ***
s(Urbanisation)   8.900  8.996 1490.4  <2e-16 ***
s(Density)        8.985  9.000 1758.4  <2e-16 ***
s(Poverty)        8.956  8.999 1470.2  <2e-16 ***
s(Poor_Sanitation) 8.979  9.000 1327.0  <2e-16 ***
s(Unemployment)   8.993  9.000 2423.5  <2e-16 ***
s(Timeliness)     8.352  8.864  600.7  <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =  0.976   Deviance explained = 0.669
UBRE = 13.899  Scale est. = 1         n = 1671

model_poisson$aic
34047.36

# Excerpt from residual check
gam.check(model_poisson)

                    k`  edf k-index p-value
```

```
s(Indigenous)      9.00 8.96    0.39  <2e-16 ***
s(Illiteracy)      9.00 8.99    0.41  <2e-16 ***
s(Urbanisation)    9.00 8.90    0.41  <2e-16 ***
s(Density)         9.00 8.98    0.39  <2e-16 ***
s(Poverty)         9.00 8.96    0.39  <2e-16 ***
s(Poor_Sanitation) 9.00 8.98    0.40  <2e-16 ***
s(Unemployment)    9.00 8.99    0.39  <2e-16 ***
s(Timeliness)      9.00 8.35    0.43  <2e-16 ***
---

#check summary
summary(model_nb_2)

Family: Negative Binomial(6.146)
Link function: log

Formula:
TB ~ offset(log(Population)) + s(Indigenous) + s(Urbanisation) +
    s(Density) + s(Poverty) + s(Poor_Sanitation) + s(Unemployment) +
    s(Timeliness)

Parametric coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -8.42863    0.01094  -770.6   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
                    edf Ref.df Chi.sq  p-value
s(Indigenous)     1.518  1.833  21.13 2.08e-05 ***
s(Urbanisation)   6.610  7.752  23.73  0.00167 **
s(Density)        4.578  5.667 147.64  < 2e-16 ***
s(Poverty)        5.771  6.945  21.36  0.00394 **
s(Poor_Sanitation) 6.119  7.293  76.07  < 2e-16 ***
s(Unemployment)   5.776  6.977  64.21  < 2e-16 ***
s(Timeliness)     4.106  5.103  66.42  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =   0.86   Deviance explained = 0.439
-REML = 7234.9  Scale est. = 1         n = 1671

model_nb_2$aic
14389.56

# Excerpt from residual check
gam.check(model_nb_2)

                    k`  edf k-index p-value
s(Indigenous)      9.00 1.52    0.49  <2e-16 ***
s(Urbanisation)    9.00 6.61    0.50  <2e-16 ***
s(Density)         9.00 4.58    0.50  <2e-16 ***
s(Poverty)         9.00 5.77    0.49  <2e-16 ***
s(Poor_Sanitation) 9.00 6.12    0.50  <2e-16 ***
s(Unemployment)    9.00 5.78    0.50  <2e-16 ***
s(Timeliness)      9.00 4.11    0.56  <2e-16 ***
---

# check summary
summary(spatial.model.2)

Family: Negative Binomial(12.246)
Link function: log
```

```
Formula:
TB ~ offset(log(Population)) + s(Indigenous) + s(Urbanisation) +
    s(Density) + s(Poor_Sanitation) + s(Unemployment) + s(Poverty) +
    s(Timeliness) + te(lon, lat, k = 20)

Parametric coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -8.467186   0.008485  -997.9   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
                       edf  Ref.df  Chi.sq  p-value
s(Indigenous)        3.700   4.346   19.53 0.000922 ***
s(Urbanisation)      5.188   6.221   52.90  < 2e-16 ***
s(Density)           4.107   5.012   38.40 1.58e-06 ***
s(Poor_Sanitation)   5.367   6.412   27.55 0.000174 ***
s(Unemployment)      4.132   5.125   79.61  < 2e-16 ***
s(Poverty)           6.716   7.729   42.69  < 2e-16 ***
s(Timeliness)        2.445   3.053   46.59  < 2e-16 ***
te(lon,lat)        139.341 174.137 1088.66  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =  0.926   Deviance explained = 0.699
-REML = 6987.8  Scale est. = 1         n = 1671

spatial.model.2$aic
13650.8

#Excerpt from residual check
gam.check(spatial.model.2)

                       k`    edf k-index p-value
s(Indigenous)        9.00   3.70    0.63  <2e-16 ***
s(Urbanisation)      9.00   5.19    0.61  <2e-16 ***
s(Density)           9.00   4.11    0.63  <2e-16 ***
s(Poor_Sanitation)   9.00   5.37    0.61  <2e-16 ***
s(Unemployment)      9.00   4.13    0.62  <2e-16 ***
s(Poverty)           9.00   6.72    0.61  <2e-16 ***
s(Timeliness)        9.00   2.45    0.66  <2e-16 ***
te(lon,lat)        399.00 139.34    0.65  <2e-16 ***
---

# check summary
summary(spatio.temporal.model.2)

Family: Negative Binomial(12.299)
Link function: log

Formula:
TB ~ offset(log(Population)) + s(Indigenous) + s(Urbanisation) +
    s(Density) + s(Poor_Sanitation) + s(Unemployment) + s(Poverty) +
    s(Timeliness) + te(lon, lat, k = 20) + Year.asFactor

Parametric coefficients:
                    Estimate Std. Error  z value Pr(>|z|)
(Intercept)       -8.4532889  0.0144206 -586.197   <2e-16 ***
Year.asFactor2013 -0.0005816  0.0201595   -0.029    0.977
Year.asFactor2014 -0.0417054  0.0202005   -2.065    0.039 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Approximate significance of smooth terms:
                       edf  Ref.df  Chi.sq  p-value
s(Indigenous)        3.702   4.347   19.56 0.000912 ***
s(Urbanisation)      5.197   6.230   53.05  < 2e-16 ***
s(Density)           4.107   5.011   38.45 1.89e-06 ***
s(Poor_Sanitation)   5.374   6.418   27.61 0.000170 ***
s(Unemployment)      4.145   5.140   80.03  < 2e-16 ***
s(Poverty)           6.722   7.734   42.72  < 2e-16 ***
s(Timeliness)        2.460   3.070   46.59  < 2e-16 ***
te(lon,lat)        139.707 174.516 1093.31  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =  0.926   Deviance explained = 0.701
-REML = 6991.1  Scale est. = 1         n = 1671

spatio.temporal.model.2$aic
13647.93

#Excerpt from residual check
gam.check(spatio.temporal.model.2)

                       k`    edf k-index p-value
s(Indigenous)        9.00   3.70    0.63  <2e-16 ***
s(Urbanisation)      9.00   5.20    0.60  <2e-16 ***
s(Density)           9.00   4.11    0.62  <2e-16 ***
s(Poor_Sanitation)   9.00   5.37    0.61  <2e-16 ***
s(Unemployment)      9.00   4.15    0.61  <2e-16 ***
s(Poverty)           9.00   6.72    0.61  <2e-16 ***
s(Timeliness)        9.00   2.46    0.66  <2e-16 ***
te(lon,lat)        399.00 139.71    0.64  <2e-16 ***
---
```