MTHM506 - Statistical Data Modelling: Individual Project

James R Lewis

March 2025

Contents

Declaration of AI Assistance: I have used OpenAI's ChatGPT tool in creating this report.

AI-supported/AI-integrated use is permitted in this assessment. I acknowledge the following uses of GenAI tools in this assessment:

1. I have used GenAI tools to check and debug my code.
2. I have used GenAI tools to proofread and correct grammar or spelling errors.
3. I have used GenAI tools to give me feedback on a draft.

I declare that I have referenced use of GenAI outputs within my assessment in line with the University referencing guidelines.

1 Introduction

- **Objective:** Clearly state the aim of the analysis—to quantify TB risk across Brazil (2012-2014), identify socio-economic covariates affecting TB rates, and understand spatial, temporal, and spatio-temporal structures.
- **Relevance:** Discuss the public health importance and resource allocation implications.
- **Data Overview:** Briefly introduce the dataset, its source, and key variables (e.g., TB counts, socio-economic covariates, spatial and temporal identifiers).

2 Exploratory Data Analysis (EDA)

- **Descriptive Statistics:** Summarise key covariates and TB rates (use one table for this).
- **Visual Exploration:** Present maps or graphs that show:
  - TB case distributions over time and across regions.
  - Potential relationships between TB rates and socio-economic covariates.
- **Spatial-Temporal Trends:** Use the provided `plot.map` function to visualise spatial TB distributions across years.

3 Methodology

- **GAM Framework:**
  - Introduce GAMs as an extension of GLMs (refer to Topic 3 notes).
  - Discuss the semi-parametric nature, choice of smoothing, and penalisation for smoothness.
  - Explain the model structure, including:
    - Response variable: TB rate (TB cases per unit population).
    - Covariates: Socio-economic factors and spatial-temporal indicators.
    - Link function and distribution assumption (likely Poisson or Negative Binomial, given count data)
- **Model Formulation:** Provide the mathematical formulation of your GAMs, including how spatial and temporal structures are modelled.
- **Justification of Choices:** Explain why GAMs are appropriate, especially for non-linear, spatial-temporal relationships.

4 Model Fitting

- **Model Development Steps:**

  - **Initial Model:** Fit a base model with socio-economic covariates.
  - **Spatial and Temporal Effects:** Extend the model to include smooth terms for spatial (latitude, longitude) and temporal (year) effects.
  - **Spatio-Temporal Interaction:** Test for interactions if appropriate.
- **Model Selection and Evaluation:** Discuss criteria used for selecting the best model (e.g., AIC, GCV, residual analysis).
- **Diagnostics:** Present key diagnostics (e.g., residual plots, deviance explained).

5 Results

- **Summary of Findings:** Interpret the key model outputs:
  - Which covariates significantly influence TB rates?
  - How do spatial and temporal trends appear?
  - Are there high-risk regions or periods for TB?
- **Visual Representation:**
  - Include maps showing predicted TB risks.
  - Temporal trend plots.
  - Any significant interaction effects (if modelled).
- **Implications:** Discuss practical implications, particularly for resource allocation.

6 Critial Review

- **Model Limitations:** Reflect on any limitations (e.g., assumptions, data gaps, potential overfitting).
- **Alternative Approaches:** Briefly discuss other modelling approaches that could be considered in the future.
- **Uncertainty Discussion:** Highlight any uncertainties in parameter estimation, especially regarding smooth terms.

7 Conclusion

- **Summary of Key Insights:** Recap significant findings and their implications.
- **Policy Recommendations:** Suggest actionable steps for health authorities based on the analysis.

7.1 Appendix:

- **Commented R Code:** Include all key code used for analysis and modelling.
- **Additional Figures/Tables:** Any extra plots or results not included in the main text.
- **Model Summaries:** Full output of final model summaries.

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.

Tables:

1. Summary Stats Table (gt table, look at website)
2. Model Comparison (Tweedie vs nb)
3. Table for regions iwth highest rate of TB, and infers from this

Plots:

1. Residual Diagnostic plots: Residual vs fitted QQ plots and 1 other? Influece plot?
2. Covariate plot 1-8 covaraite impact on rate of TB per unit pop. Compare the raw mean to predicted mean.
3. Covariate effect plot: partial effects plots to show significant covariates
4. Spatial structure explaining risk: plot.map for this. General plot without temporal noise
5. Temporal structure explaining risk: Covariate charts/significance for each year
6. temporal-Spatial structure explaining risk: plot.map for this. 3 of them, Look at differences

The health authorities want to allocate resources for hospitals to cope with the TB cases , so they would like to know if there are regions where the rate of TB per unit population is high and where you would recommend allocating these resources.

```
library(fields)
```

```
## Warning: package 'fields' was built under R version 4.4.3
```

```
## Loading required package: spam
```

```
## Warning: package 'spam' was built under R version 4.4.3
```

```
## Spam version 2.11-1 (2025-01-20) is loaded.
## Type 'help( Spam)' or 'demo( spam)' for a short introduction
## and overview of this package.
## Help for individual functions is also obtained by adding the
## suffix '.spam' to the function name, e.g. 'help( chol.spam)'.
```

```
##
## Attaching package: 'spam'
```

```
## The following objects are masked from 'package:base':
##
##     backsolve, forwardsolve
```

```
## Loading required package: viridisLite
```

```
##
## Try help(fields) to get started.
```

```
library(maps)
```

```
## Warning: package 'maps' was built under R version 4.4.3
```

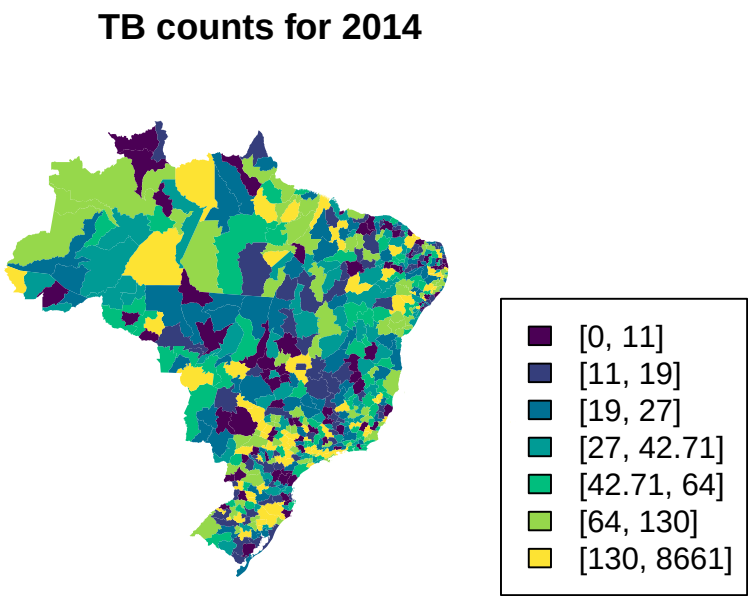
```
##
## Attaching package: 'maps'
```

```
## The following object is masked from 'package:purrr':
##
##     map
```

```
library(sp)
```

```
## Warning: package 'sp' was built under R version 4.4.3
```

```
# PLotting map of cases
plot.map(TBdata$TB[TBdata$Year==2014],n.levels=7,main="TB counts for 2014")
```



Structure:

1. Exploratory Data Analysis (EDA)

- **Descriptive Statistics:** Summary of key covariates and TB case rates (per unit population).

Write about Table 1

And implications going into this.

7.2 2. Initial Model Fitting (GAMs)

- **Base Model:** Fit an initial GAM including socio-economic covariates.
- **Model Structure Explanation:** Detail the mathematical formulation and rationale for including each covariate.
- **Model Summary:** Present key outputs (e.g., significant predictors, smooth terms).

We will model **TB case counts** Y_i as a function of socio-economic covariates and spatial-temporal effects using **Generalized Additive Models (GAMs)**. Since TB cases are **count data**, they follow a **Poisson or Negative Binomial (NB) distribution**, which makes the following GAM structure appropriate:

7.3 Model Specification

We assume that the observed TB case counts Y_i follow either a Poisson or a Negative Binomial distribution:

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

where the expected mean μ_i is modeled as:

$$\log(\mu_i) = \beta_0 + f_1(\text{Indigenous}_i) + f_2(\text{Illiteracy}_i) + \cdots + f_p(\text{Timeliness}_i) + f_{\text{spatial}}(\text{lon}_i, \text{lat}_i) + f_{\text{temporal}}(\text{Year}_i) + \log(\text{Pop}$$

Here:

- $f_j(\cdot)$ represents smooth functions capturing non-linear effects of the socio-economic covariates.
- $f_{\text{spatial}}(\text{lon}, \text{lat})$ models spatial variability.
- $f_{\text{temporal}}(\text{Year})$ accounts for temporal trends.
- The $\log(\text{Population}_i)$ term acts as an offset to adjust for population size.

This formulation allows us to estimate the impact of socio-economic factors on TB case counts while incorporating spatial and temporal structures in the data.

7.3.1 Choice of Distribution: Poisson vs. Negative Binomial

7.3.1.1 Limitations of the Poisson Model The Poisson model assumes that the mean and variance of the response variable are equal:

$$E[Y] = \text{Var}(Y)$$

However, this assumption is often violated in real-world count data due to **overdispersion**, where the variance exceeds the mean. Overdispersion can lead to:

- **Inflated test statistics**, increasing the risk of Type I errors (false positives).

- **Misleading inferences**, as confidence intervals may be too narrow.

7.3.1.2 Evidence of Overdispersion The computed dispersion statistic for the TB data is **2223.535**, which is significantly greater than 1. This confirms severe overdispersion, violating the equidispersion assumption of the Poisson model:

$$\text{Var}(Y) \gg E[Y]$$

Using a Poisson model under such conditions would underestimate variability, leading to unreliable statistical conclusions.

7.3.1.3 Why the Negative Binomial Model is More Appropriate The **Negative Binomial (NB) model** extends the Poisson distribution by introducing an **overdispersion parameter** θ, which adjusts the variance function to account for excess dispersion:

$$\text{Var}(Y_i) = \mu_i + \frac{\mu_i^2}{\theta}$$

where:

- μ_i is the expected number of TB cases for region i, - θ controls the degree of overdispersion, allowing for greater variability than the Poisson model.

Given the extreme overdispersion in the TB data, the Negative Binomial model is the **statistically robust choice**, as it:

- **Relaxes the restrictive assumption of equal mean-variance** while remaining within the exponential family framework.

- **Provides more reliable parameter estimates** by accounting for unobserved heterogeneity across microregions.

- **Prevents misleading inference** by correcting for variance inflation.

7.3.1.4 Justification for Directly Using a Negative Binomial GAM

Rather than initially fitting a Poisson GAM and later diagnosing overdispersion, we **directly proceed with a Negative Binomial GAM**. This ensures:

- **Model validity**, by selecting an appropriate variance structure from the outset.

- **Robust inference**, reducing the risk of biased estimates and misleading conclusions.

- **Improved accuracy**, as the model better captures the heterogeneity in TB case counts across microregions.

By making this adjustment, we ensure a more reliable and interpretable analysis of TB risk in Brazil.

7.4 3. Offset for Population

Since TB cases depend on the **population size** in each region, we introduce an **offset** to model TB risk per capita:

$$\log(\mu_i) = \text{linear predictor} + \log(\text{Population}_i)$$

This ensures that the response variable is scaled appropriately, effectively modeling the **rate of TB cases per 100,000 people** rather than just raw counts.

7.5 Negative Binomial GAM for TB Case Counts

Given the substantial **overdispersion** in the TB case counts, we proceed with a **Negative Binomial (NB) GAM**. The model accounts for **spatial, temporal, and socio-economic effects** while incorporating an **offset term** to model TB risk per unit population.

7.5.1 Mathematical Formulation

The **Negative Binomial GAM** assumes that the response variable Y_i (TB case count in region i) follows a **Negative Binomial distribution**:

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

where:

- μ_i is the expected number of TB cases in region i. - θ is the **overdispersion parameter**.

The **log link function** ensures a **multiplicative relationship** between the covariates and the expected count:

$$\log(\mu_i) = \beta_0 + f_1(\text{Indigenous}_i) + f_2(\text{Illiteracy}_i) + \cdots + f_p(\text{Timeliness}_i) + f_{\text{spatial}}(\text{lon}_i, \text{lat}_i) + f_{\text{temporal}}(\text{Year}_i) + \log(\text{Pop}$$

The **offset term** $\log(\text{Population}_i)$ ensures that the model accounts for diffent population sizes.

7.6 Transforming Predictions to TB Incidence Rates

Since the **Negative Binomial GAM** inherently predicts **TB case counts**, we manually convert the **fitted values** into **TB incidence rates per 100,000 population** using:

$$\text{Predicted TB Rate} = \left(\frac{\text{Predicted TB Cases}}{\text{Population}} \right) \times 100000$$

This transformation ensures that the estimated values are correctly **interpreted as TB incidence rates** rather than raw counts.

##########################

7.7 Understanding the Model Output

Even though we include an **offset term** log(Population) in the model to account for different **population sizes**, the model still fundamentally **predicts TB case counts**.

7.7.1 Mathematical Formulation

The mathematical form of the model is:

$$\log(E[Y_i]) = \beta_0 + f_1(\text{Indigenous}_i) + f_2(\text{Illiteracy}_i) + \cdots + f_p(\text{Timeliness}_i) + f_{\text{spatial}}(\text{lon}_i, \text{lat}_i) + f_{\text{temporal}}(\text{Year}_i) + \log(\text{P}$$

where: - $E[Y_i]$ = expected **TB case count** in microregion i, - $f_j(\cdot)$ = **smooth functions** of socio-economic covariates, - $f_{\text{spatial}}(\cdot)$ = **spatial effect**, - $f_{\text{temporal}}(\cdot)$ = **temporal effect**, - $\log(\text{Population}_i)$ = **offset term** to adjust for population differences.

Since the model uses a **log link function**, the expected case count is given by:

$$E[Y_i] = \text{Population}_i \times e^{\text{Linear Predictor}}$$

7.7.2 Implication

- The model's predictions are in terms of **TB case counts**, adjusted for population.
- It **does not directly predict TB incidence rates**.
- To obtain **TB rates**, we must **manually convert** the predicted case counts.

7.8 2. Why Does This Happen Despite the Offset?

The **offset does not change** the response variable from **counts to rates**. Instead, it ensures that the estimated **case counts are proportional to the population size**.

7.8.1 Key Insights

- If two regions have the **same covariate values** but **different populations**, the model will predict **higher TB case counts** for the region with the larger population.
- The offset ensures that **TB risk per person is correctly estimated**, but the raw predictions remain in **case counts**.

Thus, while the model **adjusts for population**, it does not inherently provide **incidence rates**, requiring us to **explicitly transform** the predictions.

##########################

7.8.2 Fitting the Model in R

We use the `mgcv` package to fit a **Negative Binomial GAM**, specifying the **offset for population size**:

7.9 3. Model Refinement and Improvement

- **Add Complexity:** Introduce spatial and temporal smooths to capture unexplained variation.
- Explain why some terms are included and some not. why some are formatted as they as.. Illiterecy seems to have no significant correlation as seen from a plot.
- **Interaction Terms:** If necessary, test for spatio-temporal interactions.
- **Summary Table:** Present a table comparing models (AIC, deviance explained, etc.) to show the progression and improvements.

7.10 4. Model Diagnostics

- **Residual Diagnostics:** Plot residuals to assess model fit (QQ plots, residual vs fitted, etc.).
- **Check Assumptions:** Comment on model assumptions like distributional assumptions, independence, and overfitting.

7.11 5. Covariate Effect Plots (Already created)

- **Effect Plots for Covariates:** Show smooth effect plots (e.g., `plot.gam` in R) for key socio-economic variables to visualise their impact on TB risk.
- **Interpretation:** Discuss how each covariate affects TB risk and the strength of their influence.

7.12 6. Spatial and Temporal Structure Analysis

- **Spatial Smoother Visualization:** Plot the estimated spatial effects to identify regions with high or low TB risk.
- **Temporal Smoother Visualization:** Show how TB rates have evolved over time.

7.13 7. Geospatial Plots

- **Risk Maps:** Use the `plot.map` function to produce maps of predicted TB risks across microregions.
- **Highlight High-Risk Areas:** Identify and discuss regions with consistently higher TB rates.

7.14 9. Model Comparison

- **Performance Summary:** Compare final models based on criteria like AIC, deviance explained, and residual diagnostics.
- **Final Model Selection:** Clearly state which model is preferred and why.

7.15 10. Conclusion and Recommendations

- **Key Insights:** Summarise the main findings regarding socio-economic influences, spatial-temporal risks, and high-risk areas.
- **Policy Implications:** Discuss recommendations for health authorities, such as where to allocate resources.