A Statistical Analysis of Socio-Economic and Spatio-Temporal Factors on Tuberculosis Risk in Brazil

Introduction

The dataset TBdata contains the yearly number of Tuberculosis (TB) cases in each of Brazil's 557 microregions during the years 2012-2014. The dataset also contains the coordinates and population of each microregion along with 8 socio-economic variables: Indigenous, illiteracy, urbanisation, density, poverty, sanitation, unemployment and timeliness. I have fit a Generalized additive model (GAM) to the dataset, with TB as the response variable, to understand the socio-economic and spatio-temporal factors that effect the risk of TB in Brazil.

Preliminary Data Analysis

Plotting the relationship between the socio-economic variables with the rate of TB, shown in Fig. 1 and 2, helps to provide an understanding of the relationships between the variables and TB risk. Fig 3. shows the rate of TB in each microregion, highlighting how TB rates differ significantly across Brazil. I turned the temporal variable, year, into a categorical variable and plotted the distributions of the TB rates for each year, shown in Fig. 4. The distributions are very similar, showing that the TB rates remain fairly constant over the time period.

Model Development

The number of TB infections in each microregion is a form of count data. Therefore, I first created a GAM using the Poisson distribution to model the response variable. Model analysis showed that this model did not fit the data well compared to the saturated model due to the restricted variance of the Poisson distribution. Therefore, I used a negative binomial distribution in the GAM as it can also model count data and has a dispersion term that ensures the model fits the data.

During model development, the socio-economic variable illiteracy was shown to be statistically insignificant and was therefore dropped from the model. The categorical variable year was also shown to be statistically insignificant and was also dropped from the model.

Model Description

The final model, displayed below, has an offset of the number of hundred thousand people in each microregion. Therefore, the model predictions will be the rate of TB per hundred thousand people, which provides us with a proxy measure for the risk of TB. The model uses a log-link function to ensure that the mean TB rate is always positive.

$$Y_i \sim NB(\mu_i, \theta) \quad (Y_i \text{ independent})$$

$$\log(\mu_i) = \log(\texttt{HunThou}_i) + f_1(\texttt{Indigenous}_i) + f_2(\texttt{Urbanisation}_i) + f_3(\texttt{Density}_i) +$$

$$f_4(\texttt{Poverty}_i) + f_5(\texttt{Poor-Sanitisation}_i) + f_6(\texttt{Unemployment}_i) + f_7(\texttt{Timeliness}_i) +$$

$$f_8(\texttt{lon}_i, \texttt{lat}_i)\texttt{Year}_{1i} + f_9(\texttt{lon}_i, \texttt{lat}_i)\texttt{Year}_{2i} + f_{10}(\texttt{lon}_i, \texttt{lat}_i)\texttt{Year}_{3i}$$

One dimensional smooth functions of the statistically significant socio-economic variables are used in the link function. The variables lon and lat provide the coordinates of each microregion. The link function has a two dimensional function of the variables lon and lat for each year so that spatio-temporal factors are incorporated into the model. The smooth functions were fit to the data using cubic splines and given enough knots to ensure they have adequate flexibility, as shown in Fig. 5.

A χ^2 goodness of fit test showed that the model fits the data well compared to the saturated model. Fig. 6 and 7 show that the model residuals are roughly normally distributed around the mean, showing that the model is not missing any structure within the data.

Results

Socio-Economic Factors

The variables poverty, timeliness, sanitation levels and unemployment measure different aspects of economic development. Fig. 8-11 show the smooth functions of these variables. The graphs show that areas with low levels of economic development have higher risks of TB. Fig. 12 and 13 show that an increase in levels of urbanisation and population density increase the risk of TB. This shows that TB rates will be higher in cities, particularly in the densely populated areas. Fig. 14 shows that microregions with higher proportions of indigenous people have significantly higher rates of TB.

Spatial Effects

Fig. 15-17 show the spatial effects on TB risk for the years 2012-2014, and the average spatial effect over the three years is shown in Fig. 18. On average, with the effects of the other covariates removed, the risk of TB is higher in Western Brazil, with the central western and most southern areas of Brazil having particularly high effects on TB risks. The mountainous areas North of Guanabara Bay along with central and northern Brazil have negative effects on TB risk.

Spatio-Temporal Effects

Fig. 19 and 20 show the differences in the spatial effects between years 2012-2013 and 2013-2014 respectively. The heatmap in Fig. 19 shows that North and Central Brazil experienced the largest increase in TB risk during 2012-2013. The risk of TB was reduced in South and East Brazil. The heatmap in Fig. 20 shows that, during 2013-2014, largely the opposite effect was experienced, with increased TB risk in the North and Centre of Brazil and reduced TB risk in the South and East of Brazil.

Critical Review

The effects of socio-economic factors were modelled as one-dimensional smooth functions in the model. Therefore, potentially significant interactions between the factors were not included. The model could be extended to include a range of higher dimensional smooth functions.

The changes in the spatial effects in 2013-2014 were almost the opposite of that in 2012-2013, therefore no long term spatio-temporal trends were identified. However, it is difficult to draw strong conclusions when the dataset contains measurements at only three time points. Spatio-temporal analysis could be improved if there was access to data over a longer time period.

Figures

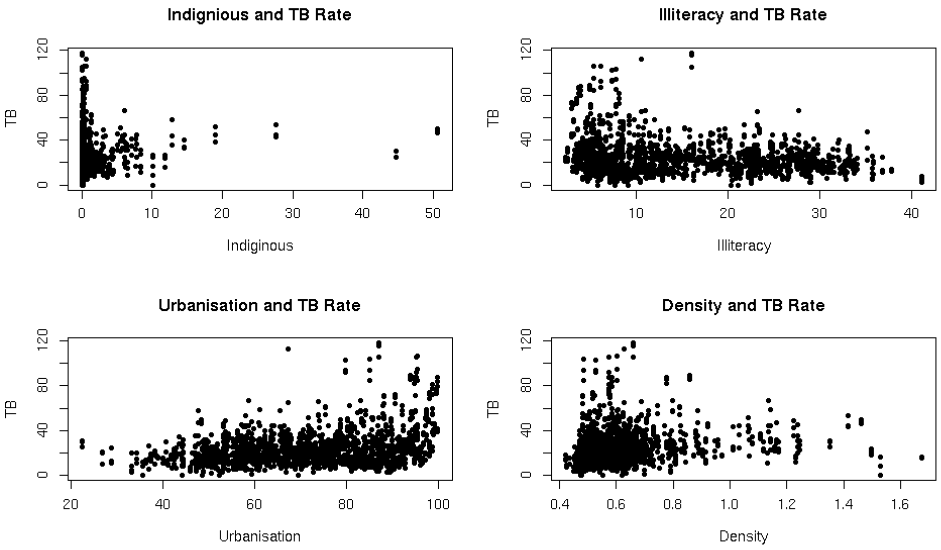


Figure 1: Scatter plots of four socio-economic variables and the rate of TB.

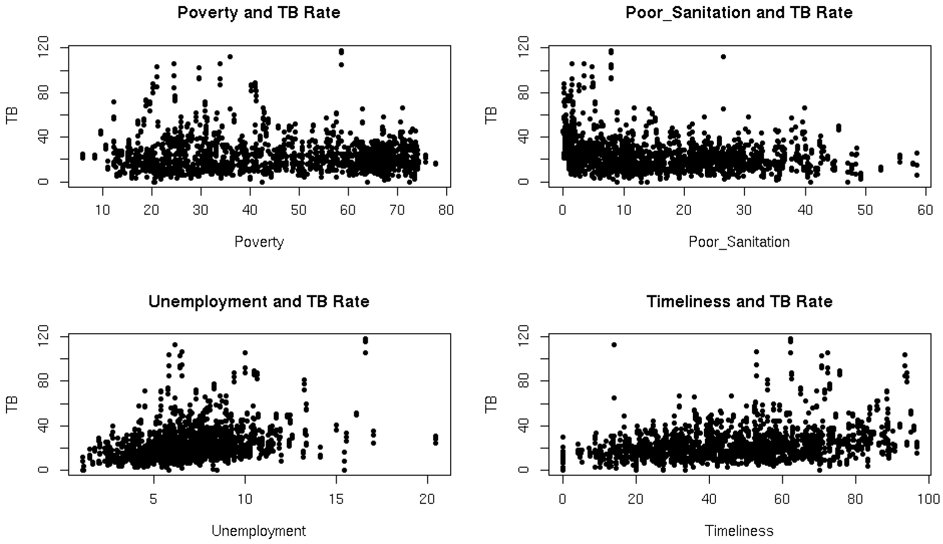


Figure 2: Scatter plots of the other four socio-economic variables and the rate of TB.

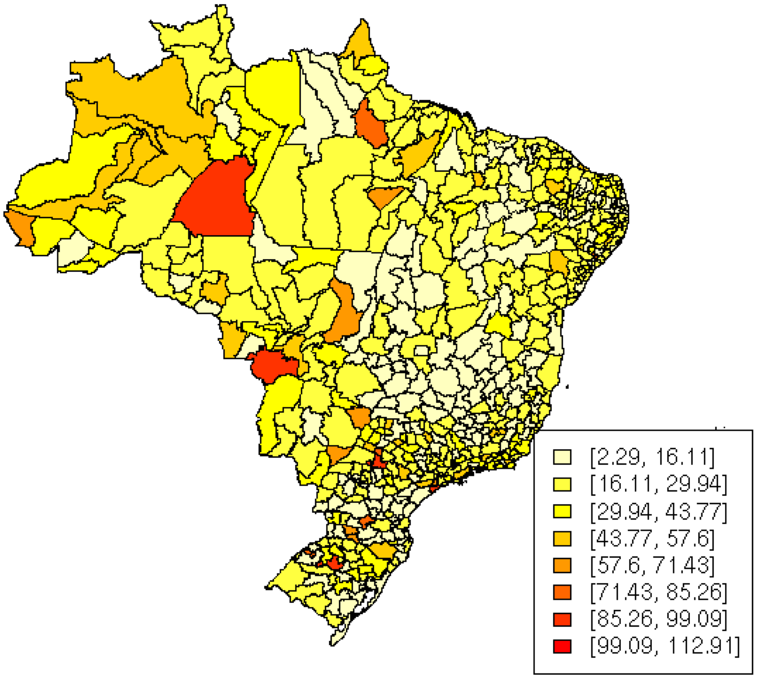


Figure 3: Rate of TB per hundred thousand people over years 2012-2014.

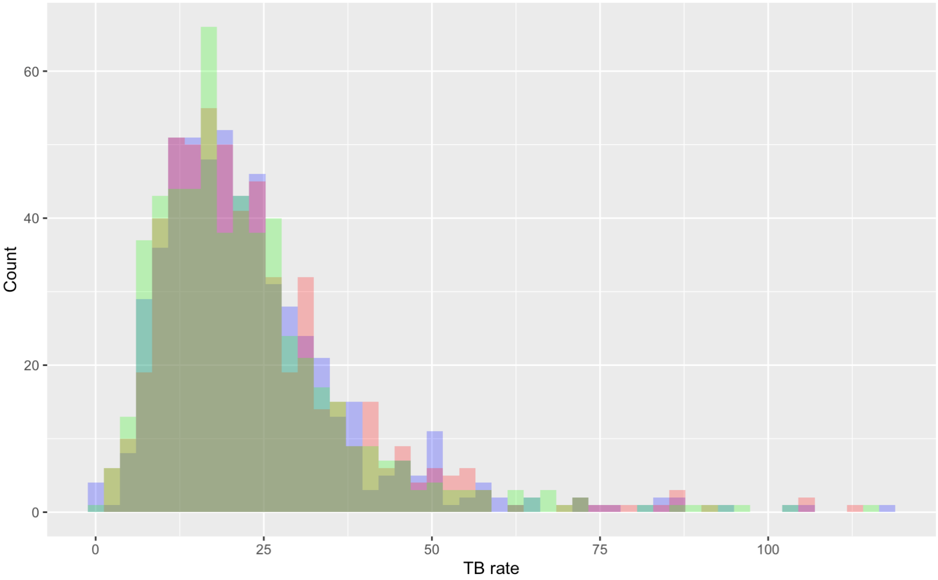
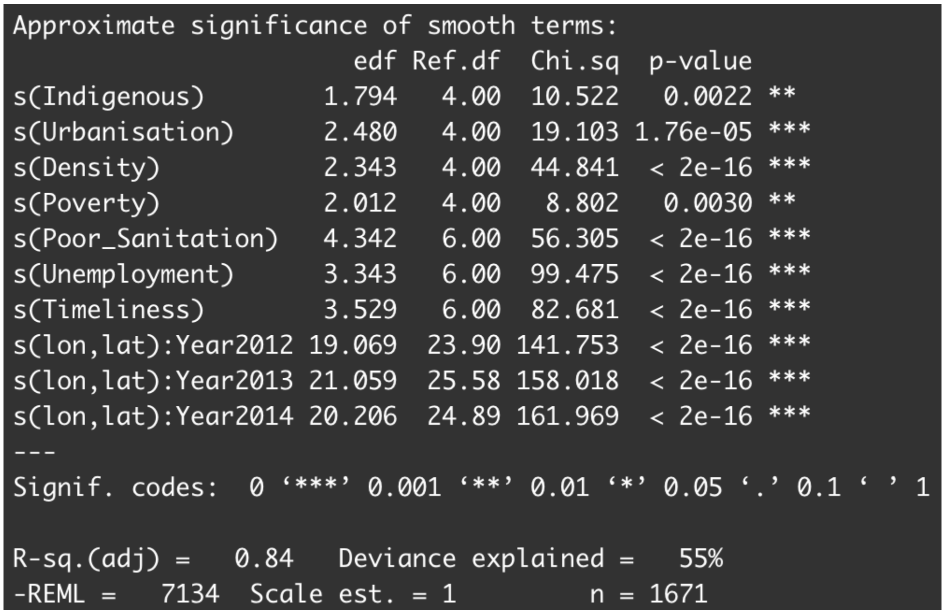


Figure 4: The distributions of TB risk for each year.



```
Approximate significance of smooth terms:
                       edf Ref.df  Chi.sq  p-value    
s(Indigenous)        1.794   4.00  10.522   0.0022 ** 
s(Urbanisation)      2.480   4.00  19.103 1.76e-05 ***
s(Density)           2.343   4.00  44.841  < 2e-16 ***
s(Poverty)           2.012   4.00   8.802   0.0030 ** 
s(Poor_Sanitation)   4.342   6.00  56.305  < 2e-16 ***
s(Unemployment)      3.343   6.00  99.475  < 2e-16 ***
s(Timeliness)        3.529   6.00  82.681  < 2e-16 ***
s(lon,lat):Year2012 19.069  23.90 141.753  < 2e-16 ***
s(lon,lat):Year2013 21.059  25.58 158.018  < 2e-16 ***
s(lon,lat):Year2014 20.206  24.89 161.969  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =   0.84   Deviance explained =   55%
-REML =   7134  Scale est. = 1         n = 1671
```

Figure 5: The summary of the model. Obtained by using summary(Model1).

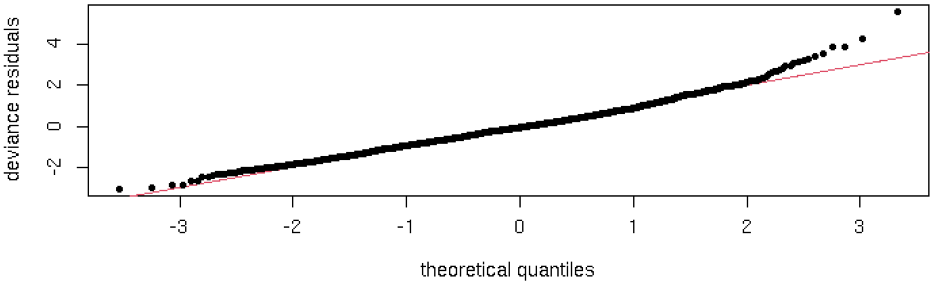


Figure 6: QQ plot of the model residuals.

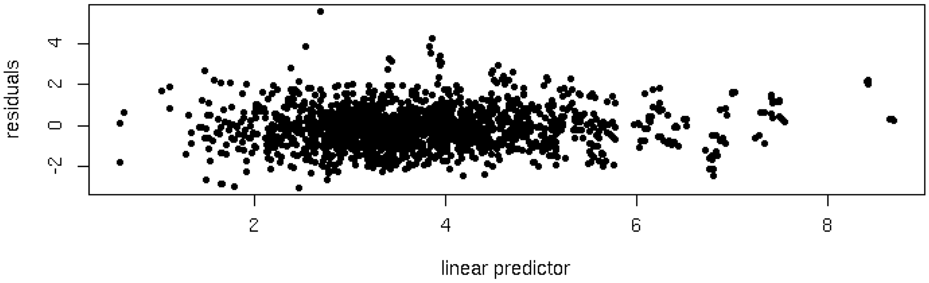


Figure 7: A plot of the residuals against the fitted values.

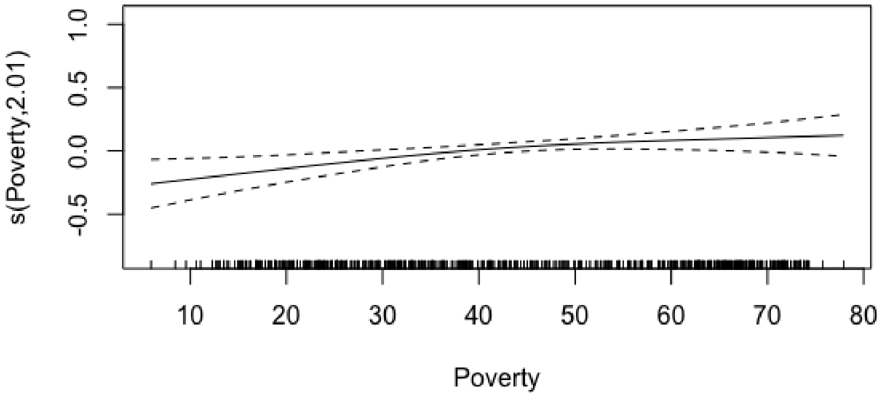


Figure 8: The smooth function of the effect of poverty on TB risk, along with the standard error of the function.

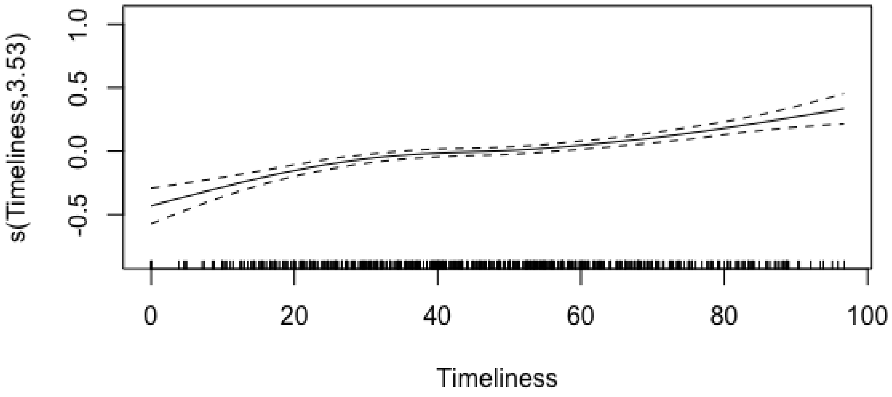


Figure 9: The smooth function of the effect of timeliness on TB risk, along with the standard error of the function.

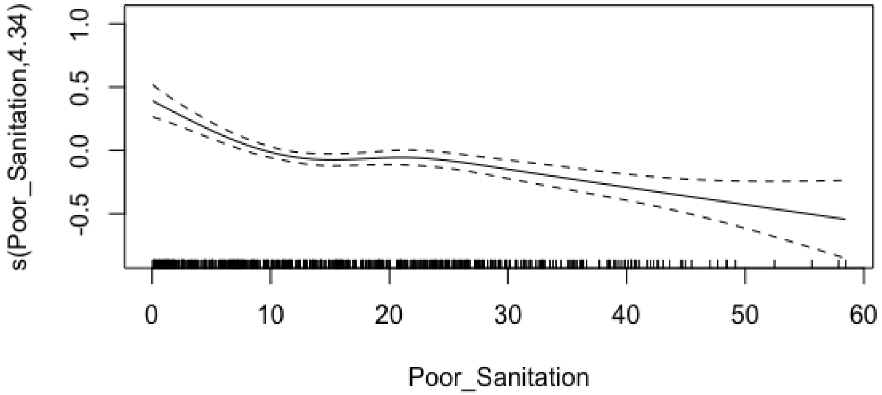


Figure 10: The smooth function of the effect of poor sanitation on TB risk, along with the standard error of the function.

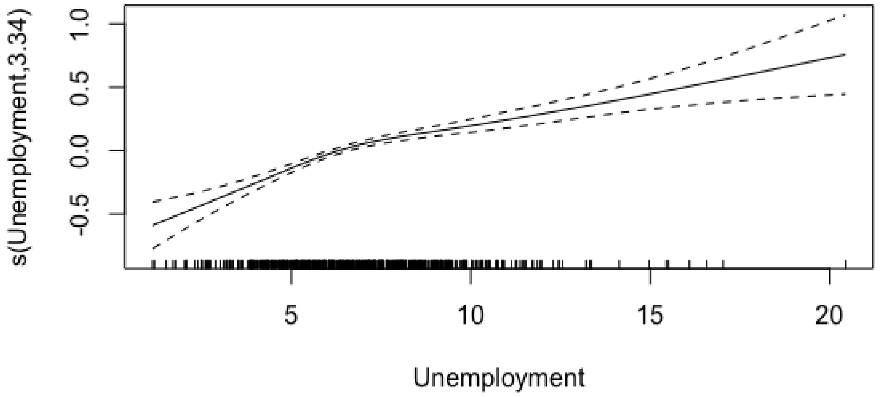


Figure 11: The smooth function of the effect of unemployment on TB risk, along with the standard error of the function.

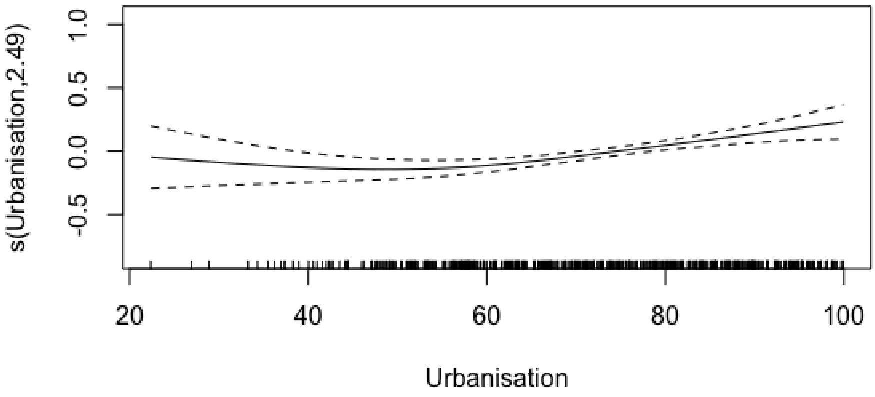


Figure 12: The smooth function of the effect of urbanisation on TB risk, along with the standard error of the function.

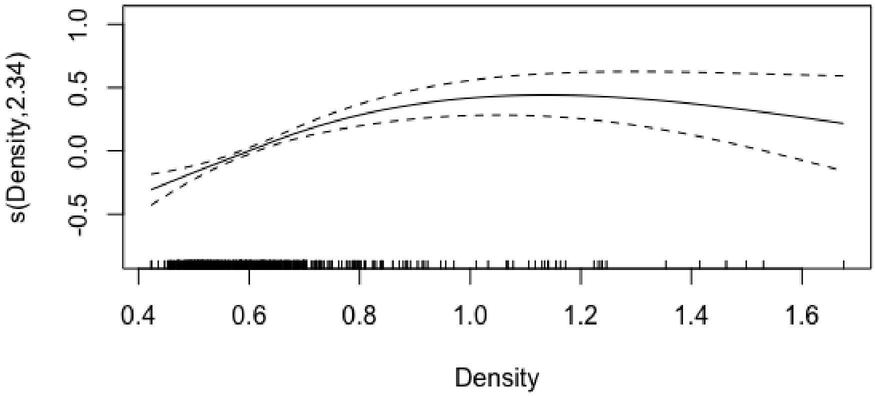


Figure 13: The smooth function of the effect of population den on TB risk, along with the standard error of the function.

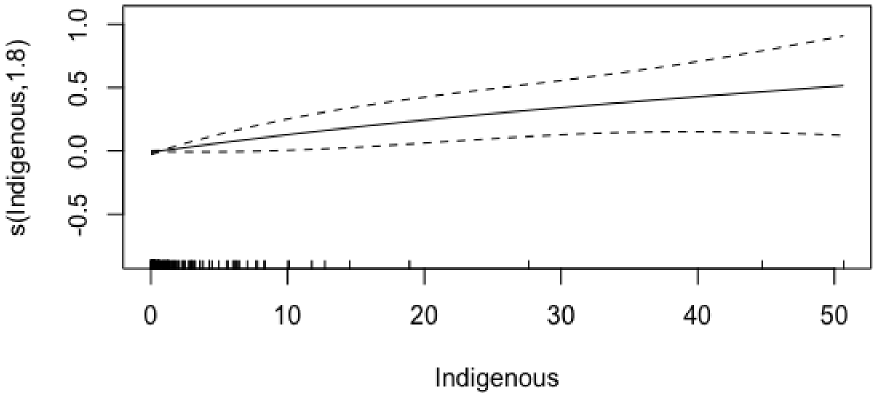


Figure 14: The smooth function of the effect of Indigenous on TB risk, along with the standard error of the function.

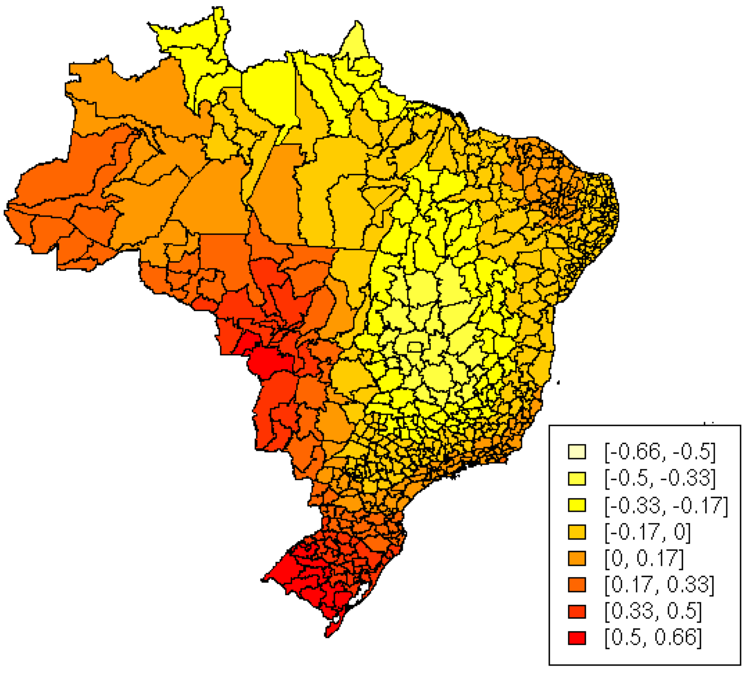


Figure 15: A heatmap showing the effect of location on TB risk in 2012.

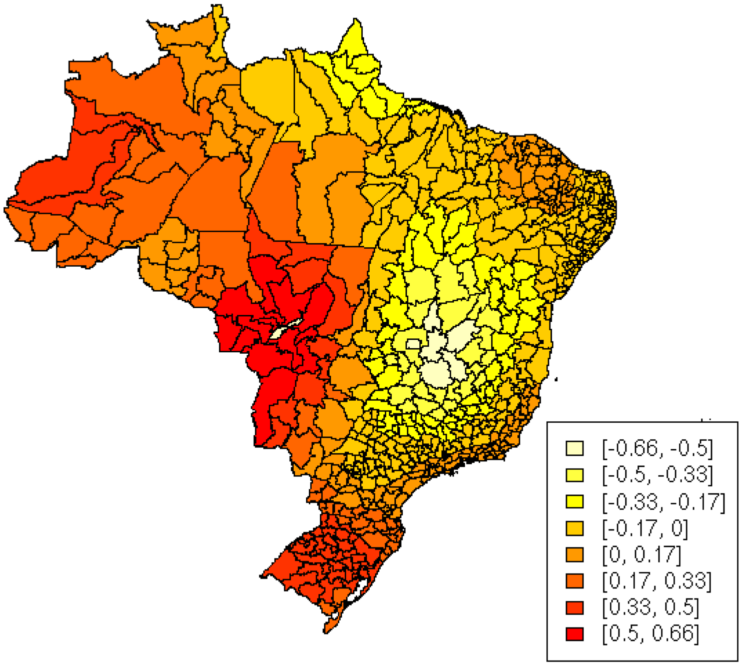


Figure 16: A heatmap showing the effect of location on TB risk in 2013.

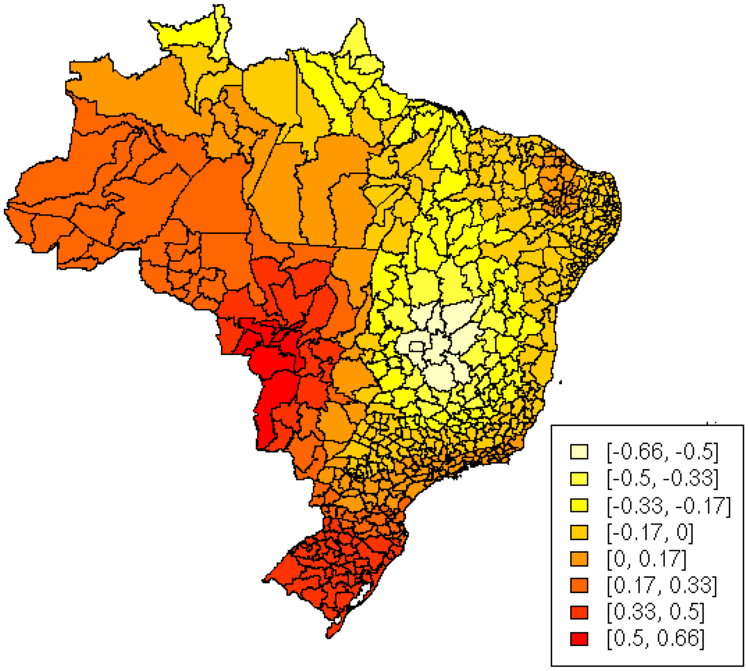


Figure 17: A heatmap showing the effect of location on TB risk in 2014.

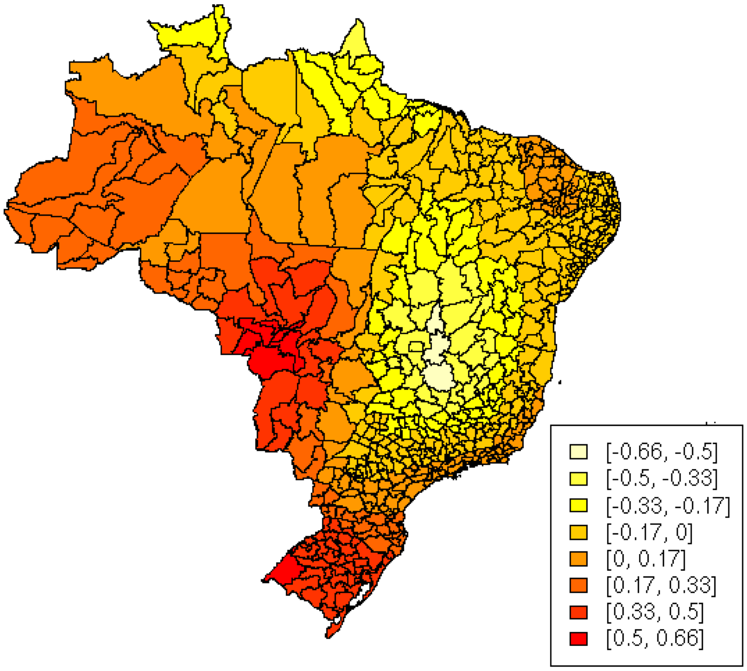


Figure 18: A heatmap showing the overall effect of location on TB risk over the years 2012-2014.

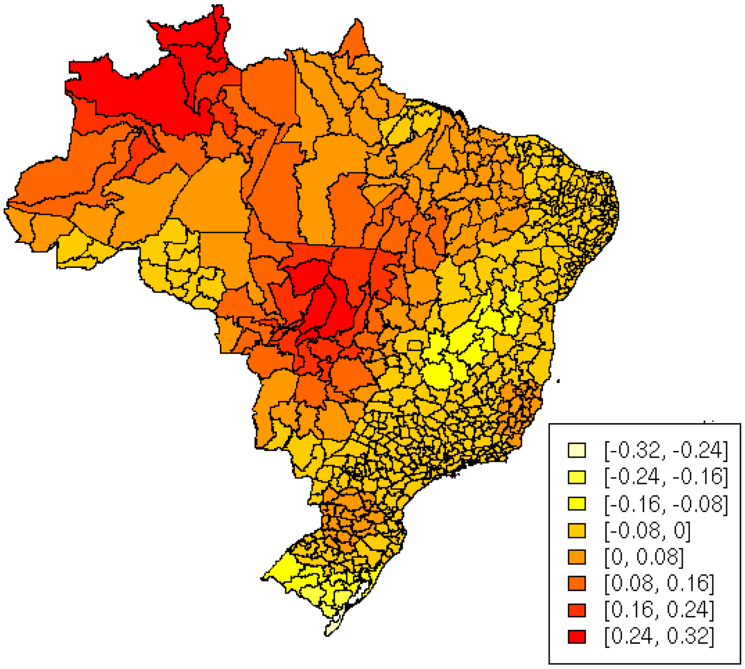


Figure 19: A heatmap showing the change in location effect on TB risk from 2012-2013.

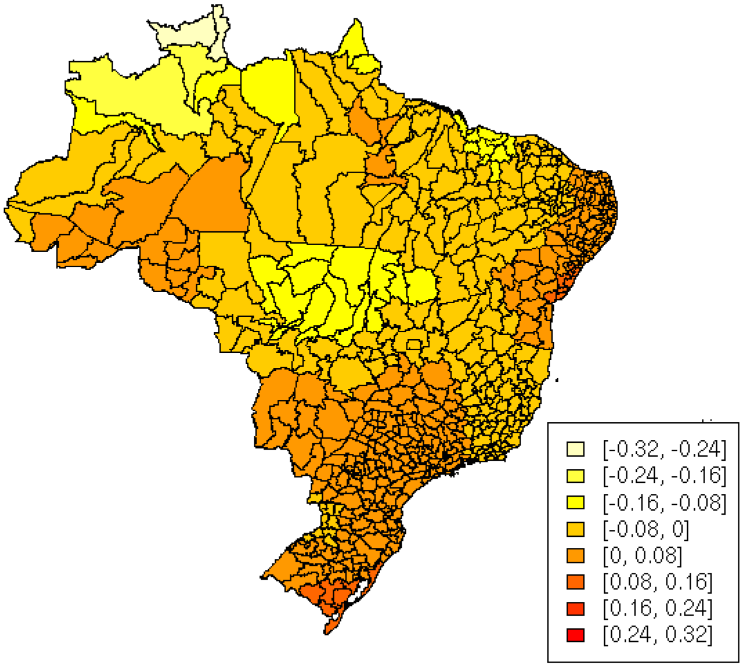


Figure 20: A heatmap showing the change in location effect on TB risk from 2013-2014.

R Code

```r
load('TB.RData') # load the dataset
TBdata$Year <- as.factor((TBdata$Year)) # ategorical variable
TBdata$HunThou <- TBdata$Population / 100000 # TB risk
TBdata <- transform(TBdata, rate = TB/HunThou)

library(mgcv) # load packages
library(lme4)
library(ggplot2)
library(tibble)

# define functions

# plots a map of Brazil with the effect of location
# on TB risk
plt.map.range1 <- function(x,n.levels,main="",cex=1){
  cols <- rev(heat.colors(n.levels)) # reverse heat map
  n <- 557 # number of coordinates
  range <- c(-0.66, 0.66) # stationary range
  splits <- seq(from = range[1], to = range[2],
                length.out = n.levels+1)
  Q <- quantile(x,probs=seq(0,1,len=n.levels+1))
  col <- rep(cols[1],n)
  for(i in 2:n.levels){ # assign coordinates to one of n.levels
    col[x>=splits[i] & x<splits[i+1]] <- cols[i]
  }
  legend.names <- c() # create legend
  for(i in 1:n.levels){
    legend.names[i] <- paste("[",round(splits[i],2), ",␣",
                             round(splits[i+1],2),"]",sep="")
  }
  plot(brasil_micro,col=col,main=main, usePolypath = FALSE)
  legend('bottomright',legend=legend.names,fill=cols,cex=cex)
}

# plots a map of Brazil with the change in effect of location
# on TB risk
plt.map.range2 <- function(x,n.levels=4,main="",cex=1){
  cols <- rev(heat.colors(n.levels)) # reverse heat map
  n <- 557 # number of coordinates
  range <- c(-0.32, 0.32)  # stationary range
  splits <- seq(from = range[1], to = range[2],
                length.out = n.levels+1)
  Q <- quantile(x,probs=seq(0,1,len=n.levels+1))
  col <- rep(cols[1],n)
  for(i in 2:n.levels){ # assign coordinates to one of n.levels
    col[x>=splits[i] & x<splits[i+1]] <- cols[i]
  }
  legend.names <- c() # create legend
  for(i in 1:n.levels){
    legend.names[i] <- paste("[",round(splits[i],2),",␣",
                             round(splits[i+1],2),"]",sep="")
  }
  plot(brasil_micro,col=col,main=main, usePolypath = FALSE)
  legend('bottomright',legend=legend.names,fill=cols,cex=cex)
}

# plots a map of Brazil with the rate of TB in each microregion
plt.map.range3 <- function(x,n.levels,main="",cex=1){
  cols <- rev(heat.colors(n.levels)) # reverse heat map
  n <- 557 # number of coordinates
  range <- c(min(x), max(x)) # stationary range
  splits <- seq(from = range[1], to = range[2],
                length.out = n.levels+1)
  Q <- quantile(x,probs=seq(0,1,len=n.levels+1))
  col <- rep(cols[1],n)
  for(i in 2:n.levels){ # assign coordinates to one of n.levels
    col[x>=splits[i] & x<splits[i+1]] <- cols[i]
  }
  legend.names <- c() # create legend
  for(i in 1:n.levels){
    legend.names[i] <- paste("[",round(splits[i],2), ",␣",
                             round(splits[i+1],2),"]",sep="")
  }
  plot(brasil_micro,col=col,main=main, usePolypath = FALSE)
  legend('bottomright',legend=legend.names,fill=cols,cex=cex)
}

# preliminary data analysis

BetaData <- tibble(x = TBdata$rate[1:557],
                   y = TBdata$rate[558:1114],
                   h = TBdata$rate[1115:1671])
ggplot(BetaData) +
  geom_histogram(aes(x=x), bins =50, alpha=0.3,
                 fill="blue") +
  geom_histogram(aes(x=y), bins =50, alpha=0.3,
                 fill = "red") +
  geom_histogram(aes(x=h), bins =50, alpha=0.3,
                 fill = 'green') +
  xlab('TB␣rate') +
  ylab('Count')

x11(width=10,height=6)
par(mfrow=c(2,2))
plot(TBdata$Indigenous,TBdata$rate,pch=20,lwd=1,
     xlab="Indiginous", ylab="TB",
     main ="Indignious␣and␣TB␣Rate")
plot(TBdata$Illiteracy,TBdata$rate,pch=20,lwd=1,
     xlab="Illiteracy", ylab="TB",
     main ="Illiteracy␣and␣TB␣Rate")
plot(TBdata$Urbanisation,TBdata$rate,pch=20,lwd=1,
     xlab="Urbanisation", ylab="TB",
     main ="Urbanisation␣and␣TB␣Rate")
plot(TBdata$Density,TBdata$rate,pch=20,lwd=1,
     xlab="Density", ylab="TB",
     main ="Density␣and␣TB␣Rate")

x11(width=10,height=6)
par(mfrow=c(2,2))
plot(TBdata$Poverty,TBdata$rate,pch=20,lwd=1,
     xlab="Poverty", ylab="TB",
     main ="Poverty␣and␣TB␣Rate")
plot(TBdata$Poor_Sanitation,TBdata$rate,pch=20,lwd=1,
     xlab="Poor_Sanitation", ylab="TB",
     main ="Poor_Sanitation␣and␣TB␣Rate")
plot(TBdata$Unemployment,TBdata$rate,pch=20,lwd=1,
     xlab="Unemployment", ylab="TB",
     main ="Unemployment␣and␣TB␣Rate")
plot(TBdata$Timeliness,TBdata$rate,pch=20,lwd=1,
     xlab="Timeliness", ylab="TB",
     main ="Timeliness␣and␣TB␣Rate")

all <- list(TBdata$rate[1:557],
            TBdata$rate[558:1114],
            TBdata$rate[1115:1671])
mean_all <- rowMeans(simplify2array(all))
x11()
plt.map.range3(mean_all,n.levels=8,
main="Rate␣of␣TB␣per␣hundred␣thousand␣people")

# Use a negative binomial GAM model to measure TB risk.
# All significant covariates have been used and
# spatio-temporal effects included.
Model1 <- gam(TB ~ offset(I(log(HunThou))) +
                  s(Indigenous,k=5,bs="cs") +
                  s(Urbanisation,k=5,bs="cs") +
                  s(Density,k=5,bs="cs") +
                  s(Poverty,k=5,bs="cs") +
                  s(Poor_Sanitation,k=7,bs="cs") +
                  s(Unemployment,k=7,bs="cs") +
                  s(Timeliness,k=7,bs="cs") +
                  s(lon,lat,k=30, by=Year),
                data=TBdata, family=nb(link="log"))

# model fit analysis
```

```
summary(Model1)
plot(Model1) # plots of smooth functions
sc.deviance <- Model1$deviance/Model1$sig2
1-pchisq(sc.deviance,Model1$df.residual) # fits
x11(width=16,height=6)  # the residual plots look good
par(mfrow=c(2,2))       # and each smooth function has
gam.check(Model1,pch=20) # enough degrees of freedom

# analysis of smooth functions
preds <- predict(Model1, type = 'terms', se.fit=F) # predict
year2012 <- preds[1:557,] # obtains TB effects for each year
year2013 <- preds[558:1114,]
year2014 <- preds[1115:1671,]
cord.preds.2012 <- as.numeric(year2012[,8]) # location effect
cord.preds.2013 <- as.numeric(year2013[,9]) # for each year
cord.preds.2014 <- as.numeric(year2014[,10])
List <- list(cord.preds.2012, cord.preds.2013, cord.preds.2014)
cord.preds.mean <- rowMeans(simplify2array(List)) # overall
cord.change.1 <- cord.preds.2013 - cord.preds.2012 # change
cord.change.2 <- cord.preds.2014 - cord.preds.2013

# the following six plots are heat maps showing the
# effect of location on TB risk
x11()
plt.map.range1(cord.preds.2012,n.levels=8,
               main="Effect␣of␣Location␣on␣Risk␣in␣2012")
x11()
plt.map.range1(cord.preds.2013,n.levels=8,
               main="Effect␣of␣Location␣on␣Risk␣in␣2013")
x11()
plt.map.range1(cord.preds.2014,n.levels=8,
               main="Effect␣of␣Location␣on␣Risk␣in␣2014")
x11()
plt.map.range1(cord.preds.mean,n.levels=8,
               main="Overall␣Effect␣of␣Location␣on␣Risk")
x11()
plt.map.range2(cord.change.1,n.levels=8,
               main="Change␣of␣Location␣Effect␣2012␣-␣2013")
x11()
plt.map.range2(cord.change.2,n.levels=8,
               main="Change␣of␣Location␣Effect␣2013␣-␣2014")
```