Quantifying Spatio-temporal Risk of TB in Brazil

Daniel Burke, Arthur Finch, Matthew Found, Charlie Wilkinson

1 Introduction & Problem Overview

In this report, we use Generalized Additive Models (GAMs) to quantify spatial, temporal, and spatio-temporal risk from tuberculosis (TB) in Brazil. The *TBdata* dataframe contains measurements of TB levels and several other variables as outlined in *Appendix A*. In this dataset we have annual response and covariate measurements for 2012 to 2014 where Brazil is divided into 557 administrative microregions and all variables are given per microregion. We define TB risk to be a continuous variable as the rate of TB cases per $100,000$ individuals. It is important to note that both population and TB cases are count data bounded at zero, where TB cases intuitively cannot exceed population. Poisson and Negative Binomial models may be useful here and are often used when dealing with count or bounded nature datasets. In order to comprehensively quantify spatio-temporal TB risk, we decompose our consideration into sub-problems:

- Do the socio-economic covariates have a significant effect on TB risk?
- Spatial, Temporal, Spatio-temporal structure of systematic risk not explained by covariates.
- Which are the regions with high TB risk and how should medical resources be allocated?

We use GAM in order to explore and answer the sub-problems listed above. A GAM is an extension of the Generalized Linear Model (GLM) that relaxes the assumption that the mean of the response variable must be a linear combination of the predictor variables[6]. Instead of assuming a strictly linear relationship, a GAM models the linear predictor (η) as the sum of some smooth nonlinear functions of the predictor variables $f(\cdot)$:

$$g(\mu_i) = \eta_i = \mathbf{z}'\beta + f_1(x_{1,i}) + ... + f_p(x_{p,i})$$

The smooth functions enable the GAM to capture nonlinear relationships between the predictors and the response variable without explicitly specifying the functional form of their relationship. This inherent adaptability makes GAMs a very useful tool in our efforts to explore the dataset and answer the questions posed. We can experiment with different functions of spatial, temporal, and socioeconomic covariates in relation to our TB risk response variable.

2 Model Fitting

Based on the nature of the dataset we elect to implement a Poisson distribution. In the first instance we analysed plots comparing TB case counts with the other variables and variable interactions highlighted by color scale to guide our model construction. The discovery of non-linear patterns such as an exponential increase in TB cases at the highest levels of Urbanisation prompted the introduction of a logarithmic mean function through the (link='log') argument which allows our model to better capture these behaviours. Moreover, we observed a linear relationship between Population and TB cases, prompting the addition of a log-transformed offset term into the model to reflect the proportional increase of TB cases with population whilst aligning with the mean function. With the foundations of our model built, we began the iterative process of adding variables to the model as smooth terms. We cyclically repeated steps 1-4 of the following process:

1. Fit the model and view summary. Primarily consider significance of added terms to assess importance and deviance explained percentage as a measure of goodness of fit to justify added complexity.

2. Plot covariate smooths produced by model. These indicate the effect a term has on the response.

3. Plot and assess the residuals of the model e.g. using QQ plot, histogram of residuals, residuals vs linear predictions, response vs fitted values to comment on model performance.

4. Assess 'gam.check' summary for potential knot adjustments.

In our initial plots we observed that Urbanisation and Timeliness appear to have a significant impact on TB cases, and we subsequently added these into our model as smooth terms. Each of the smooth terms uses cubic splines with 5 knots for model flexibility. In completing the above process we found that both Urbanisation and Timeliness were significant but that the model returned poor QQ plots, and that the residuals vs linear predictor plots exhibited funneling as shown in the first row of Fig. 1. The covariate smooth plots and GAM summary showed that Urbanisation was positively impacting TB cases at high Urbanisation values as previously found during our initial plots. K-index values were low ~ 0.37, and increasing the number of knots didn't cause meaningful change so we decided to increase model complexity.

Urbanisation, which is the shift towards greater proportions of urban living [1], correlates with TB cases, which is spread through the air [5]. Following the logical connection of TB transmission to densely populated and unhygienic areas commonly associated to urban environments that we have already shown has an impact, we added Density and Poor Sanitation to further improve the model using the same smooth term structure as before. With these additions, our model's explained deviance increased from 36.5% to 54%, indicating a better fit. Analysis of the covariate smooths showed that Density and Poor Sanitation positively influenced TB cases, although Poor Sanitation levels paradoxically had a negative impact, possibly due to limited data points at higher levels or the difficulty in preventing airborne transmission with personal hygiene. Despite improvements in the residual plots, row 2 in Fig. 1 demonstrates funneling in the linear predictor plot and deviations in the tails of the QQ plot. Additionally, the k-index was well below 1, and despite efforts to improve this with extra knots, we found it took in excess of 100 knots to see k-index values approach 1. Therefore, we increased knots to 20 for improved model performance and opted to further increase model complexity.

In the next iteration of our model we added Poverty and Unemployment smooth terms, significantly increasing the explained deviance from 54% to 77.6%. We observed both of these terms to be significant. The covariate smooths show Unemployment to have a strong positive effect on TB cases, and Poverty to have a varied but overall positive impact. We observed marginal improvements in model fit based on the residual plots. The k-index for each smooth term was still low $\sim 0.43 - 0.47$ and as such we doubled the number of knots from 20 to 40 for each smooth term and saw significant improvements in the model fit based on the residuals plots shown in row 3 of Fig. 1. Despite this the k-index values only saw a $\sim 22\%$ increase, suggesting motive for adding further model complexity.

By incorporating latitude and longitude for a spatial component through a thin plate spline, we enhanced our model as shown in row 4 of Fig. 1. With the residuals behaving as expected, we optimised the number of knots for each smooth term to approach a k-index of 1. Subsequently, the summary indicated all terms were significant and our model's explained deviance increased from 77.6% to 99.4%, indicating an excellent fit. It is worth noting that adding a temporal dimension with the Year variable, both individually and by Region, showed no effect on TB cases and thus TB risk due to the flat lines at 0 in the smooth plots we observed. Therefore, it was omitted from our final model. Mathematically, our final model takes the following form:

$$g(\mu_i) = \eta_i = \mathbf{z}'\beta + f_1(x_{1,i}) + f_2(x_{2,i}) + f_3(x_{3,i}) + f_4(x_{4,i}) + f_{5,6}(x_{5,i}, x_{6,i}) + f_7(x_{7,i}) + f_8(x_{8,i})$$

where x_1 = Timeliness, x_2 = Urbanisation, x_3 = Poverty, x_4 = Density, x_5 = Latitude, x_6 = Longitude, x_7 = Unemployment and x_8 = Poor sanitation.

3 Results & Discussion

3.1 Covariates

As shown in Section 2, all covariates used in our final model have significant effect on the TB case rate. We evaluate the individual effect of each of the covariates on TB risk by plotting their smooth functions. Within Fig. 2 we observe that Urbanisation and Density share a positive trend as their values increase and generally have a positive effect on the rate of TB. This agrees with our initial thoughts, whereby more densely populated areas will be vulnerable to the airborne spread of TB. Likewise, Poverty shares a similar positive trend, where we observe increasing TB risk with a greater poverty index. Investigations regarding the relationship between TB and socioeconomic characteristics within Brazil found the disease to be most prevalent in the countries poorest areas, with results confirming that TB is determined by the population's living conditions [3].

As unemployment and poverty are closely related socioeconomic factors, we would expect similar plots for the two covariates. Whilst both have a distinguished initial spike, which implies a strong negative risk of TB when these variables are equal to zero, unemployment remains around zero until around the 10% rate. Whilst there is a positive effect for the larger values, there is a large drop in data potentially harming the credibility of any inference.

Timeliness exhibits a slight positive trend across the values, where risk is impartial for quicker reporting times, yet positive for the slower values. This variable is used as a proxy for healthcare resources, where surrounding evidence suggests that structural and organisational obstacles faced in Brazil's healthcare (primarily in the Northeast) hinder the performance of health services when treating TB [4]. The final covariate we investigate is poor sanitation, which demonstrates a strong negative trend. This opposes our initial intuition, as we would expect poor sanitation to increase the risk of TB. Without additional context regarding the extent to which poor sanitation refers to personal hygiene, we are unable to drawn meaningful conclusions to the impact of the variable on the risk of TB.

3.2 Spatial Analysis

Fig. 4 a) shows the smooth function as a heatmap over Longitude and Latitude over Brazil, positive values indicate an increased TB risk in those areas and vice versa for negative values. One region of clearly increased risk is on the Western border that Brazil shares with Bolivia and Paraguay. These regions are quite rural and contain mostly collections of small towns, which may not have access to proper sanitation or medicine, or even education about how disease spreads. This is in agreement with what we observe in Fig. 4 b) which is our model's predicted TB risk. We can also see an increased TB risk on the South-Eastern coastline, the most populous area of Brazil, with the two largest cities, São Paulo and Rio de Janiero. These regions may be more at risk due to a higher density and greater poverty inside the favelas of the major cities. This increase of the smooth function is once again reflected in our model's predictions, with the maximum prediction appearing in São Paulo. It is worth noting that this is a smooth function of latitude and longitude only, and the effects of other covariates, though not directly influencing the spatial smooth function, may be present in the output due to the socioeconomic status in that area, such as the poverty in the favelas discussed above. That being said, it is clear that some regions area at greater risk of TB than others, and this map could be used by health authorities to guide the allocation of resources to deal with the disease.

3.3 Temporal Analysis

The only covariate in our model which changes in time is Population, and all other other socioeconomic variables are invariant over the three years, meaning that an extensive temporal analysis cannot be carried out in this case. When observing the TB risk distribution by year shown in Fig. 3, we see that the changes are minute, further supporting that the only covariate exhibiting change across the years is population. Further challenges arise in the coarse granularity of time in the dataset when trying to understand the structure of systematic risk. It would be common in Epidemiology to have more frequent response and covariate measurement intervals. Due to the annual granularity of time, we are unable to conduct time-series analysis by quarter or season and are thus unable to quantify short term trends or understand within-year fluctuations. Moreover we only actually have three unique time points: 2012, 2013, and 2014. Looking at TB Risk for all regions year-by-year we observe fairly constant mean, median, and standard deviation of near 23, 20, and 15 respectively. There is a marginal increase in TB Risk overall in 2013 which is also the only year where all regions in the dataset have non-zero TB risk, but an ANOVA test shows that this risk increase is not statistically significant. These findings are reiterated when looking at density plots, where all three years show similar single-peaked density curves with positive skew. On an individual basis, looking region-by-region we observe that in some places TB Risk increases year-on-year, while it decreases or remains constant in others. Whilst we are able to observe these spatio-temporal trends, we are unable to observe significant purely temporal trends in TB risk over time.

3.4 Spatio-Temporal Analysis

One other source of systematic risk in space and time not captured in the dataset is that of large organised events. Major events in Brazil during this period such as the World Cup and Confederations Cup could have had impacts on TB Risk as super-spreading events in the host regions, potentially responsible for a substantial majority of secondary TB infections [2]. Our inability to investigate these potentially significant events as well as seasonal fluctuations limits the completeness of our spatio-temporal analysis.

4 Conclusion & Critical Review of Analysis

In this report we aimed to quantify the risk TB across Brazil over a three year time period and evaluate the different factors that significantly impact the disease risk. To do so, we utilised a GAM that included socioeconomic covariates such as urbanization, density, poverty, unemployment, timeliness of reporting, and poor sanitation, as well as spatial and temporal components. Investigation into the socioeconomc factors suggested that Density, Urbanisation and Poverty significantly affect the rate of TB risk, exhibiting a positive trend with high values. This aligned with surrounding research investigating socioeconomic status and TB. Interestingly, poor sanitation demonstrated a negative trend in affecting TB risk, which with different data, could be studied further. Another improvement to our model could be including more time dependant covariates such as the Year, to further our understanding how the disease changes over time. Our spatial analysis uncovered regions with higher TB risk, specifically along the Western border and the densely populated areas along the Southeastern coastline. This information could guide the allocation of resources and targeted interventions by health authorities. Our temporal analysis was limited by the short time-frame and annual granularity of the data. Whilst we observed some marginal increases in TB risk for 2013, the difference was not statistically significant. Future investigation may benefit from having access to data with more granular time resolution, such as monthly measures over a larger timescale, or to include time variation of socioeconomic variables. Such data would allow quantification of spatial and temporal patterns, giving better insight that would further improve prevention strategies for the health authorities in Brazil.

Figures

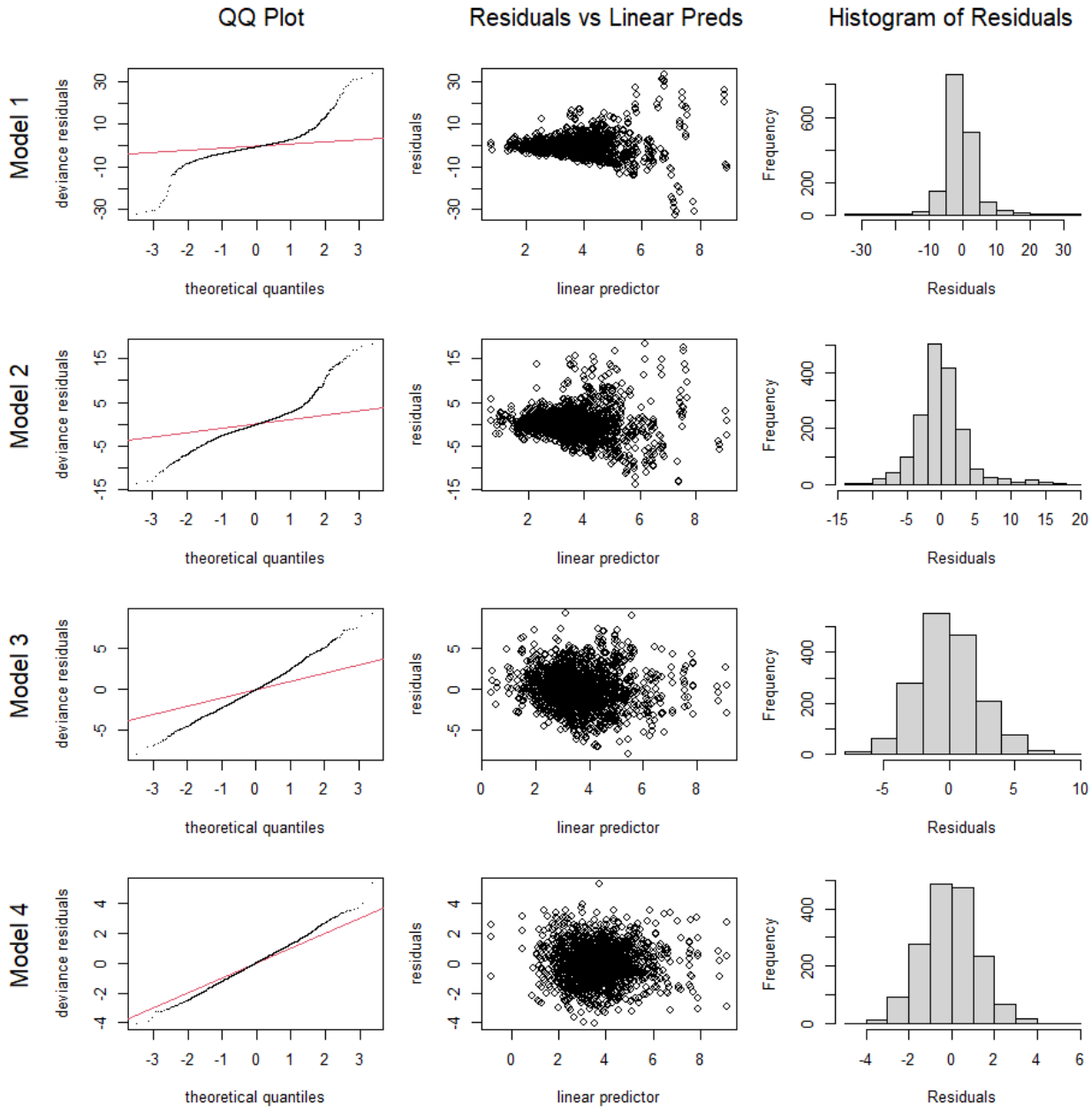

Figure 1: Plot of residuals during the model building phase. Each column represents the labeled residual plot and each row corresponds to a respective model, showing the iterative process of using the residuals to justify the inclusion of additional covariates. The figure shows how the residuals improve towards their desired output during the model building.

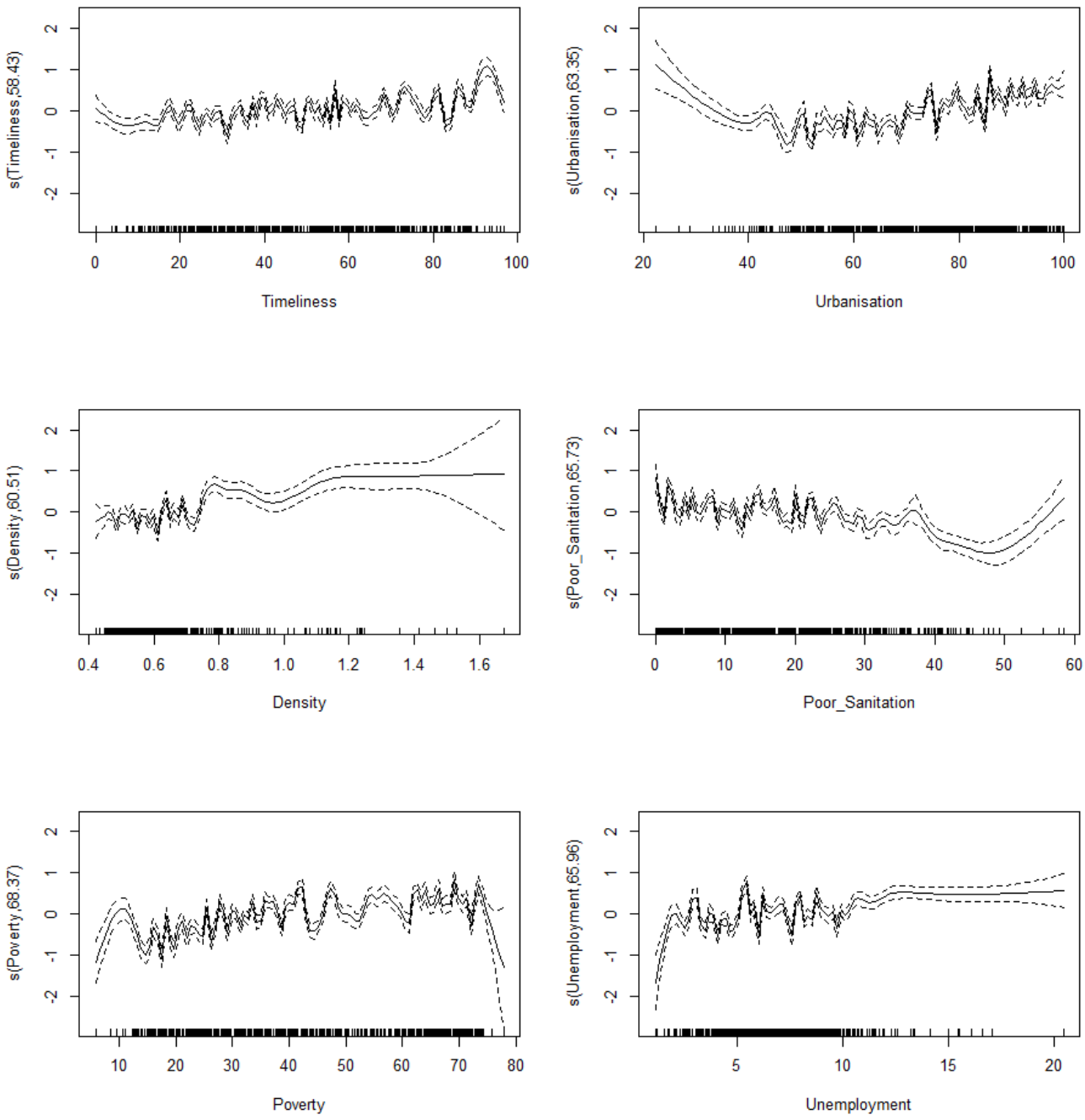

Figure 2: Plot of smooth functions of model covariates (excludes smooth function of Longitude & Latitude, see Fig. 4)

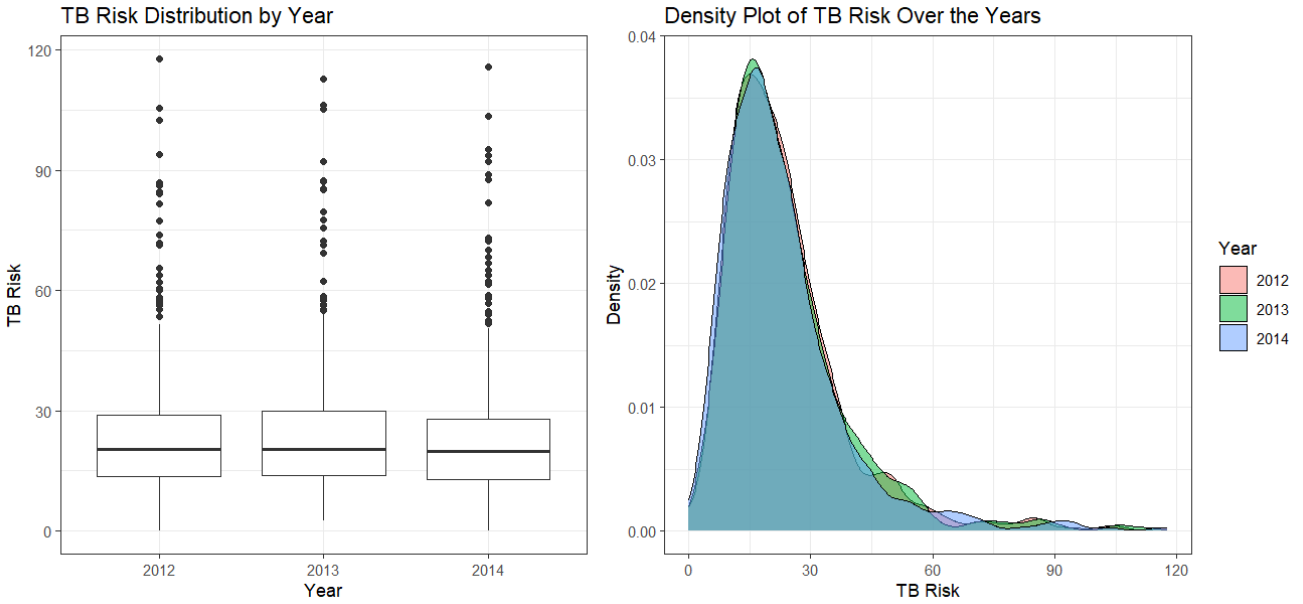

Figure 3: Plots of Temporal TB Risk. Left hand box plot represents TB risk distribution for each year. Right hand plot is analogous, further showing the distribution through a density plot. In both plots we observe similar distributions year-to-year with a right hand skew.

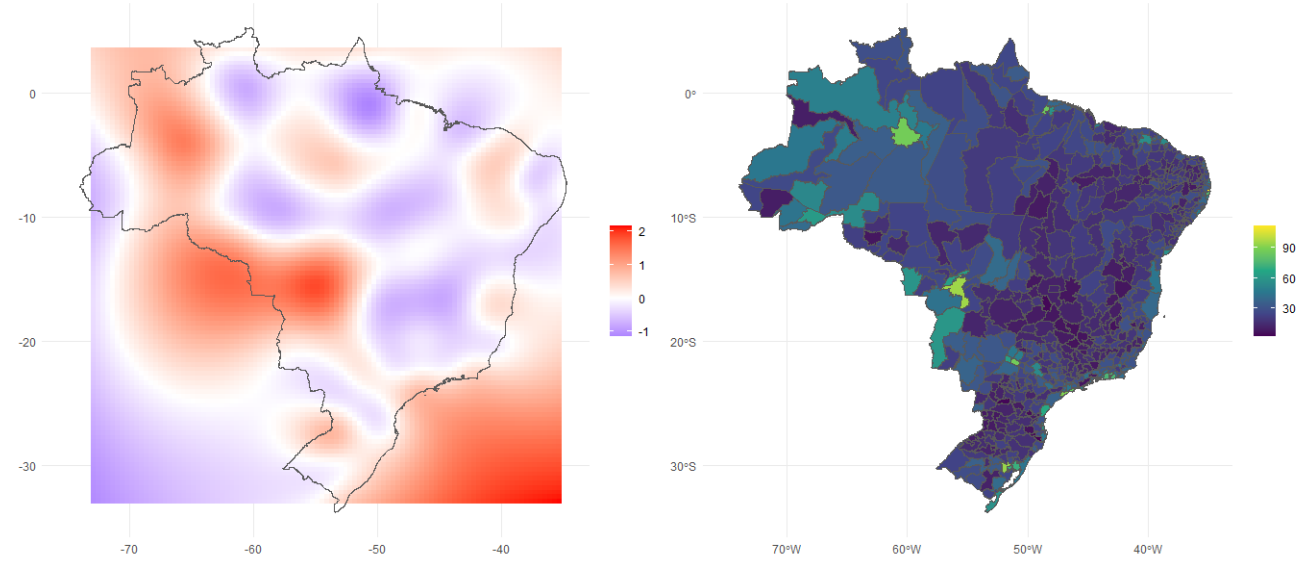

Figure 4: Plots of Spatial TB Risk. Left hand plot represents predicted Longitude/Latitude Smooth Function effect. 'Redder' areas represent positive effects on TB risk and 'Bluer' areas represent negative effects on TB risk. Right hand plot represents our model's predicted TB Risk per region.

References

[1] Emilie Alirol, Laurent Getaz, Beat Stoll, François Chappuis, and Prof Louis Loutan. Urbanisation and infectious diseases in a globalised world. *The Lancet Infectious Diseases*, 11(2):131–141, February 2011.

[2] Yohannes A. Melsew, Manoj Gambhir, Allen C. Cheng, et al. The role of super-spreading events in mycobacterium tuberculosis transmission: evidence from contact tracing. *BMC Infectious Diseases*, 19(1):244, 2019.

[3] Maria de Lourdes Sperli Geraldes Santos, Silvia Helena Figueiredo Vendramini, Claudia Eli Gazetta, Sonia Aparecida Cruz Oliveira, and Tereza Cristina Scatena Villa. Poverty: socioeconomic characterization at tuberculosis. *Revista latino-americana de enfermagem*, 15:762–767, 2007.

[4] Tereza CS Villa, Antônio Ruffino-Netto, Lucia M Scatena, Rubia LP Andrade, Maria EF Brunello, Jordana A Nogueira, Pedro F Palha, Lenilde D Sá, Marluce MA Assis, Silvia HF Vendramini, et al. Health services performance for tb treatment in brazil: a cross-sectional study. *BMC Health Services Research*, 11:1–8, 2011.

[5] P. Wilson. Is natural ventilation a useful tool to prevent the airborne spread of tb? *PLoS Med*, 4(2):e77, February 2007.

[6] Simon N. Wood. *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC, New York, 2nd edition, 2017. eBook published on May 3, 2017.

Appendix A

Table 1: **Dataset Variables**

Variable	**Description**	**Nature**
Indigenous	Proportion of indigenous population	Continuous
Illiteracy	Illiteracy level	Continuous
Urbanisation	Rate of urbanisation	Continuous
Density	Average dwellers per room	Continuous
Poverty	Poverty level	Continuous
Poor Sanitation	Indicator of sanitation levels	Continuous
Unemployment	Unemployment levels	Continuous
Timeliness	Average time between TB diagnosis and report	Continuous
Year	Year of metric measurement	Discrete
TB	Number of TB cases	Discrete
Population	Number of people	Discrete
Region	Unique ID to distinguish regions	Categorical
lat	Latitude of centriod	Continuous
lon	Longitude of centriod	Continuous

Appendix B

```
# CODE EXAMPLE OF MODEL BUILDING ITERATIVE PROCESS

# First model
model1 <- gam(TB ~ s(Timeliness, k=5, bs='cs') + s(Urbanisation, k=5, bs='cs') +
↪  offset(log(Population)), family=poisson(link='log'), data=TBdata)

# Residual plots and additional summary info
summary(model1)
par(mfrow=c(1,2))
plot(model1)
par(mfrow=c(2,2))
gam.check(model1)

# Second model
model2 <- gam(TB ~ s(Timeliness, k=20, bs='cs') + s(Urbanisation, k=20, bs='cs') +
↪  s(Density, k=20, bs='cs') + s(Poor_Sanitation, k=20, bs='cs') +
↪  offset(log(Population)), family=poisson(link='log'), data=TBdata)

# Residual plots and additional summary info
summary(model2)
par(mfrow=c(2,2))
plot(model2)
par(mfrow=c(2,2))
gam.check(model2)

# Third Plot
model3 <- gam(TB ~ s(Timeliness, k=40, bs='cs') + s(Urbanisation, k=40, bs='cs') +
↪  s(Density, k=40, bs='cs') + s(Poor_Sanitation, k=40, bs='cs') + s(Poverty, k=40,
↪  bs='cs') + s(Unemployment, k=40, bs='cs') + offset(log(Population)),
↪  family=poisson(link='log'), data=TBdata)

# Residual plots and additional summary info
summary(model3)
par(mfrow=c(3,2))
plot(model3)
par(mfrow=c(2,2))
gam.check(model3)

# Final plot
model4 <- gam(TB ~ s(Timeliness, k=60, bs='cs') + s(Urbanisation, k=66, bs='cs') +
↪  s(Density, k=62, bs='cs') + s(Poor_Sanitation, k=68, bs='cs') + s(Poverty, k=70,
↪  bs='cs') + s(Unemployment, k=67, bs='cs') + s(lat, lon, k=65, bs='tp') +
↪  offset(log(Population)), family=poisson(link='log'), data=TBdata)

# Residual plot and additional summary info
```

```
summary(model4)
par(mfrow=c(4,2))
plot(model4)
par(mfrow=c(2,2))
gam.check(model4)
```

```
# CODE EXAMPLE OF SPATIAL PREDICTION GEOPLOTS

# predicting risk and plotting against known data
risk_preds = predict(model7, newdata=TBdata, type='link') #this returns log of
↪  predicted cases
TBdata$pred_risk = (exp(risk_preds) / TBdata$Population) * 100000   #turn predicted
↪  cases into predicted risk (TB cases per 100000)

# create ggplot object
smooth_plot <- ggplot()+
  geom_tile(data=latlong,aes(x=lon,y=lat,fill=est))+ #tile to create heatmap
  xlim(-75,-35)+
  xlab("")+
  ylab("")+
  scale_fill_gradient2(low = "blue",mid="white", high = "red", name=" ")+
  geom_sf(data = brasil_outline,fill=NA,size=50)+
  theme_minimal()+
  theme(plot.title = element_text(hjust = 0.5))

pred_plot <- ggplot()+
  geom_sf(data=df,aes(fill=pred_risk),size=0.25)+
  xlim(-75,-35)+
  labs(fill=" ")+
  scale_fill_viridis_c()+
  theme_minimal()

(smooth_plot+pred_plot)+plot_layout(ncol=2)
```