MTHM506 Project

Quantifying spatio-temporal risk from TB in Brazil

For this project, you are required to conduct an independent analysis of tuberculosis (TB) data originating from Brazil, using Generalized Additive Models (GAMs). A substantial proportion of the marks are allocated to independent learning, specifically in the use and understanding of GAMs that goes well beyond the scope of what was taught in the module.

Brazil is divided into 557 administrative microregions and the available data comprises of counts of TB cases in each microregion for each of the years 2012-2014. The R workspace `TB.RData` contains the relevant data set in the form of a dataframe `TBdata`, which contains the following variables:

No Change

1. ~~`Indigenous`~~: proportion of indigenous population (the higher the number, the more the indigenous population) in each microregion;
2. ~~`Illiteracy`~~: a continuous measure of illiteracy levels per microregion, the higher the number the more the illiteracy;
3. `Urbanisation`: the rate of urbanisation of a microregion;
4. `Density`: dwelling density (average dwellers per room) in each microregion;
5. `Poverty`: a continuous measure of poverty in each microregion (the higher the number, the higher the poverty levels);
6. `Poor_Sanitation`: a continuous indicator of sanitation levels in each microregion (the higher the number the poorer the sanitation);
7. `Unemployment`: unemployment levels in each microregion (high values indicate more unemployment);
8. `Timeliness`: timeliness of notification, the average amount of time between diagnosing a TB case and reporting it to the health system - this is a proxy measure for the amount of resources in each microregion.
9. `Year`: The year (2012–2014);
10. `TB`: The number of TB cases in each microregion for the corresponding year;
11. `Population`: The number of people living in each microregion and year;
12. `Region`: A unique ID number to distinguish the 557 regions;
13. `lon`: the longitude of the centroid of the microregion;
14. `lat`: the latitude of the centroid of the microregion;

The aim of this project is to use this data set to quantify TB risk across Brazil over the 3 years, where risk is defined as the rate of TB cases per unit population. The health authorities would like to know whether any of the socio-economic covariates (1–8 above) are significantly affecting the rate of TB per unit population, and if so, in what way. In addition, they would like to understand the a) spatial, b) temporal, and c) spatio-temporal structure of risk that is not explained by covariates 1–8 (i.e. the spatio-temporal effects).

You are expected to use the GAM framework to analyse the data and thus attempt to answer the questions above. You should find chapter 7 of Wood (2017) useful, as well as the `R` help file on `gam.models` (obtained by typing `?gam.models` in `R`, after having loaded the library `mgcv`).

Note, the workspace `TB.RData` also contains an object `brasil_micro` and a function `plot.map`. The latter uses the former to produce a map of Brazil to plot a vector a values. You may find this useful when exploring the data and presenting the results. Specifically, the function `plot.map` takes the following arguments:

- `x`: a vector of 557 values in the order given in `TBdata` (one for each microregion) to be plotted;
- `n.levels`: number of categories that the range of `x` will be divided into, to produce the colour scheme. These are calculated as equidistant empirical quantiles, default is 4;
- `main`: a character string to be used as a title on the plot.
- `cex`: a number used to manipulate the size of the legend, default is 1.

You can of course edit this function as you see fit. As an example demonstrating its use, Figure 1 shows the TB counts in each microregion for the year 2014, using:

```
> library(fields)
> library(maps)
> library(sp)
> plot.map(TBdata$TB[TBdata$Year==2014],n.levels=7,main="TB counts for 2014")
```

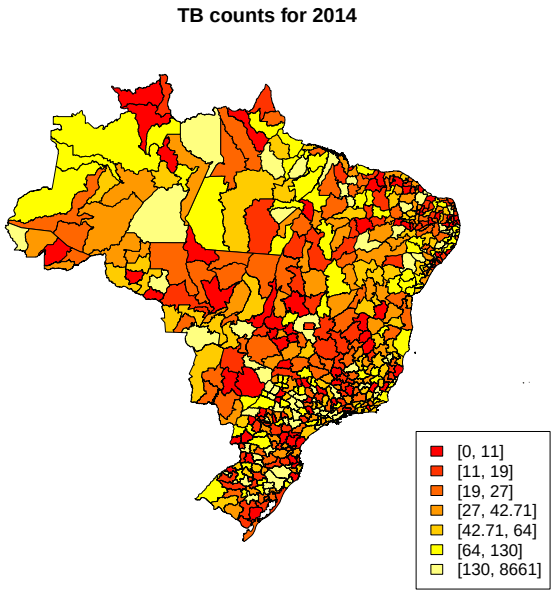

Figure 1: Map of counts across microregions for the year 2014.

Instructions and guidelines for producing the report

A professional word processed report presenting your analysis will be required to be submitted for assessment by the indicated deadline. Marks achieved in this assignment will contribute 50% of the final module mark. Your answer should comprise at most **two sides of text** followed by as many pages of figures or tables or R code that you feel is appropriate. Throughout your report you must use A4 paper and a font size of at least 11 points, while lines must be single spaced. No credit will be awarded to additional pages of text.

The project is designed to assess your knowledge obtained from the taught aspect of the module, as well as your ability to learn new material. There are 50 marks, and a brief outline of the marking criteria is given below with approximate marks:

- Understanding and exploration of both the problem and the data (5).

Modeling —
- Thoroughness and rigour, e.g. clear mathematical description of models (10).
- Clear exposition of the steps you took in model fitting and exposition of a final model (10).

Plots & Inference —• Clear presentation and interpretation of results (15).

Evaluation —• Critical review of the analysis (5).

- Clarity and conciseness in writing and tidy presentation of R code and associated plots (5).

This is an individual research project and while I will happy provide support (particularly with respect to R), you are expected to work on this by yourselves.

References

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

Model Fitting: [10m]
— Saturated/Null Model
— Model Extension/Reduction
— Final Model Fully Explained

Mathematical: [10m]
— Starting Model Formulae
— Rigor: Offset by Pop, Time, and Region
— Final Model Formulae

Presentation [15+5m]
— Brazil Map Plot
— Plots of Smooths
—