Risk Analysis of TB in Brazil

Module: COMM511 **Student ID: 710064461**

Introduction

Mycobacterium tuberculosis causes tuberculosis (TB), is an infectious disease. This study intends to investigate the spatial, temporal and spatial-temporal risk structure of TB cases per unit population in Brazil, looking at the region, illiteracy, sanitation, unemployment, poverty, and population density between 2012 and 2014 to better understand the disease's spread.

The TB data has 1671 observations with no null values, the data has various features i.e. Indigenous (proportion of indigenous population in the region), Illiteracy which measures the literacy level in the people of the region, Urbanization, Density which tells us the average people living in the room, Poverty (measures the poverty level for each region), Poor Sanitation measuring the sanitation level of the regions, Unemployment, Timeliness that tells us the time taken to diagnose TB and report it to the health system, Year, Population, TB(no of TB cases in the region per year), Region, lon(longitude), lat (latitude).

Data Analysis

The summary statistics of the TB data (Fig1) of Brazil tell the maximum Indigenous proportion of people is 50.6, the highest illiteracy rate out of the 557 regions is 41% with an average mean of 14.8%, the maximum density of people living in a room is 1.6 with the mean of 0.6, poverty is highest with 77.8 and the mean value is 44.37, the maximum unemployment rate is 20 with the mean no of people being unemployed as 6.9 and the time is taken to report TB to authorities after its diagnosis is more with 96 as the maximum.

On observing the covariates in the plot (Fig 2), there is a spike in the Indigenous population at the beginning, less no of people with a high illiteracy rate, urbanization is increasing in most of the regions, and a high density of people around 0.6 living in a room. We also observe the normality in the timeliness. To get more insight into these covariates we plot the correlation chart between these covariates to understand their relationship (Fig 3). The TB cases are increasing with the increase of population in a region. There is a high positive correlation between Illiteracy and Poverty as both the factor are linked to each other. Density and urbanization are negatively correlated as urbanization increases the density of people staying in a room is decreasing. Poverty decreases as urbanization rises, with a negative 0.75 correlation.

In plotting some of the covariates on the ggplot we see, poor sanitation has a high influence on poverty and urbanization. As urbanization is increasing poor sanitation is decreasing which shows the development in the individual household (Fig 4). Poverty is directly linked to poor sanitation, poorer households have the worst sanitation facility (Fig 5). The decrease in illiteracy in the population is increasing the timeliness or the resources of the people to report the TB cases to the health authorities (Fig 6). With the increase of urbanization, the timeliness is also increasing depicting that the means and resources of the regions are increasing (Fig 7). When we look at the histogram of TB cases, we can see that there is a higher frequency of small numbers of TB cases Fig 9. We will fit the Poisson model to the TB data because it is an unbounded distribution with a Poisson graph. Between 2012 and 2014, the number of cases of tuberculosis in Brazil grew in 2013, then fell in 2014, with a peak in 2013 (Fig 10). In each of the 7 regions, we see for the region red the TB cases in 2012 were [0,10] which were increased to

[1,12] in 2013 and came down to [0,11] in 2014 (Fig 10). We can also observe the TB cases per 1000 population in brazil from 2012-to 2014 in Fig 11.

Model

The mathematical formula for modeling the Poisson distribution is shown in Fig 12. The $\log(z_i)$ is the offset of the population, and β_0 is the intercept with f(.) as the smooth function. The x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8 are the covariates of Indigenous, Illiteracy, Urbanization, Density, Poverty, Poor Sanitation, Unemployment, and Timeliness respectively. The f(lon, lat, x_{4i}) is the ti() interaction of longitude, latitude, and density.

Experiments

We created a base model with offset and above-mentioned covariates, as the smooth function, with factor term fYear and the interaction term of lon, lat. On fitting a Poisson model with the log link and cubic regression spline 'cr' in model1. The initial k of the covariates in smooth function was small which did not result in the significance of these covariates, increased k value to 60 which gave a significant p-value, and the difference between k' and edf was also greater than 1 in gam.check() Fig13. In model3 we added ti() interaction term in longitude, latitude with density, to observe the effect of density over the years on the regions via 'by' parameter. The gam.check() resulted in significant for the interaction term added for the model in Fig 14. We also tried changing the bs to a cubic spline(model4) but the AIC was the same as that of the model3. To check for the overdispersion in the Poisson model, we tried modeling the data using the quassipoisson model (model5) and negative binomial model (model6) but the covariates results were not significant Fig15. We also tested combining longitude and latitude with additional interaction parameters such as urbanization and poor sanitation. In the model summary, none of them were significant, and the AIC of the model was increased.

The final model is the model3(Fig16) with Poisson and interaction term with lon, lat, and density over time. In this model, the Normal QQ plot shows a straight line with all the data falling in the line. In the residual vs linear predictor plot, the data is randomly scattered around 0. The histogram of the residual shows the normal distribution and the response vs fitted plot shows the data in a straight line of 0, which shows our model fits the data well (Fig 17). We tried to calculate the AIC of all these models Fig 18 which shows that model3 has the lowest AIC of 11926, this can also be confirmed by checking for overdispersion which is 0 Fig 19, so the Poisson model is a good fit. In summary, all the smooth terms of the model are highly significant including the interaction term with the year 2013 and the fixed term for fyear Fig20. The Fig21 shows the result of all the covariates affecting TB in our model with the contour plot and the C.I. for each covariate.

Conclusion

In the study, we observed the 8 covariates are affecting the rate of TB in Brazil. The TB cases increased initially from 2012 and then decreased in the regions in 2014. To reduce TB cases, poverty, population density, illiteracy, unemployment, and inadequate sanitation should all be reduced. With urbanization and literacy, the time it takes to report cases of tuberculosis improves, resulting in a reduction in TB cases.

Critical Review

We could observe the interaction of urbanization and poor sanitation is not much with the spatial-temporal structure in the model. The correlation is much higher among the covariates than the TB. In the year 2012, the density has no bearing on geographic coordinates, which could be investigated further.

```
   Indigenous          Illiteracy       Urbanisation        Density          Poverty       Poor_Sanitation    Unemployment
 Min.   : 0.01034   Min.   : 2.336   Min.   :22.34   Min.   :0.4223   Min.   : 5.923   Min.   : 0.0466   Min.   : 1.128
 1st Qu.: 0.06366   1st Qu.: 6.683   1st Qu.:58.45   1st Qu.:0.5166   1st Qu.:26.229   1st Qu.: 6.3903   1st Qu.: 5.145
 Median : 0.10577   Median :11.516   Median :72.66   Median :0.5840   Median :42.603   Median :13.9129   Median : 6.782
 Mean   : 0.84307   Mean   :14.802   Mean   :71.96   Mean   :0.6212   Mean   :44.371   Mean   :16.4490   Mean   : 6.930
 3rd Qu.: 0.23973   3rd Qu.:22.844   3rd Qu.:86.16   3rd Qu.:0.6585   3rd Qu.:63.907   3rd Qu.:24.9953   3rd Qu.: 8.405
 Max.   :50.64623   Max.   :41.137   Max.   :99.93   Max.   :1.6751   Max.   :77.883   Max.   :58.4328   Max.   :20.438

   Timeliness         Year            TB          Population          Region          lon               lat
 Min.   : 0.00   2012:557   Min.   :  0   Min.   :  23966   11001  :   3   Min.   :-72.86   Min.   :-32.865
 1st Qu.:31.29   2013:557   1st Qu.: 17   1st Qu.: 105054   11002  :   3   1st Qu.:-50.95   1st Qu.:-22.278
 Median :48.36   2014:557   Median : 35   Median : 180423   11003  :   3   Median :-46.31   Median :-15.889
 Mean   :47.67              Mean   :125   Mean   : 357768   11004  :   3   Mean   :-46.42   Mean   :-15.240
 3rd Qu.:62.58              3rd Qu.: 76   3rd Qu.: 315440   11005  :   3   3rd Qu.:-40.64   3rd Qu.: -7.380
 Max.   :96.69              Max.   :9097   Max.   :14597964   11006  :   3   Max.   :-34.95   Max.   :  3.486
                                          (Other):1653
```

Fig1: TB Data Summary

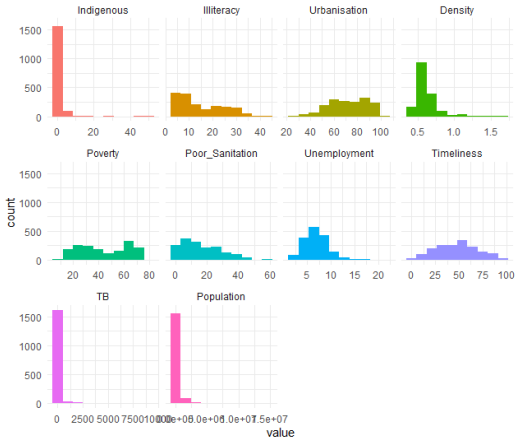

Fig2: TB Covariates

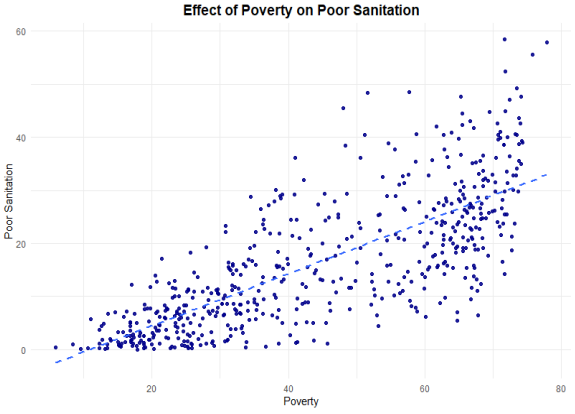

Fig5: Poverty and Poor Sanitation

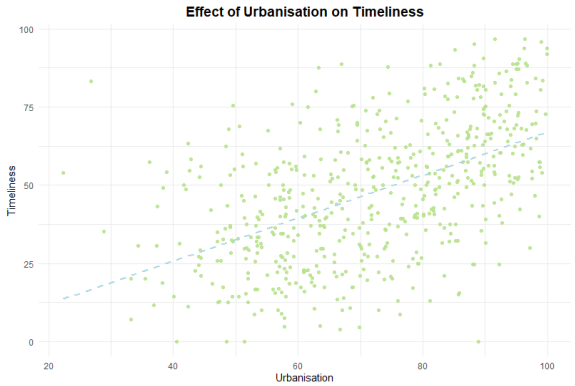

Fig7: Urbanization and Timeline

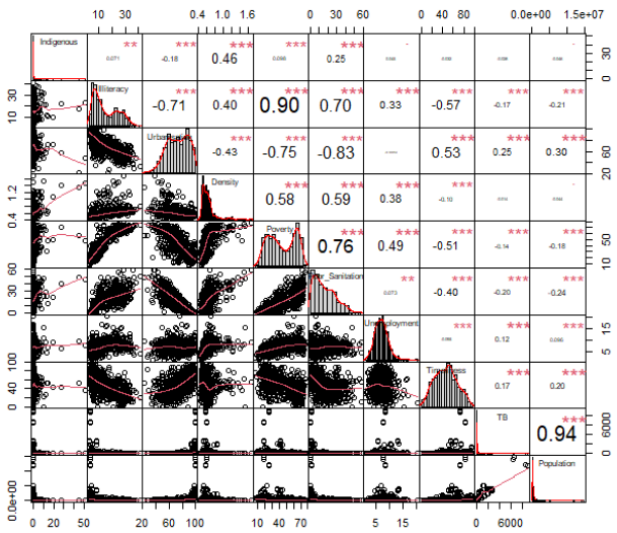

Fig3: Correlation Chart

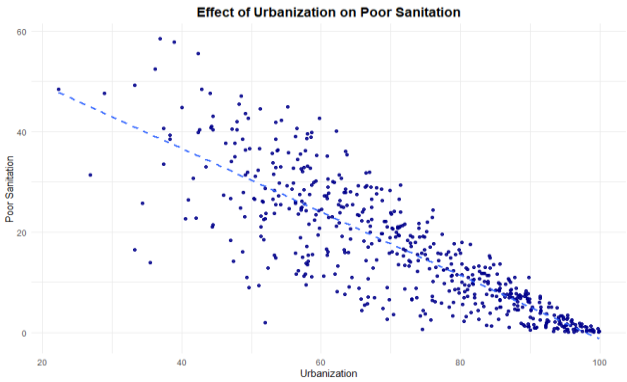

Fig4: Urbanization and Poor Sanitation

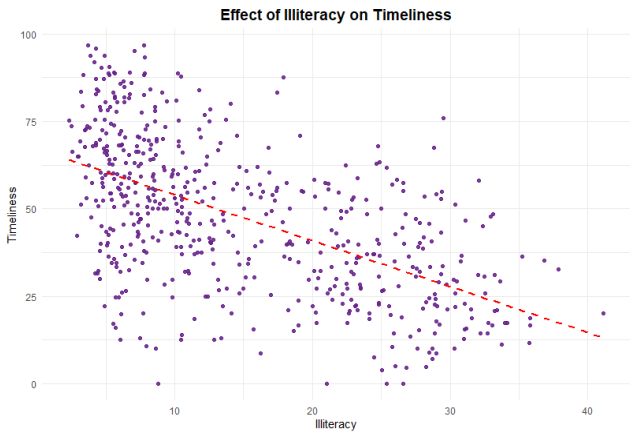

Fig6: Illiteracy and Timeliness

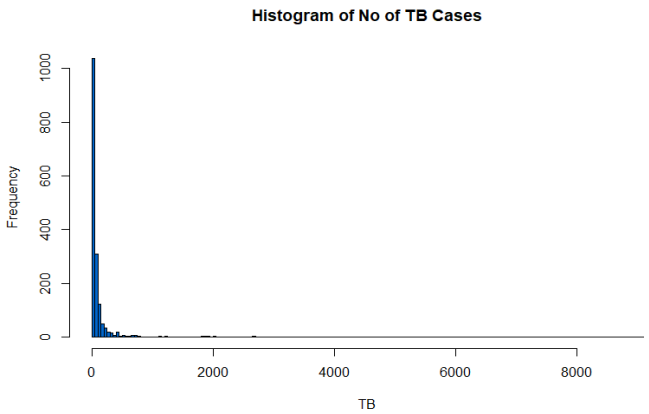

Fig9: TB Cases Yearwise

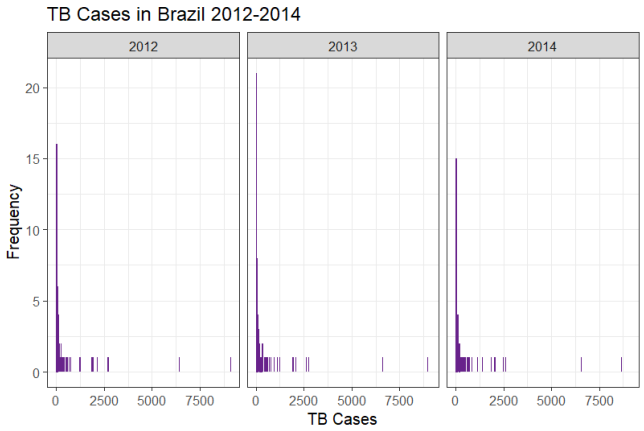

Fig8: Histogram of TB Cases

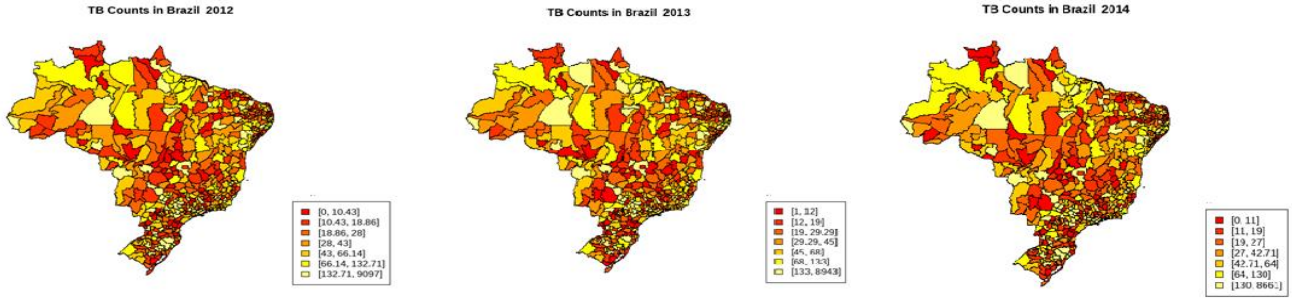

Fig10: TB Cases in Brazil

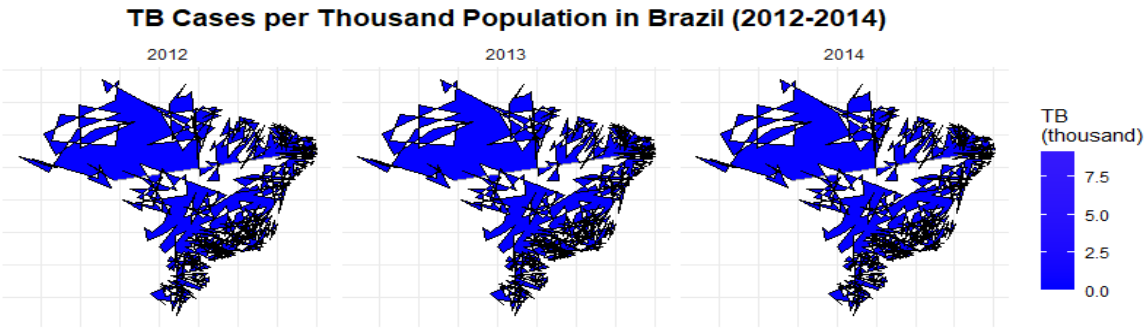

Fig11: TB Cases per thousand in Brazil

$$Y_i \sim Pois(\lambda_i = z_i \rho_i)$$

$$\log(\lambda_i) = \log(z_i) + \log(\rho_i)$$

$$\log(\rho_i) = \beta_0 + f(x_1 i) + f(x_2 i) + f(x_3 i) + f(x_4 i) + f(x_5 i) + f(x_6 i) + f(x_7 i) + f(x_8 i) + f(lon, lat) + f(lon, lat, x_4 i)$$

Fig12: Mathematical Formula

	k'	edf	k-index	p-value
s(Indigenous)	59.00	59.00	1.34	1
s(Illiteracy)	59.00	52.43	1.30	1
s(Urbanisation)	59.00	54.40	1.31	1
s(Density)	69.00	62.30	1.33	1
s(Poverty)	79.00	79.00	1.30	1
s(Poor_Sanitation)	79.00	79.00	1.31	1
s(Unemployment)	79.00	75.41	1.30	1
s(Timeliness)	59.00	57.54	1.24	1
s(lon,lat)	29.00	29.00	1.32	1
ti(lon,lat,Density):fYear2012	64.00	14.29	1.33	1
ti(lon,lat,Density):fYear2013	64.00	7.64	1.33	1
ti(lon,lat,Density):fYear2014	64.00	1.00	1.33	1

Fig14 Model3 Gam Check

	k'	edf	k-index	p-value
s(Indigenous)	59.0	59.0	1.35	1
s(Illiteracy)	59.0	41.1	1.31	1
s(Urbanisation)	59.0	57.8	1.32	1
s(Density)	69.0	69.0	1.34	1
s(Poverty)	79.0	77.9	1.31	1
s(Poor_Sanitation)	79.0	78.8	1.32	1
s(Unemployment)	79.0	79.0	1.30	1
s(Timeliness)	59.0	59.0	1.25	1
s(lon,lat)	29.0	26.9	1.32	1

Fig13: Model1 Gam Check

	k'	edf	k-index	p-value
s(Indigenous)	29.00	2.38	0.53	<2e-16 ***
s(Illiteracy)	29.00	3.24	0.53	<2e-16 ***
s(Urbanisation)	29.00	5.34	0.53	<2e-16 ***
s(Density)	29.00	3.95	0.53	<2e-16 ***
s(Poverty)	29.00	1.93	0.54	<2e-16 ***
s(Poor_Sanitation)	29.00	6.52	0.52	<2e-16 ***
s(Unemployment)	29.00	3.78	0.54	<2e-16 ***
s(Timeliness)	39.00	3.58	0.61	<2e-16 ***
s(lon,lat)	29.00	26.06	0.50	<2e-16 ***
ti(lon,lat,Density):fYear2012	64.00	1.00	0.53	<2e-16 ***
ti(lon,lat,Density):fYear2013	64.00	1.00	0.53	<2e-16 ***
ti(lon,lat,Density):fYear2014	64.00	1.01	0.53	<2e-16 ***

Fig15 Model6 Gam Check

```
model3 <- gam( TB ~ offset(I(log(Population))) + s(Indigenous, bs="cr", k= 60) + s(Illiteracy, bs="cr", k= 60) + s(Urbanisation, bs="cr", k= 60)+
                 + s(Density, bs="cr", k= 70) + s(Poverty, bs="cr", k= 80) + s(Poor_Sanitation,bs="cr", k= 80) + s(Unemployment,bs="cs", k= 80)+
                 s(Timeliness,bs="cs",k=60) + s(lon,lat) + fYear+ ti(lon, lat,Density, by=fYear,bs=c('tp','cr'), k=c(50,30)),
                 data = TBdata,
                 family= poisson(link = 'log'))
```

Fig16 Model3

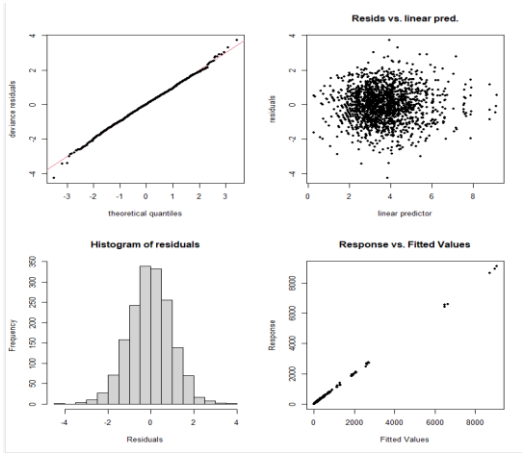

Fig17 Model3 Gam Check Plot

```
> AIC(model1, model3, model4, model5, model6)
             df      AIC
model1 551.41532 11974.99
model3 574.00924 11926.85
model4 574.01025 11926.85
model5 181.74625       NA
model6  73.65426 14034.97
```

Fig18 AIC Score

```
> overdispersion = model3$deviance / model3$df.residuals
> overdispersion
numeric(0)
```

Fig19 Overdispersion

```
Parametric coefficients:
             Estimate Std. Error  z value Pr(>|z|)
(Intercept) -8.597937   0.076353 -112.608  < 2e-16 ***
fYear2013   -0.018446   0.007871   -2.344   0.0191 *
fYear2014   -0.048093   0.007549   -6.370 1.88e-10 ***
---
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)                59.000 59.000 365.357  < 2e-16 ***
s(Illiteracy)                52.428 53.463 285.980  < 2e-16 ***
s(Urbanisation)              54.399 55.503 492.386  < 2e-16 ***
s(Density)                   62.300 63.559 404.196  < 2e-16 ***
s(Poverty)                   79.000 79.000 461.870  < 2e-16 ***
s(Poor_Sanitation)           79.000 79.000 577.772  < 2e-16 ***
s(Unemployment)              75.412 79.000 606.379  < 2e-16 ***
s(Timeliness)                57.539 59.000 530.052  < 2e-16 ***
s(lon,lat)                   29.000 29.000 132.078  < 2e-16 ***
ti(lon,lat,Density):fYear2012 14.286 18.411  27.757 0.070787 .
ti(lon,lat,Density):fYear2013  7.641  8.753  32.167 0.000229 ***
ti(lon,lat,Density):fYear2014  1.003  1.005   1.936 0.164349
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Fig20 Model3 Summary

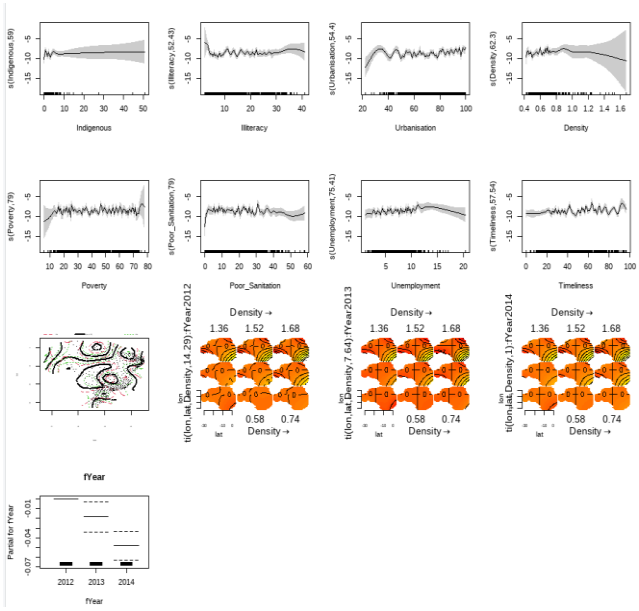

Fig21 Model3 Plot

Appendix

```
# Library
library(ggplot2)
library(PerformanceAnalytics)
library(mgcv)
library(psych)
library(funModeling)
library(Hmisc)

theme_set(theme_minimal())
theme_update(plot.title = element_text(hjust = 0.5,size = 14, face = "bold"))

nrow(TBdata)
head(TBdata,5)
names(TBdata)
str(TBdata)
is.null(TBdata)
TBdata$fYear <- as.factor(TBdata$Year)
TBdata$fRegion <- as.factor(TBdata$Region)

# EDA
summary(TBdata)
TB_sub <- subset(TBdata, select = -c(Year, Region, lon, lat, fYear, fRegion))
eda <- function(my_data)
{
  freq(my_data)
  print(profiling_num(my_data))
  plot_num(my_data)
  describe(my_data)
}

eda(TB_sub)
chart.Correlation(TB_sub, histogram = TRUE, method = "pearson")

# Multivariate Analysis
ggplot(TBdata, aes(Urbanisation,Poor_Sanitation))+
  geom_point(alpha=0.5,color="darkblue")+
  labs(x = "Urbanization", y = "Poor Sanitation",
       title="Effect of Urbanization on Poor Sanitation")+
  geom_smooth(method=lm,se=FALSE,linetype="dashed")

ggplot(TBdata, aes(Poverty,Poor_Sanitation))+
  geom_point(alpha=0.5,color="darkblue")+
  labs(x = "Poverty", y = "Poor Sanitation",
       title="Effect of Poverty on Poor Sanitation")+
  geom_smooth(method=lm,se=FALSE,linetype="dashed")

ggplot(TBdata, aes(Illiteracy,Timeliness))+
  geom_point(alpha=0.5, color="darkorchid4")+
  labs(x = "Illiteracy", y = "Timeliness", title="Effect of Illiteracy on Timeliness")+
  geom_smooth(method=lm,se=FALSE,linetype="dashed",color="red")

ggplot(TBdata, aes(Urbanisation,Timeliness))+
```

```
  geom_point(alpha=0.5, color="#b9e38d")+
  labs(x = "Urbanisation", y = "Timeliness", title="Effect of Urbanisation on Timeliness")+
  geom_smooth(method=lm,se=FALSE,linetype="dashed",color="lightblue")

hist(TB_sub$TB,
     main="Histogram of No of TB Cases", xlab='TB', col="#0066CC", breaks=300)

ggplot(TBdata,aes(TB)) +
  geom_bar(color = "darkorchid4") +
  facet_wrap( ~ Year ) +
  labs(title = "TB Cases in Brazil 2012-2014",
       y = "Frequency",
       x = "TB Cases") + theme_bw(base_size = 15)


ggplot(data = TBdata, # the input data
       aes(x = lon, y = lat, fill = TB/1000, group = Year)) +
  geom_polygon(aes(group = Year), colour = "grey60") +
  geom_path(colour="black", lwd=0.5) +
  coord_equal() +
  facet_wrap(~ Year) +
  scale_fill_gradient2(low = "blue", mid = "grey", high = "red",
                       midpoint = 150, name = "TB\n(thousand)") +
  labs(title = "TB Cases per Thousand Population in Brazil (2012-2014)")+
  theme(axis.text = element_blank(),
        axis.title = element_blank(),
        axis.ticks = element_blank())

x <- c(2012,2013,2014)
for (val in x){
  plot.map(TBdata$TB[TBdata$Year==val],n.levels=7,main=paste("TB Counts in Brazil ", val))

}

## Models
# Base model
model1 <- gam( TB ~ offset(I(log(Population))) + s(Indigenous, bs="cr", k= 60) +
                 s(Illiteracy, bs="cr", k= 60) + s(Urbanisation, bs="cr", k= 60)+
                 +s(Density, bs="cr", k= 70) +s(Poverty, bs="cr", k= 80) +
                 s(Poor_Sanitation,bs="cr", k= 80) + s(Unemployment,bs="cs", k= 80)+
                 s(Timeliness,bs="cs",k=60) + s(lon,lat) + fYear,
               data = TBdata,
               family= poisson(link = 'log'))

par(mfrow=c(2,2))
gam.check(model1, pch=20)

# Interaction term of Urbanization and Poor Sanitation with long, lat
model2 <- gam( TB ~ offset(I(log(Population))) + s(Indigenous, bs="cr", k= 60) +
                 s(Illiteracy, bs="cr", k= 60) + s(Urbanisation, bs="cr", k= 60)+
                 + s(Density, bs="cr", k= 70) + s(Poverty, bs="cr", k= 80) +
                 s(Poor_Sanitation,bs="cr", k= 80) + s(Unemployment,bs="cs", k= 80)+
                 s(Timeliness,bs="cs",k=60) + s(lon,lat) + fYear+
```

```
                ti(lon, lat, Urbanisation ,bs=c('tp','cr'), k=c(50,30))+
                ti(lon, lat, Poor_Sanitation ,bs=c('tp','cr'), k=c(50,30)),
              data = TBdata,
              family= poisson(link = 'log'))

par(mfrow=c(2,2))
gam.check(model2, pch=20)
summary(model2)

# Interaction term of Density with long, lat
model3 <- gam( TB ~ offset(I(log(Population))) + s(Indigenous, bs="cr", k= 60) +
                s(Illiteracy, bs="cr", k= 60) + s(Urbanisation, bs="cr", k= 60)+
                + s(Density, bs="cr", k= 70) + s(Poverty, bs="cr", k= 80) +
                s(Poor_Sanitation,bs="cr", k= 80) + s(Unemployment,bs="cs", k= 80)+
                s(Timeliness,bs="cs",k=60) + s(lon,lat) + fYear+
                ti(lon, lat,Density, by=fYear,bs=c('tp','cr'), k=c(50,30)),
                data = TBdata,
                family= poisson(link = 'log'))

par(mfrow=c(2,2))
gam.check(model3, pch=20)
summary(model3)


model4 <- gam( TB ~ offset(I(log(Population))) + s(Indigenous, bs="cs", k= 60) +
                s(Illiteracy, bs="cs", k= 60) + s(Urbanisation, bs="cs", k= 60)+
                +s(Density, bs="cs", k= 70) +s(Poverty, bs="cs", k= 80) +
                s(Poor_Sanitation,bs="cs", k= 80) + s(Unemployment,bs="cs", k= 80)+
                s(Timeliness,bs="cs",k=60) + s(lon,lat) + fYear+
                ti(lon, lat,Density, by=fYear,bs=c('tp','cr'), k=c(50,30)),
              data = TBdata,
              family= poisson(link = 'log'))

par(mfrow=c(2,2))
gam.check(model4, pch=20)

# Quasipoisson
model5 <- gam( TB ~ offset(I(log(Population))) + s(Indigenous, bs="cr", k= 30) +
                s(Illiteracy, bs="cr", k= 30) + s(Urbanisation, bs="cr", k= 30)+
                +s(Density, bs="cr", k= 30) +s(Poverty, bs="cr", k= 30) +
                s(Poor_Sanitation,bs="cr", k= 30) + s(Unemployment,bs="cs", k= 30)+
                s(Timeliness,bs="cs",k=40) + s(lon,lat) + fYear+
                ti(lon, lat,Density, by=fYear,bs=c('tp','cr'), k=c(50,30)),
              data = TBdata,
              family= quasipoisson("log"),
              method = "REML")

par(mfrow=c(2,2))
gam.check(model5, pch=20)

## Negative Binomial
model6 <- gam( TB ~ offset(I(log(Population))) + s(Indigenous, bs="cr", k= 30) +
                s(Illiteracy, bs="cr", k= 30) + s(Urbanisation, bs="cr", k= 30)+
```

```
                  +s(Density, bs="cr", k= 30) +s(Poverty, bs="cr", k= 30) +
                  s(Poor_Sanitation,bs="cr", k= 30) + s(Unemployment,bs="cs", k= 30)+
                  s(Timeliness,bs="cs",k=40) + s(lon,lat) + fYear+
                  ti(lon, lat,Density, by=fYear,bs=c('tp','cr'), k=c(50,30)),
                data = TBdata,
                family= nb(link="log"),
                method = "REML")

par(mfrow=c(2,2))
gam.check(model6, pch=20)


AIC(model1, model3, model4, model5, model6)
AIC(model2, model3)

# Overdispersion in Poisson model
overdispersion = model3$deviance / model3$df.residuals
overdispersion

plot(model3, shade=TRUE, seWithMean=TRUE, pages=1, all.terms=TRUE,
     shift = coef(model3)[1], rug=TRUE)
```