# o teu pai de 4

## Quarto

Quarto enables you to weave together content and executable code into a finished document. To learn more about Quarto see https://quarto.org.

Let the model be :

$$Log(\lambda_i) = offset(log(Population_i)) + f_1(Indigenous_i) +$$

$$f_2(Illiteracy_i) + f_3(Urbanisation_i) + f_4(Density_i) + f_5(Poverty_i) +$$

$$f_6(PoorSanitation_i) + f_7(Unemployment_i) + f_8(Timeliness_i) +$$

$$f_9(Year_i) + f_{10}(Year_i, lon_i) + f_{11}(lat_i, Year_i),$$

$$u_i = E[TB_i] \quad and \quad TB_i \sim Pois(\lambda_i)$$

$$\log(\mu_i) = \beta_0 + f_1(x_{i1}) + f_2(x_{i2}) + ... + f_k(x_{ik}) + \log(\phi) + \log(y_i + r_i)$$

## Running Code

When you click the **Render** button a document will be generated that includes both content and the output of embedded code. You can embed code like this:

## List of Figures

## Introduction:

Tuberculosis (TB) is a bacterial disease that primarily affects the lungs but can also impact other parts of the body. It is a significant public health problem, with approximately 10 million cases reported globally in 2020. Brazil is a high-burden country for TB, with an estimated 96000 cases. The purpose of this report is to determine whether various socio-economic variables impact the rate of TB per unit population in Brazil between 2012 and 2014. We will analyse data from 537 micro-regions in Brazil, including the latitude and longitude of each region and the year in which the data was collected. @who_tb_2020.

The socio-economic variables that were recorded for each micro-region in our data set are as follows: the level of illiteracy, urbanisation, poverty, unemployment, and sanitation; the proportion of indigenous population; the dwelling density; and finally, a proxy indicator of the amount of resources in the form of the average amount of time between diagnosing a TB case and reporting it to the health system. In addition, the latitude and longitude of the respective 537 micro-regions, as well as the year in which the data was obtained, are supplied for each of these values. Because of this, we will be better able to explain the geographical, temporal, and spatio-temporal structure of any systematic risk that is not described by the covariates.

## Exploratory Data Analysis:

A simple exploration of our covariates and their potential relationships with the rate of tuberculosis in each microregion of Brazil was carried out before attempting any type of formal statistical analysis or regression on the data. Several of the findings were unexpected, such as the observation that a lower degree of sanitation did not appear to be associated with a higher rate of tuberculosis cases. The same was true in regard to the levels of poverty as it can been from figure 1 and figure 2.

Nevertheless, the issue with attempting to infer statistical associations in such a straightforward manner is that we are unable to take into consideration the possibility of changes occurring in other variables for each of the data points. This is the primary reason why we need to use a formal model to investigate the impact of our covariates on the incidence of tuberculosis in relation to the total population.

When developing statistical models, there is a trade-off between interpretability and flexibility. Linear models are often preferred due to their simplicity and ease of interpretation, but they may not accurately represent complex non-linear relationships in the data, such as in the case of tuberculosis data. On the other hand, machine learning models like neural networks or boosted trees can provide accurate predictions but are difficult to interpret and require a large amount of data. As a middle ground, generalized additive models (GAMs) were chosen as the preferred framework for analysis in this case. GAMs provide a balance between interpretability and flexibility, making them a suitable choice for this analysis.

## Model Selection:

At first, the idea was to combine a log link and a Poisson generalised additive model. This was because we were working with count data in the form of TB cases broken down by each microregion. Nonetheless, it was essential to standardise this count because the populations in each region were distinct from one another. After doing some study on the topic, we discovered that using the log of the population in the model is the most effective way to carry out a population standardised regression. While we were trying to fit our model in R, it was necessary for us to include an offset for the log of the population. This provides us with the tuberculosis rate per capita:

We defined our model as :

$$Log(\lambda_i) = offset(log(Population_i)) + f_1(Indigenous_i) +$$

$$f_2(Illiteracy_i) + f_3(Urbanisation_i) + f_4(Density_i) + f_5(Poverty_i) +$$

$$f_6(PoorSanitation_i) + f_7(Unemployment_i) + f_8(Timeliness_i) +$$

$$f_9(Year_i) + f_{10}(Year_i, lon_i) + f_{11}(lat_i, Year_i),$$

$$u_i = E[TB_i] \;\; and \;\; TB_i \sim Pois(\lambda_i)$$

Confirming that our model was accurate was the next step in the process of developing our model. Unfortunately, a QQ-plot revealed that the quantiles in our data did not closely resemble what they would look like if they followed a theoretical poisson distribution as it can be seen from figure 4. This was the conclusion that could be drawn from the plot. After calculating a Pearson estimate of our dispersion parameter to determine whether or not this lack of fit was due to over-dispersion, the results were unequivocal: the parameter was almost 9,33 (when it should have been 1), which indicated that our data was indeed over dispersed for a Poisson model.

An alternative model to Poisson is a Negative Binomial model. The Negative Binomial (NB) model is likewise suitable for use with count data; however, it differs from the Poisson regression in that it contains an additional parameter that can alter the variance independently from the mean.

Because of this, it is more flexible than the Poisson model, and as a result, it can fit data with a greater degree of fluctuation. A NB model was fitted to the data making use of the exact same specification as was used before. We can count ourselves fortunate that the QQ-plot as seen in figure 4 for the revised model showed a good fit (the majority of the dots fell on the y=x line), and the estimate for the dispersion parameter was 1,133. The AIC was also reduced for our newly developed model. We arrived at the following model after making adjustments to the choice of rank for each of our smooth terms (covariates), until we were certain that they were not too low and edf was not too similar to k'.

## Model Fitting and Results:

The most difficult aspect of using this model was figuring out how to account for the interplay that exists between time and place. It was required to add space and time as a product smooth, despite the fact that the majority of the other covariates were modelled using univariate smooth functions. After doing some research on the topic, we came to the conclusion that the interaction term should be incorporated into the model as a tensor product smooth utilising the "te" term in the mgcv package. Because space and time are measured on such distinct scales, this particular sort of interaction works best in situations in which the two important variables are on different scales.

The negative binomial gam model output shows the statistical significance of various predictors on the incidence of tuberculosis (TB) in a given population. The model assumes a negative binomial distribution with a log link function, indicating that the response variable, TB, has a count distribution and that the logarithm of the expected counts is modeled as a linear combination of the predictors.

The intercept term in the model is statistically significant ($p<0.001$), indicating that there is a significant difference in the TB incidence rate even when all other predictor variables are set to zero. The negative coefficient on the intercept suggests that there is a baseline level of protection against TB in the population.

Among the predictors, all except s(Year) are statistically significant at the 0.05 level. Indigenous status, Illiteracy, Poverty, and Timeliness have a statistically significant positive association with the incidence of TB. Urbanisation, Density, Poor Sanitation, and Unemployment are also significantly associated with TB but have a negative effect on incidence.

The s(Year) term has a statistically significant association with TB incidence but its effect is not linear. The term is modeled using a spline with a small number of degrees of freedom (k=3). The p-value for the term is <2e-16, indicating a strong association with TB incidence.

The interaction terms, te(lon,Year) and te(lat,Year), which model the effect of longitude and latitude on TB incidence with respect to year, respectively, are also statistically significant. Both terms have positive effects on TB incidence, indicating that the risk of TB increases with increasing longitude or latitude, depending on the year.

Overall, the model has an adjusted R-squared value of 0.895, indicating that the predictors in the model explain 53.4% of the variability in TB incidence. The scale estimate is 1, suggesting that the model is well-calibrated. The negative binomial distribution is appropriate for modeling the count response variable, and the log link function is appropriate for modeling the logarithm of the expected counts.

## Model Evaluation and Discussion:

First, lets discuss the spatial findings of our analysis. As we can see from the yearly graph, figure 5, the analysis shows that there is a gradual decrease of TB cases per capita from 2012 to 2013 at the national level. In contract from 2013 to 2014 the data displays very little decrease of TB risk on a national level.

Secondly, our spatial analysis discovered that the risk of Tb increases the more you deviate from the geological Eastern central region of Brazil, with the risk of Tb increasing up to 33% at its maximum on the south-western border region of Brazil. This increase in Tb risk is much more prolonged longitude wise mainly due to the center of lower TB risk is nearly centrally located in latitude, but skewed significantly to the east in longitude. From the figure 6 we can also observed the risk of TB increases quicker on the latitude scale then on the longitude scale.

Thirdly, merging the 2 previous analysis into the spatio-temporal we can see a more significant decrease of cases overall from 2012 to 2013 in all microregions which is very evident on the figure 3. In the geographical center of Brazil and the Amazonian outskirts in both years where the coastal regions have a some improvements from 2012 to 2013 but the risk of TB was either nearly unchanged or did increase from 2013 to 2014.

Looking at the figure 7 we can identify that regions with the lowest longitude and latitude have significantly higher TB risk, however regions with longitude bigger than the center of the lowest risk demonstrate a risk level that is quite similar to that of the low-risk area. Regions with latitude bigger than the lowest center have a gradual increase in TB risk. These effects have remained stable during the 3 years analysed

## Conclusion:

In summary, the gam model identifies several significant predictors of TB incidence, including Indigenous status, Illiteracy, Poverty, and Timeliness, as well as Urbanisation, Density, Poor Sanitation, and Unemployment. The interaction terms with latitude and longitude also play a significant role in predicting TB incidence. The model can be used to identify populations at risk for TB and to develop interventions to reduce TB incidence in these populations.

In conclusion, we discuss some of the restrictions imposed in this study. The relatively little time frame under consideration presents the most evident limitation of the study. If determining any temporal structure of systematic risk was the objective, then additional time ought to elapse prior to the data being examined and modelled in order to allow for the passage of time. Because the incidence and rate of tuberculosis are probably affected by a wide variety of other factors that are not accounted for in our dataset, additional covariates could also be added to the study. Additional limitations of our study are associated with the application of a generalised additive model (GAM), including the risk that our model overfits our data, as well as its high computing cost and complexity.

## Figures

## Code Appendix

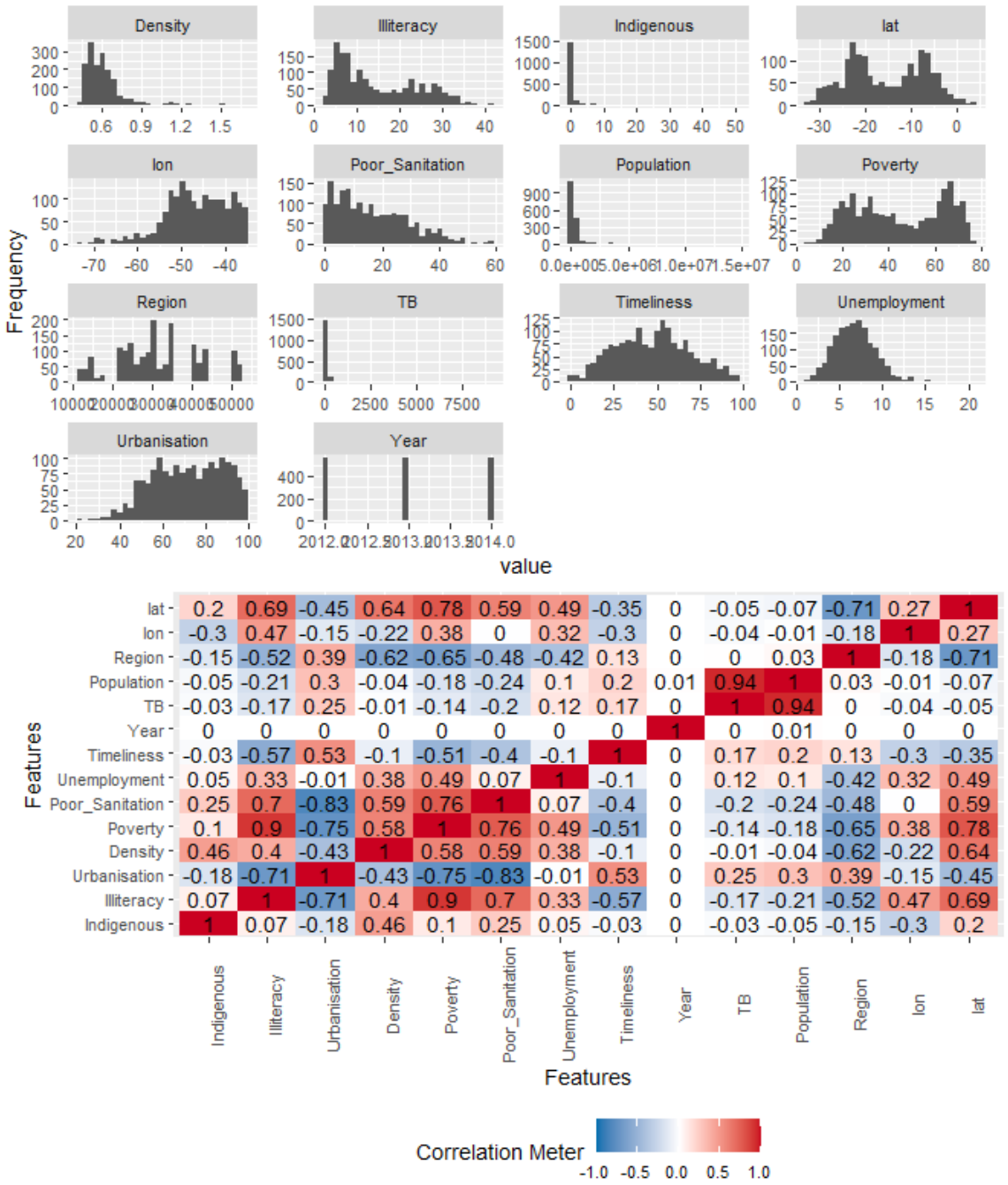


Fig 1: Brazil TB 2012-2014 data exploratory Analysis

| | Indigenous | Illiteracy | Urbanisation | Density | Poverty | Poor_Sanitation | Unemployment |
|---|---|---|---|---|---|---|---|
| X | Min. : 0.01034 | Min. : 2.336 | Min. :22.34 | Min. :0.4223 | Min. : 5.923 | Min. : 0.0466 | Min. : 1.128 |
| X.1 | 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 |
| X.2 | Median : 0.10577 | Median :11.516 | Median :72.66 | Median :0.5840 | Median :42.603 | Median :13.9129 | Median : 6.782 |
| X.3 | Mean : 0.84307 | Mean :14.802 | Mean :71.96 | Mean :0.6212 | Mean :44.371 | Mean :16.4490 | Mean : 6.930 |
| X.4 | 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 |
| X.5 | 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 |
|---|---|---|---|---|---|---|---|
| X | Min. : 0.00 | Min. :2012 | Min. : 0 | Min. : 23966 | Min. :11001 | Min. :-72.86 | Min. :-32.865 |
| X.1 | 1st Qu.:31.29 | 1st Qu.:2012 | 1st Qu.: 17 | 1st Qu.: 105054 | 1st Qu.:24007 | 1st Qu.:-50.95 | 1st Qu.:-22.278 |
| X.2 | Median :48.36 | Median :2013 | Median : 35 | Median : 180423 | Median :31028 | Median :-46.31 | Median :-15.889 |
| X.3 | Mean :47.67 | Mean :2013 | Mean : 125 | Mean : 357768 | Mean :31329 | Mean :-46.42 | Mean :-15.240 |
| X.4 | 3rd Qu.:62.58 | 3rd Qu.:2014 | 3rd Qu.: 76 | 3rd Qu.: 315440 | 3rd Qu.:41007 | 3rd Qu.:-40.64 | 3rd Qu.: -7.380 |
| X.5 | Max. :96.69 | Max. :2014 | Max. :9097 | Max. :14597964 | Max. :53001 | Max. :-34.95 | Max. : 3.486 |

Fig 2: Table of summaries of the variables

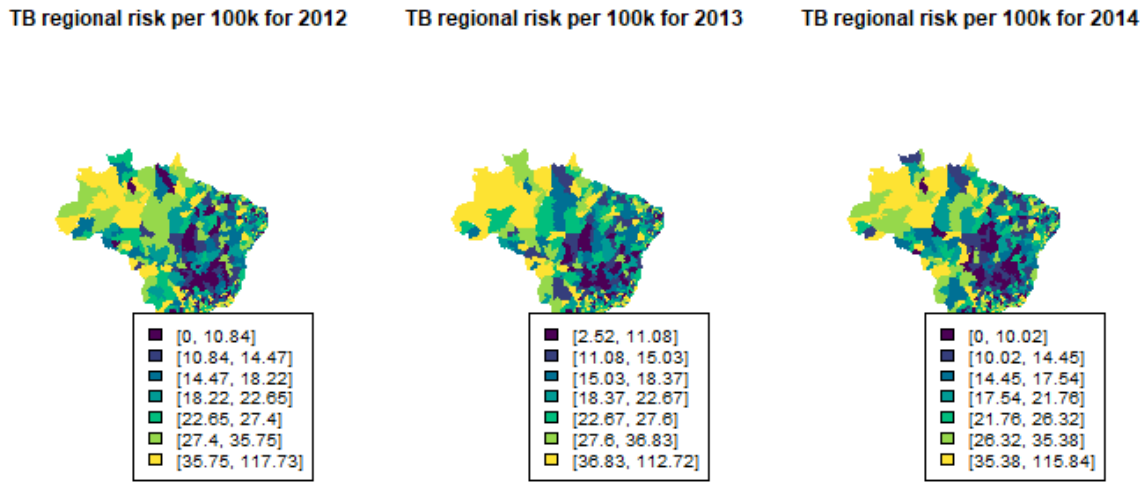


Fig 3: Temporal-Spatial Analysis of risk of TB per 100000 per Year per region

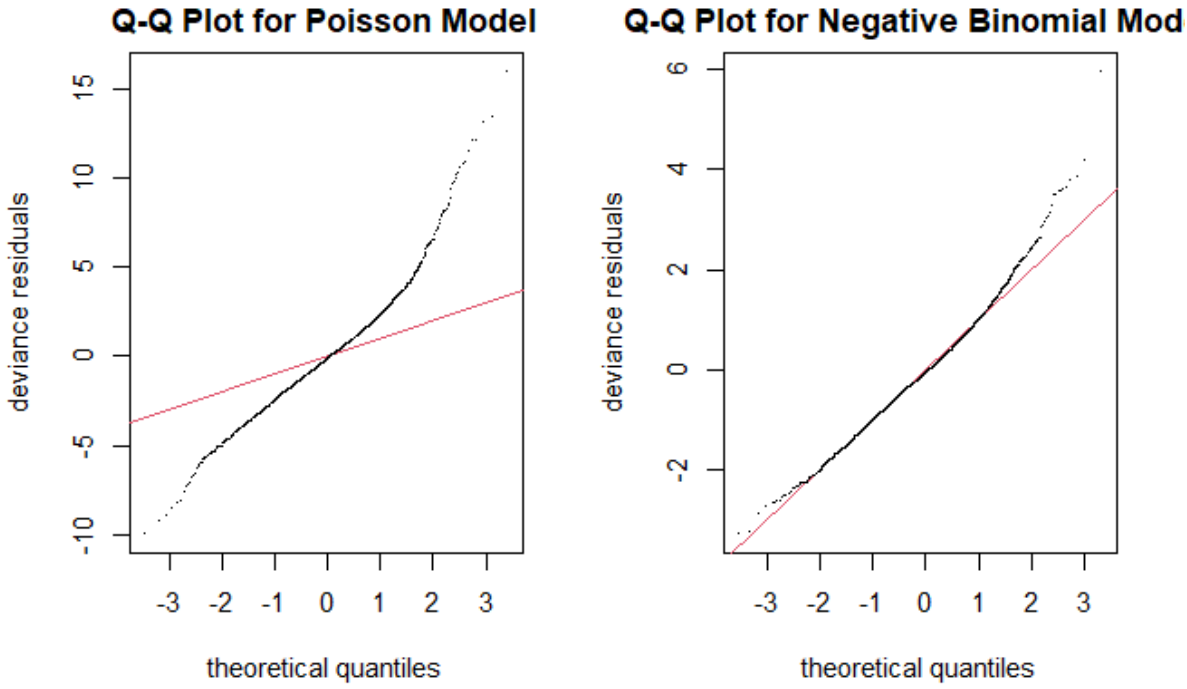


Fig 4: QQ-Plot of our distributions

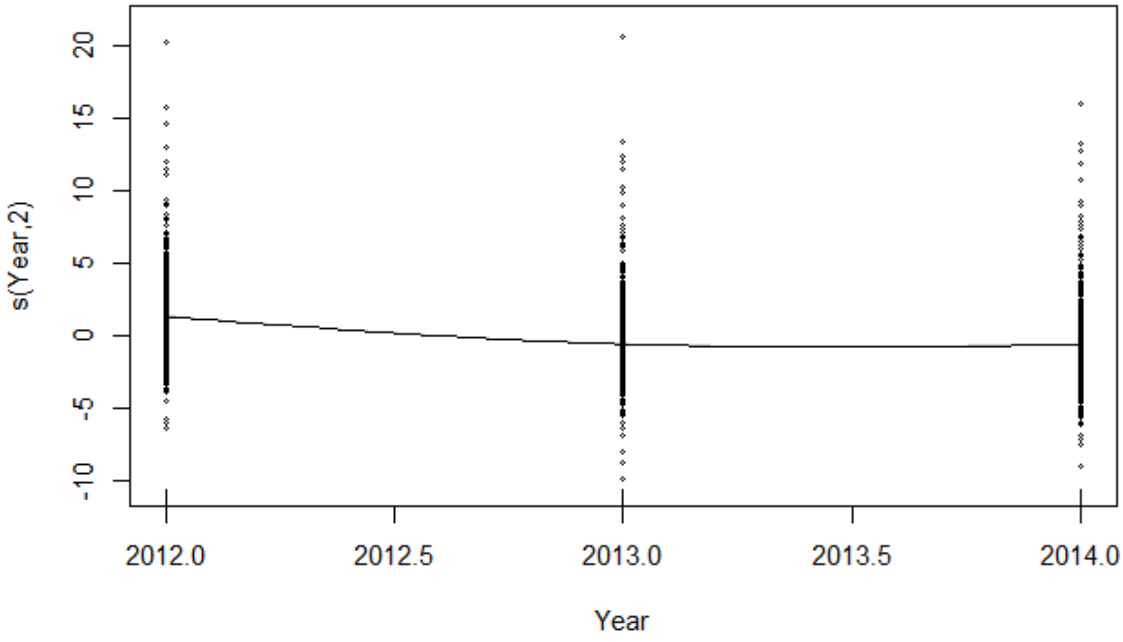


Fig 5: Temporal Analysis of risk of TB per Year

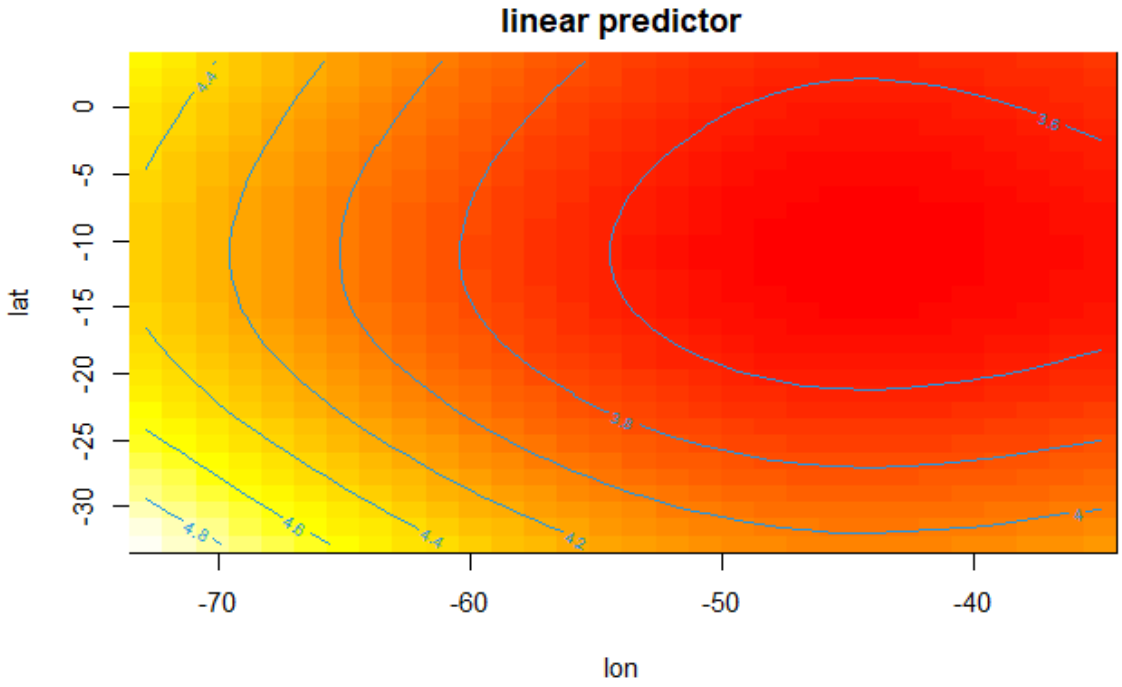


Fig 6: Spatial Analysis of risk of TB per Year per region

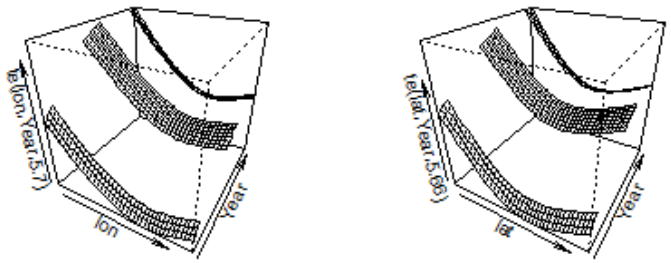


Fig 7: Temporal-Spatial Analysis of risk of TB per 100000 per Year per region