Modelling Spatio-Temporal Risk of Tuberculosis in Brazil Using Generalised Additive Models

James Lewis

21 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.
4. I have used GenAI tools to assist formatting figures and tables.

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

1 Introduction

Brazil is among the top 30 countries with the highest tuberculosis (TB) rates globally (Taveres, 2024), with substantial regional disparities driven by socio-economic and geographical inequalities. Understanding these spatio-temporal patterns, and identifying socio-economic factors influencing TB risk, is critical for public health planning.

This study aims to: (1) quantify TB incidence per 100,000 population across Brazil's 557 microregions (2012–2014); (2) identify significant socio-economic predictors of TB; and (3) assess residual spatial, temporal, and spatio-temporal patterns in TB risk. We apply ***Generalised Additive Models*** (GAMs), which flexibly model non-linear effects and spatial structure using penalised regression splines (Wood, 2017).

2 Data and Modelling Approach

This analysis uses the *TBdata* dataset, which records annual TB case counts and population estimates for 557 microregions from 2012 to 2014. Each observation corresponds to a region-year combination, yielding a balanced panel of 1,671 records.

Eight socio-economic covariates are included: proportion of Indigenous population, Illiteracy rate, Urbanisation, Dwelling Density, Poverty, Poor Sanitation, Unemployment, and Timeliness of TB notification. See **Table 1** for summary statistics. These variables capture structural inequality and living conditions across regions, with wide variation observed — for instance, Indigenous population share ranges from **0.01**% to **50.65**%, and Poverty from **5.92**% to **77.88**%. Spatial coordinates (*longitude* and *latitude*) and a time indicator (*Year*) allow for analysis of both geographic clustering and evolving temporal trends in TB incidence.

The response is TB case count, with population included as a *log-offset* to model incidence rates. This standardises for regional population differences, aligning the analysis with TB risk per capita. Preliminary analysis showed **overdispersion** (statistic of **2223**), making Poisson models inappropriate; a **Tweedie distribution** was used instead, flexibly bridging Poisson and Gamma models and accommodating zero-inflation and skewness.

GAMs extend GLMs by modelling non-linear covariate effects via *penalised regression splines*, enabling flexible, interpretable estimation of smooth terms. Spatial smooths were fit using t*hin-plate splines*, and time-varying effects were incorporated using `by = Year` interaction terms.

3 Model Specification and Fitting

This section outlines the mathematical formulation and model fitting process for estimating TB risk across Brazil's microregions. The response variable *Yi* denotes TB case counts, with population included as a *log-offset*. A ***Tweedie distribution*** was chosen to model overdispersed count data with skewness and excess zeros.

The final model is expressed as:

$$Y_i \sim \text{Tweedie}(\mu_i, \phi, p)$$

$$\log(\mu_i) = \log(\text{Population}_i) + \beta_1 \text{Year}_i + \beta_2 \text{Indigenous}_i + \sum_{j=1}^{6} f_j(X_{ji}) + f_{\text{spatial}}(\text{lon}_i, \text{lat}_i)$$

Here, the term $\sum_{j=1}^{6} f_j(X_{ji})$ represents smooth functions of six socio-economic covariates: Urbanisation, Density, Unemployment, Poor Sanitation, Poverty, and Timeliness. The function $f_{\text{spatial}}(\text{lon}_i, \text{lat}_i)$ is a *bivariate thin-plate spline* capturing residual spatial structure. Smoothness was controlled using penalised splines, with selection performed via *REML* and additional penalties applied using `select = TRUE`. Covariates exhibiting near-linear effects (effective degrees of freedom 1), such as Indigenous population percentage, were modelled *parametrically.* **Illiteracy** was excluded from the final model due to a **lack of statistical significance**.

To explore temporal variation in covariate and spatial effects, we fitted a second model using `by = Year` interaction smooths. This approach fits a separate smooth function for each covariate within each year, allowing effect shapes to vary across time.

$$\log(\mu_i) = \log(\text{Population}_i) + \beta_0 + \sum_{y \in \{2012,2013,2014\}} \left(\sum_{j=1}^{8} f_{j,y}(X_{ji}) + f_{\text{spatial},y}(\text{lon}_i, \text{lat}_i) \right)$$

Here, $f_{j,y}(X_{ji})$ represents the smooth effect of covariate j in year y, and $f_{\text{spatial},y}(\cdot)$ captures year-specific spatial risk structure using thin-plate splines. This structure enables analysis of evolving socio-economic and spatial patterns, though the added complexity yielded only marginal gains in model fit.

Four model variants were compared to evaluate contributions of spatial and temporal components:

- **Model 1**: Socio-economic covariates only (Negative Binomial + tensor product smooth)
- **Model 2**: Model 1 + spatial smooth (Negative Binomial)
- **Model 3**: Model 2 + Year smooth (Tweedie)
- **Model 4**: Model 3 + spatio-temporal interaction (Tweedie)

Models were evaluated using *AIC*, *adjusted* R^2, and *deviance explained*, see **Table 2**. Model 3 offered the best balance of fit and interpretability, achieving an adjusted R^2 of 0.902 and explaining 59.7% of deviance. The tensor product smooth in Model 1 increased model complexity and interpretability challenges, with limited improvement in fit, suggesting possible overfitting or over-smoothing of local effects.

Smoothness selection was performed using *REML*, which reduces overfitting relative to GCV (Wood, 2017). The `select = TRUE` option applied additional penalties, allowing exclusion of non-contributing smooths. **Figure 1** shows the QQ plot and residuals. Diagnostics showed good model fit: residuals displayed no clear patterns, QQ-plots showed acceptable distributional fit, and observed vs. fitted, see **Figure 2**, show that TB rates and predictions aligned closely, proving the models accuracy.

4 Results & Interpretation

4.1 Socio-Economic Risk Factors

Figure 3 presents the estimated smooth terms for socio-economic covariates, capturing their non-linear relationships with TB risk.

Indigenous population percentage, modelled linearly (EDF 1), showed a strong positive association with TB incidence, indicating persistent vulnerability. **Poor sanitation** had a clear protective effect, with risk decreasing steadily as coverage improved.

Poverty, **unemployment**, and **notification timeliness** also exhibited positive associations with TB rates. The effect of timeliness—where longer reporting delays increased risk—suggests diagnostic inefficiencies may contribute to transmission in under-resourced areas.

Urbanisation and **density** had more complex, non-linear patterns. Urbanisation had little effect until high levels (>80%), where risk increased sharply, likely reflecting overcrowding in informal settlements. Density increased TB risk up to a threshold, then declined—possibly due to better healthcare access in highly urbanised areas.

Overall, poverty, sanitation, and timeliness emerged as the most influential predictors, showing strong, consistent effects across the sample. As seen in the Model summary, all covariates displayed significance. These findings highlight the multidimensional nature of TB risk and the need to address both socio-economic deprivation and infrastructure deficiencies to reduce disease burden.

4.2 Spatial Risk Structure

Figure 4 displays predicted TB incidence across Brazil's microregions, based on the fitted GAM with socio-economic and spatial smooth terms. High-risk clusters persist in the north and northwest (e.g. Manaus, Cuiabá) and southeast (e.g. Santos).

These elevated risks remained after accounting for covariates, suggesting unobserved influences such as healthcare access, environmental exposure, or regional mobility. The spatial smooth effectively captured this residual heterogeneity, producing risk maps useful for geographically targeted intervention.

These clusters represent regions where TB burden is not fully explained by measured socio-economic variables. Prioritising surveillance and public health strategies in these areas—particularly those with moderate covariate profiles but high predicted risk—may help identify and address hidden structural drivers of transmission.

4.3 Temporal and Spatio-Temporal Trends

Figure 5 shows smooth terms for selected socio-economic covariates with by = Year interactions, allowing their effects on TB risk to vary between 2012 and 2014. Although some changes were observed—such as modest increases in the influence of poverty, unemployment, and Indigenous population share—overall variation was limited. Interestingly, the model's summary output shows that `poverty:Year2013` and `poverty:Year2014` (p-values of **0.019,0.001**), has lower significance than in `poverty:Year2012`(**0.0001**). This is reflected in the flatter smooth terms.

This likely reflects the short three-year study window and the annual data resolution, which may obscure more rapid or localised shifts. Smooth terms also impose regularity, potentially masking

abrupt transitions. In contrast, the effects of sanitation and density slightly weakened over time, possibly reflecting gradual urban improvements.

Figure 6 illustrates predicted annual TB Risk across regions. While the spatial pattern remained broadly stable, several interior regions showed slight increases in predicted risk, and some coastal areas declined. These patterns suggest a need for continued monitoring to detect emerging hotspots. Finer-grained temporal data—such as monthly counts or a longer time series—would improve sensitivity to evolving epidemiological dynamics.

4.4 Policy Implications and High-Risk Regions

Table 3 Model-adjusted predictions identified around **40–45 microregions** in the top quintile of TB incidence, exceeding **40 cases per 100,000**. High-risk areas were consistently located near **Manaus**, **Cuiabá**, and **Santos**, where disease burden remained elevated across all years.

These regions should be prioritised for intervention—especially where high predicted risk overlaps with **poverty**, **poor sanitation**, and **large Indigenous populations**. Targeted investment in **community-based screening**, **diagnostic access**, and **basic infrastructure** can address the structural vulnerabilities most strongly associated with TB.

Although risk patterns were stable over the study period, continued surveillance is essential. These results provide an immediate, evidence-based foundation for geographically targeted action that tackles both clinical and socio-economic drivers of TB. Prioritising these regions not only supports disease control but also advances health equity by addressing the structural inequalities that underlie TB risk.

5 Conclusion

This study used Generalised Additive Models with a Tweedie framework to estimate spatio-temporal TB risk across Brazil's 557 microregions. The model identified consistent associations between TB incidence and socio-economic vulnerability—especially in regions with overlapping deprivation, poor sanitation, and high Indigenous populations. Spatial risk maps revealed persistent high-burden clusters, supporting **geographically targeted policy recommendations**.

While the model performed robustly, limitations remain. The short study window restricts long-term inference, and the ecological design limits individual-level conclusions. Unmeasured factors such as healthcare access or internal migration may also contribute to residual spatial risk. The additive GAM structure assumes smooth, additive effects and may underrepresent abrupt changes or covariate interactions.

Despite these constraints, the analysis offers a **practical framework for identifying high-risk regions** and informing TB control strategy. Sustainable reductions in burden will depend on both clinical intervention and **investment in the underlying social and structural determinants of disease**.

6 Appendix

This appendix contains all figures, tables, and model outputs referenced in the main report.

Word Count: 1517

6.1 References:

1. Wood, S. N. (2017). *Generalized Additive Models: An Introduction with R (2nd ed.)*. CRC Press.
2. Murase et al. (2009) *Application of a generalized additive model (GAM) to reveal relationships between environmental factors and distributions of pelagic fish and krill: a case study in Sendai Bay, Japan*
3. Tavares et al. (2024) *Unsuccessful tuberculosis treatment outcomes across Brazil's geographical landscape before and during the COVID-19 pandemic: Are we truly advancing toward the sustainable development/end TB goal? Infectious Diseases of Poverty, 13(1), 17.*
4. Wood, S. N. (n.d.). Tweedie: *Tweedie exponential family models* [R documentation]. mgcv package, R Project for Statistical Computing. Retrieved March 22, 2025

6.2 Table 1 - Summary Statistics Table

Summary statistics of socio-economic covariates and TB Rate per 100,000 population.

```
# [Table A.1] Summary stats of covariates + TB Rate

TBdata$TB_Rate <- (TBdata$TB/TBdata$Population)*100000

# List of variables and their brief descriptions
covariate_labels <- c(
  "Indigenous",
  "Illiteracy",
  "Urbanisation",
  "Density",
  "Poverty",
  "Poor Sanitation",
  "Unemployment",
  "Timeliness",
  "TB Rate"
)

# Variables for summarisation
variables <- c("Indigenous", "Illiteracy", "Urbanisation", "Density",
               "Poverty", "Poor_Sanitation", "Unemployment", "Timeliness",
               "TB_Rate")

# Summarising with explicit naming
summary_stats <- TBdata %>%
  summarise(across(all_of(variables), list(
    Mean = ~mean(.x, na.rm = TRUE),
    SD = ~sd(.x, na.rm = TRUE),
    Min = ~min(.x, na.rm = TRUE),
    Q1 = ~quantile(.x, 0.25, na.rm = TRUE),
    Median = ~median(.x, na.rm = TRUE),
    Q3 = ~quantile(.x, 0.75, na.rm = TRUE),
    Max = ~max(.x, na.rm = TRUE)
  ), .names = "{.col}__{.fn}")) %>%
  pivot_longer(cols = everything(),
               names_to = c("Variable", "Statistic"),
               names_sep = "__",
               values_to = "Value") %>%
  pivot_wider(names_from = Statistic, values_from = Value) %>%
  mutate(Variable = covariate_labels)

# Create the professional GT table
summary_stats %>%
  gt() %>%
  tab_header(
```

```
    title = md("**Summary Statistics of Socio-Economic Covariates and
                TB Rate**")
  ) %>%
  fmt_number(columns = where(is.numeric), decimals = 2) %>%
  cols_label(
    Variable = "Covariate",
    Mean = "Mean",
    SD = "Std Dev",
    Min = "Min",
    Q1 = "25th Percentile",
    Median = "Median",
    Q3 = "75th Percentile",
    Max = "Max"
  ) %>%
  tab_style(
    style = cell_text(weight = "bold"),
    locations = cells_column_labels(everything())
  ) %>%
  tab_style(
    style = list(cell_fill(color = "gray95")),
    locations = cells_body(rows = Variable == "TB Rate (Cases per 100,000
                           population)")
  ) %>%
  tab_options(
    table.font.names = "Times New Roman",
    table.font.size = 13,  # Slightly larger text for better readability
    heading.align = "center",
    column_labels.font.size = 13,  # Increase column header size
    column_labels.font.weight = "bold",
    row.striping.include_table_body = TRUE,
    data_row.padding = px(10),  # Increase row padding for more height
    column_labels.border.top.width = px(3),
    column_labels.border.bottom.width = px(3),
    table.border.top.width = px(2),
    table.border.bottom.width = px(2),
    table.border.bottom.color = "black",
    table.border.top.color = "black",
    table.width = pct(100)
  ) %>%
tab_caption("Table 1: Summary Statistics of Socio-Economic Covariates
             and TB Rate")
```

Table 1: Table 1: Summary Statistics of Socio-Economic Covariates and TB Rate

Summary Statistics of Socio-Economic Covariates and TB Rate

Covariate	**Mean**	Std Dev	**Min**	25th Percentile	**Median**	75th Percentile	**Max**
Indigenous	0.84	3.53	0.01	0.06	0.11	0.24	50.65
Illiteracy	14.80	9.30	2.34	6.68	11.52	22.84	41.14
Urbanisation	71.96	16.53	22.34	58.45	72.66	86.16	99.93
Density	0.62	0.17	0.42	0.52	0.58	0.66	1.68
Poverty	44.37	19.37	5.92	26.23	42.60	63.91	77.88
Poor Sanitation	16.45	12.51	0.05	6.39	13.91	25.00	58.43
Unemployment	6.93	2.56	1.13	5.15	6.78	8.40	20.44
Timeliness	47.67	21.47	0.00	31.29	48.36	62.58	96.69
TB Rate	23.54	15.42	0.00	13.52	20.11	28.94	117.73

6.2.1 Covariate vs TB Rate Scatterplots

Not included but was used for EDA. Faceted scatterplots showing the relationship between each socio-economic covariate and the TB Rate.

```
# This isnt included as a plot, but was very informative in displaying the
# distribution of data for each covariate, and the mean line

# Convert data to long format for faceted plotting
TBdata_long <- TBdata %>%
  pivot_longer(cols = c(Indigenous, Illiteracy, Urbanisation, Density,
                        Poverty, Poor_Sanitation, Unemployment, Timeliness),
               names_to = "Covariate",
               values_to = "Value")

# Facet labels
facet_labels <- c(
  Indigenous = "Indigenous (%)",
  Illiteracy = "Illiteracy (%)",
  Urbanisation = "Urbanisation (%)",
  Density = "Dwelling Density (Dwellers/Room)",
  Poverty = "Poverty (%)",
  Poor_Sanitation = "Poor Sanitation (Higher = Worse)",
  Unemployment = "Unemployment (%)",
  Timeliness = "TB Notification Delay (Days)"
)

# Faceted Plot
ggplot(TBdata_long, aes(x = Value, y = TB_Rate)) +
  geom_point(alpha = 0.4, color = "darkblue") +  # Scatterplot points
  geom_smooth(method = "loess", color = "red", se = FALSE) +
  # Smoothed trend line
  facet_wrap(~ Covariate, scales = "free_x",
             labeller = labeller(Covariate = facet_labels),
```

```
            nrow = 4, ncol = 2) +
labs(title = "Relationship Between Socio-Economic Covariates and TB Rate",
     x = "Covariate Value",
     y = "TB Rate per 100,000") +
custom_theme+
theme(plot.title = element_text(size = 20, face = "bold", hjust = 0.5),
      strip.text = element_text(size = 16, face = "bold")
)
```

6.3 Initial Model Checks

- Check for overdispersion
- First GAM specification
- Use of tensor smooths and exploratory edf

```
# Initial Model Fitting (GAMs)

# Checking for overdispersion:
observed_var <- var(TBdata$TB)
expected_var <- mean(TBdata$TB)
dispersion <- observed_var / expected_var
print(dispersion)
```

```
## [1] 2223.535
```

```
# Ensure categorical variables are factors
TBdata$Year <- factor(TBdata$Year)
TBdata$Region <- factor(TBdata$Region)

# Defining Offset
TBdata$logPop <- log(TBdata$Population)
```

```
library(mgcv)
set.seed(123)
# Fit the Negative Binomial GAM with population offset. Base Model

# First GAM model
gam_model_initial <- gam(TB ~
                    s(Indigenous, bs = "tp", k=10) +
                    s(Illiteracy, bs = "tp", k=10) +
                    s(Urbanisation, bs = "tp", k=10) +
                    s(Density, bs = "tp", k=10) +
                    s(Poverty, bs = "cr", k=5) +
                    s(Unemployment, bs = "cr", k=5) +
                    s(Poor_Sanitation, k = 10, bs = "cr") +
                    s(Timeliness, bs = "tp", k=10) +
                    te(lon, lat, by = Year, bs = c("tp", "cr"),
                       k = c(20, 20, 5)) +
                    offset(logPop),
                  data = TBdata,
                  family = nb(link = "log"),
                  method = "REML")

# Used a mix of "cr" and "tp" splines. CR on the covariates with
# mostly linear trends, as this smooth function is better on them. (Wood 2017)

# Tested the difference in model accuracy using the different splines.

# In the initial model building, k.check was pivotal in understanding the right
```

```r
# value for k.


par(mfrow = c(2,2))
gam.check(gam_model_initial)

summary(gam_model_initial)

# Removing Illiteracy from future models as it inst a significant predictor.
# Indigenous edf is around 1, meaning the covariate has a linear trend.
# Thus a smooth function isn't needed for this.
```

6.3.1 Second Model - Negative Binomial

Summary and results for final GAM with NB family. Includes choice of smooths and rationale.

Year included as a factor, as 3 years of data isnt enough to apply a useful smooth function.

```r
# Final Negative Binomial GAM with linear/smooth terms

set.seed(123)
gam_model <- gam(TB ~
                   Year +
                   Indigenous +
                   s(Urbanisation) +
                   s(Density) +
                   s(Unemployment) +
                   s(Poor_Sanitation) +
                   s(Poverty) +
                   s(Timeliness) +
                   s(lon,lat, k = 30) +
                 offset(logPop),
                 data = TBdata,
                 family = nb(link = "log"),
                 method = "REML",
                 select=TRUE)
# **Select = TRUE** Reference to Wood 2017 page 406, this is used as "reduced tendency to unde

# I found this improved model accuacy without overfitting, so i retained this
# feature going forward

plot(gam_model, pages = 1, all.terms = TRUE, shade = TRUE)
summary(gam_model)
k.check(gam_model)

par(mfrow = c(2,2))
gam.check(gam_model)
```

6.4 Final Model - Tweedie Family

Includes model diagnostics and fitted predictions.

```
# Final Tweedie GAM model
set.seed(123)
gam_model_tw <- gam(TB ~
                    Year +
                    Indigenous +
                    s(Urbanisation) +
                    s(Density) +
                    s(Unemployment) +
                    s(Poor_Sanitation) +
                    s(Poverty) +
                    s(Timeliness) +
                    s(lon,lat) +
                    offset(logPop),
                  data = TBdata,
                  family = tw,
                  method = "REML",
                  select=TRUE)

# Learnt this model from Wood 2017, and applied it. Significantly better fit.

k.check(gam_model_tw)
```

```
##                     k'       edf   k-index p-value
## s(Urbanisation)      9  6.191294 0.5246709       0
## s(Density)           9  3.096894 0.5203579       0
## s(Unemployment)      9  4.661098 0.5275125       0
## s(Poor_Sanitation)   9  5.516808 0.5181694       0
## s(Poverty)           9  4.615654 0.5288610       0
## s(Timeliness)        9  3.805945 0.5813806       0
## s(lon,lat)          29 24.694337 0.4894340       0
```

```
summary(gam_model_tw)
```

```
##
## Family: Tweedie(p=1.666)
## Link function: log
##
## Formula:
## TB ~ Year + Indigenous + s(Urbanisation) + s(Density) + s(Unemployment) +
##     s(Poor_Sanitation) + s(Poverty) + s(Timeliness) + s(lon,
##     lat) + offset(logPop)
##
## Parametric coefficients:
##              Estimate Std. Error  t value Pr(>|t|)
## (Intercept) -8.439363   0.017281 -488.367  < 2e-16 ***
## Year2013    -0.000324   0.023324   -0.014 0.988918
```

```
## Year2014     -0.040093    0.023385   -1.714 0.086631 .
## Indigenous    0.015384    0.004297    3.580 0.000354 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                        edf Ref.df      F  p-value
## s(Urbanisation)      6.191      9  7.182  < 2e-16 ***
## s(Density)           3.097      9  3.364  < 2e-16 ***
## s(Unemployment)      4.661      9  9.638  < 2e-16 ***
## s(Poor_Sanitation)   5.517      9  6.485  < 2e-16 ***
## s(Poverty)           4.616      9  3.768 4.15e-07 ***
## s(Timeliness)        3.806      9  6.304  < 2e-16 ***
## s(lon,lat)          24.694     29 18.592  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.902   Deviance explained = 59.7%
## -REML = 7122.6  Scale est. = 0.57508   n = 1671
```

```r
# plot(gam_model_tw, pages = 1, all.terms = TRUE, shade = TRUE)

# Generate predicted TB cases from the GAM
TBdata$predicted_cases <- predict(gam_model_tw, type = "response")

# Convert to TB rate per 100,000 population
TBdata$predicted_rate <- (TBdata$predicted_cases / TBdata$Population) * 100000
```

6.5 GAM with Temporal Smooths

Model incorporating smooths by year (`s(..., by = Year)`) to examine changing effects.

```r
# GAM with Temporal/Spatial

gam_model_SP <- gam(TB ~ Year +
                      s(Indigenous, by = Year) +
                      s(Illiteracy, by = Year) +
                      s(Urbanisation, by = Year) +
                      s(Density, by = Year) +
                      s(Unemployment, by = Year) +
                      s(Poor_Sanitation, by = Year) +
                      s(Poverty, by = Year) +
                      s(Timeliness, by = Year) +
                      s(lon,lat, by = Year) +
                      offset(logPop),
                    data = TBdata,
                    family = tw,
                    method = "REML",
                    select=TRUE)
```

```r
summary(gam_model_SP)
```

```
##
## Family: Tweedie(p=1.688)
## Link function: log
##
## Formula:
## TB ~ Year + s(Indigenous, by = Year) + s(Illiteracy, by = Year) +
##     s(Urbanisation, by = Year) + s(Density, by = Year) + s(Unemployment,
##     by = Year) + s(Poor_Sanitation, by = Year) + s(Poverty, by = Year) +
##     s(Timeliness, by = Year) + s(lon, lat, by = Year) + offset(logPop)
##
## Parametric coefficients:
##              Estimate Std. Error  t value Pr(>|t|)
## (Intercept) -8.420859   0.018061 -466.243   <2e-16 ***
## Year2013     0.003441   0.025461    0.135    0.893
## Year2014    -0.040094   0.025534   -1.570    0.117
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                                   edf Ref.df     F  p-value
## s(Indigenous):Year2012      8.665e-01      9 0.710 0.003695 **
## s(Indigenous):Year2013      6.494e-01      9 0.205 0.074809 .
## s(Indigenous):Year2014      8.364e-01      9 0.561 0.009512 **
## s(Illiteracy):Year2012      4.386e-04      9 0.000 0.493095
## s(Illiteracy):Year2013      8.945e-04      9 0.000 0.500691
## s(Illiteracy):Year2014      9.963e-04      9 0.000 0.826420
## s(Urbanisation):Year2012    2.449e+00      9 1.728 9.35e-05 ***
## s(Urbanisation):Year2013    2.879e+00      9 2.150 1.67e-05 ***
## s(Urbanisation):Year2014    2.588e+00      9 1.566 0.000266 ***
## s(Density):Year2012         2.276e+00      9 1.182 0.001226 **
## s(Density):Year2013         2.130e+00      9 1.144 0.001017 **
## s(Density):Year2014         2.017e+00      9 0.937 0.003667 **
## s(Unemployment):Year2012    1.774e+00      9 3.162  < 2e-16 ***
## s(Unemployment):Year2013    2.202e+00      9 4.489  < 2e-16 ***
## s(Unemployment):Year2014    2.223e+00      9 4.287  < 2e-16 ***
## s(Poor_Sanitation):Year2012 3.981e+00      9 2.137 8.32e-05 ***
## s(Poor_Sanitation):Year2013 3.121e+00      9 1.447 0.001039 **
## s(Poor_Sanitation):Year2014 3.943e+00      9 2.642 6.71e-06 ***
## s(Poverty):Year2012         1.323e+00      9 1.328 0.000128 ***
## s(Poverty):Year2013         1.144e+00      9 0.447 0.019878 *
## s(Poverty):Year2014         8.845e-01      9 0.846 0.001451 **
## s(Timeliness):Year2012      1.535e+00      9 2.894 6.50e-07 ***
## s(Timeliness):Year2013      9.328e-01      9 1.530 8.47e-05 ***
## s(Timeliness):Year2014      9.543e-01      9 2.257 2.91e-06 ***
## s(lon,lat):Year2012         1.905e+01     29 4.983  < 2e-16 ***
```

```
## s(lon,lat):Year2013         1.993e+01     29 5.589  < 2e-16 ***
## s(lon,lat):Year2014         1.991e+01     29 5.778  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.871   Deviance explained = 57.3%
## -REML =   7225  Scale est. = 0.57076   n = 1671
```

```
# plot(gam_model_SP, pages = 1, all.terms = TRUE, shade = TRUE)
```

```
# AIC tests
AIC(gam_model)
AIC(gam_model_tw)
AIC(gam_model_SP)
```

6.6 Table 2 - Model Comparison Table

Comparison between NB, Tweedie, and temporal GAMs using AIC and deviance explained.

```
# GT table comparing NB, Tweedie, and spatio-temporal GAMs
library(gt)
library(tidyverse)

# Data preparation
model_comparison <- data.frame(
  Model = c("Tweedie", "Negative Binomial", "Tweedie (Temporal Terms)"),
  Family = c("Tweedie(p=1.666)", "NB( =8.336)", "Tweedie(p=1.688)"),
  Adj_R2 = c(0.902, 0.846, 0.871),
  Deviance_Explained = c("59.7%", "56.6%", "57.3%"),
  REML = c(7123, 7068, 7225),
  Scale_Estimate = c(0.575, 1.00, 0.570),
  AIC = c(14120, 14023, 14325)
)

model_comparison %>%
  gt() %>%
  tab_header(
    title = md("**Comparison of GAM Models**"),
    subtitle = "Tweedie vs Negative Binomial vs Tweedie with Temporal Terms"
  ) %>%
  cols_label(
    Adj_R2 = md("Adjusted R²")
  ) %>%
  fmt_number(columns = c(Adj_R2, Scale_Estimate), decimals = 3) %>%
  tab_style(
    style = cell_text(weight = "bold"),
    locations = cells_column_labels(everything())
  ) %>%
  tab_footnote(
    footnote = "Lower AIC values indicate better model fit.",
    locations = cells_column_labels(columns = AIC)
  ) %>%
  tab_options(
    table.font.names = "Times New Roman",
    table.font.size = 11,
    heading.align = "center",
    column_labels.font.size = 11,
    column_labels.font.weight = "bold",
    data_row.padding = px(6),
    table.width = pct(100),
    column_labels.border.top.width = px(2),
    column_labels.border.bottom.width = px(2),
    table.border.top.width = px(1),
```

Table 2: Table 2: Comparison of GAM Models

Comparison of GAM Models

Tweedie vs Negative Binomial vs Tweedie with Temporal Terms

Model	**Family**	Adjusted R^2	Deviance_Explained	**REML**	Scale_Estimate	AIC[1]
Tweedie	Tweedie(p=1.666)	0.902	59.7%	7123	0.575	14120
Negative Binomial	NB(=8.336)	0.846	56.6%	7068	1.000	14023
Tweedie (Temporal Terms)	Tweedie(p=1.688)	0.871	57.3%	7225	0.570	14325

[1]Lower AIC values indicate better model fit.

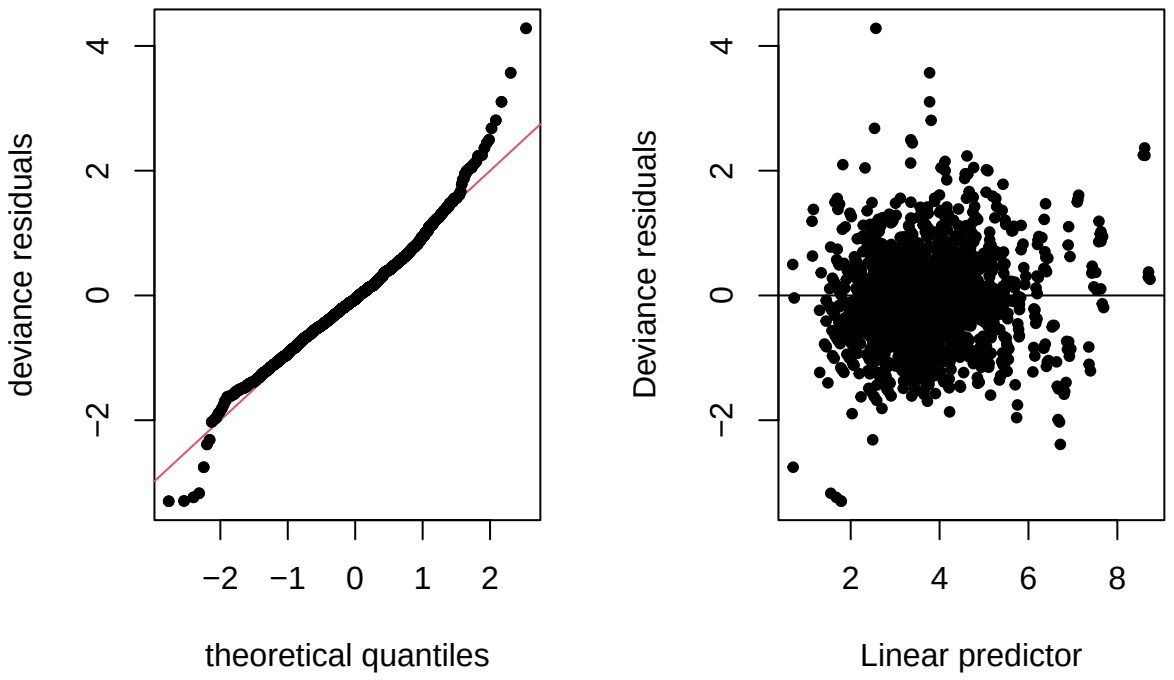
```
    table.border.bottom.width = px(1),
    table.border.bottom.color = "black",
    table.border.top.color = "black"
  ) %>%
  tab_caption("Table 2: Comparison of GAM Models")
```

6.7 Figure 1 - Diagnostic Plots

QQ plot and deviance residuals from the final Tweedie model.

Only presenting QQ plot of TW model. It showed a better fit an accuracy over the other models.

```
# Check residuals
par(mfrow=c(1,2))
# QQplot
qq.gam(gam_model_tw,pch=20)
# deviance residuals vs linear predictor
xx <-gam_model_tw$linear.predictors
yy <-residuals(gam_model_tw,type="deviance")

plot(xx,yy,pch=20,xlab="Linear predictor",ylab="Deviance residuals")
abline(h=0)
```



6.8 Figure 2 - Observed vs Fitted Plot

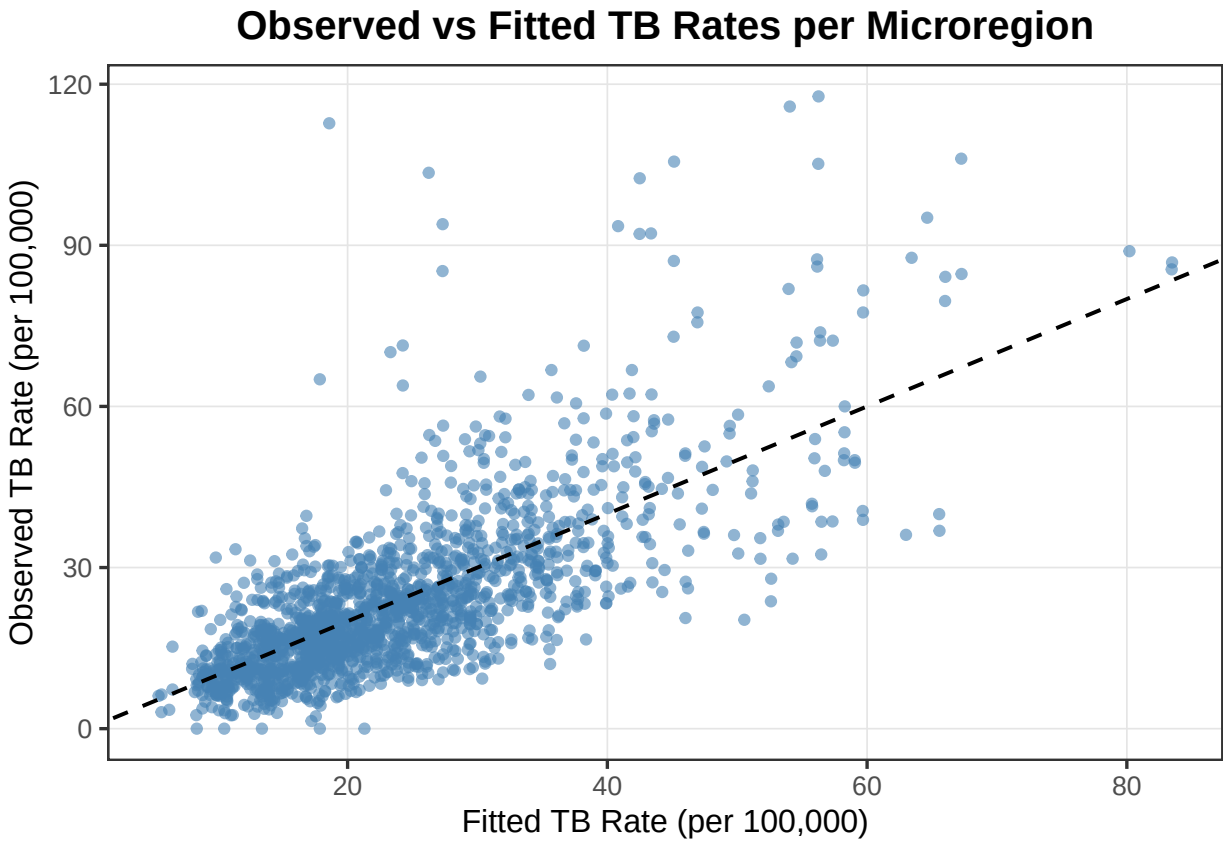
Scatterplot comparing observed and fitted TB Rates. This indicates the accuracy of the model.

```
obs_fit_plot <- ggplot(TBdata, aes(x = predicted_rate, y = TB_Rate)) +
  geom_point(alpha = 0.6, colour = "steelblue") +
  geom_abline(slope = 1, intercept = 0, linetype = "dashed", colour = "black",
              linewidth = 0.7) +
  labs(
    title = "Observed vs Fitted TB Rates per Microregion",
    x = "Fitted TB Rate (per 100,000)",
    y = "Observed TB Rate (per 100,000)"
  ) +
  theme_bw(base_size = 13) +
  theme(
    plot.title = element_text(face = "bold", size = 15, hjust = 0.5),
    axis.title = element_text(size = 12),
    axis.text = element_text(size = 10),
    panel.grid.major = element_line(size = 0.3, colour = "grey90"),
    panel.grid.minor = element_blank()
  )
```

```
## Warning: The `size` argument of `element_line()` is deprecated as of ggplot2 3.4.0.
## i Please use the `linewidth` argument instead.
```

```
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```
obs_fit_plot
```



6.9 Figure 3 - Smooth Effects of Covariates

Estimated smooth functions with 95% CIs for each covariate (Tweedie model).

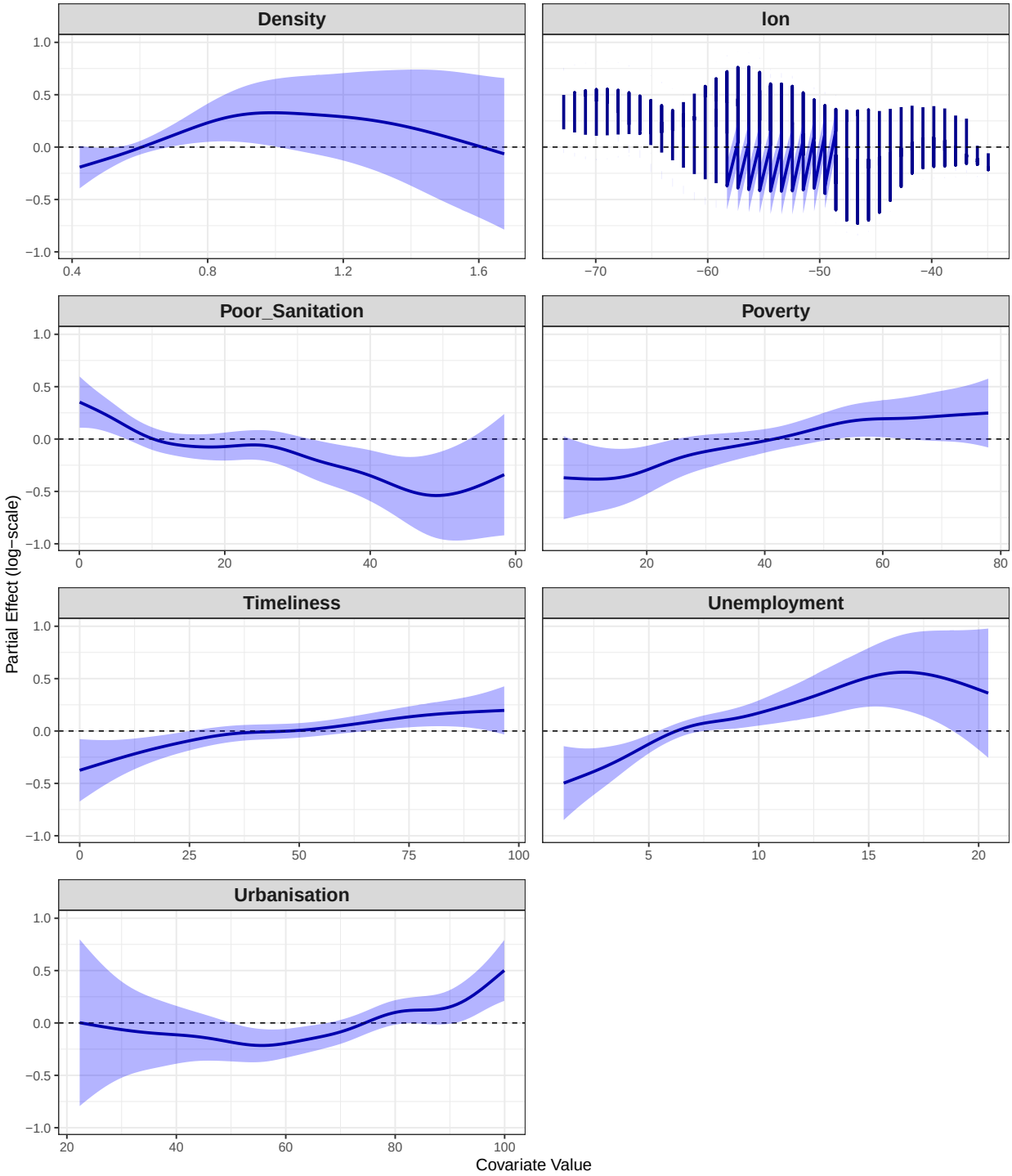
```
# Smooth terms from Tweedie GAM

# Extract smooth terms from the GAM model (gam_model_tw)
smooth_terms <- plot(gam_model_tw, pages = 1, all.terms = TRUE, shade = TRUE,
                     seWithMean = TRUE)

# Data frame to store all covariates smooth predictions
smooth_plot_data <- map_dfr(1:length(smooth_terms), function(i) {
  data.frame(
    Covariate = smooth_terms[[i]]$xlab,
    x = smooth_terms[[i]]$x,
    fit = smooth_terms[[i]]$fit,
    se = smooth_terms[[i]]$se
  )
})

# Compute confidence intervals
smooth_plot_data <- smooth_plot_data %>%
  mutate(lower_CI = fit - 1.96 * se,
         upper_CI = fit + 1.96 * se)

# Define clear and informative facet labels
facet_labels <- c(
  "s(Urbanisation)" = "Urbanisation (%)",
  "s(Density)" = "Dwelling Density (Dwellers/Room)",
  "s(Unemployment)" = "Unemployment (%)",
  "s(Poor_Sanitation)" = "Poor Sanitation (Higher = Worse)",
  "s(Poverty)" = "Poverty (%)",
  "s(Timeliness)" = "TB Notification Delay (Days)",
  "s(lon,lat)" = "Spatial Effect (Longitude & Latitude)"
)

# Filter out the 2-dimensional smooth (lon,lat), which needs separate treatment
smooth_plot_data_filtered <- smooth_plot_data %>%
  filter(Covariate != "s(lon,lat)")
```

```
# Plot using ggplot
ggplot(smooth_plot_data_filtered, aes(x = x, y = fit)) +
  geom_line(color = "darkblue", linewidth = 1.2) +
  geom_hline(yintercept = 0, linetype = "dashed", color = "black") +
  geom_ribbon(aes(ymin = lower_CI, ymax = upper_CI), alpha = 0.3,
              fill = "blue") +
  facet_wrap(~ Covariate, scales = "free_x", labeller = labeller
             (Covariate = facet_labels), nrow = 4, ncol = 2) +
  labs(title = "Figure 3:Partial Effects of Socio-Economic Covariates on TB Risk",
```

```
      subtitle = "Estimated smooth terms with 95% confidence intervals",
      x = "Covariate Value",
      y = "Partial Effect (log-scale)",
      caption = "Positive values indicate increased risk; negative values
      indicate decreased risk.") +
theme_bw(base_size = 14) +
theme(
  plot.title = element_text(size = 20, face = "bold", hjust = 0.5),
  plot.subtitle = element_text(size = 14, hjust = 0.5),
  strip.text = element_text(size = 16, face = "bold"),
  axis.title = element_text(size = 14),
  panel.spacing = unit(1, "lines")
)
```

Figure 3:Partial Effects of Socio−Economic Covariates on TB Risk

Estimated smooth terms with 95% confidence intervals



Positive values indicate increased risk; negative values indicate decreased risk.

6.10 Figure 4 - Spatial Structure of TB Risk

Map of predicted TB Risk across microregions, using spatial smooths.

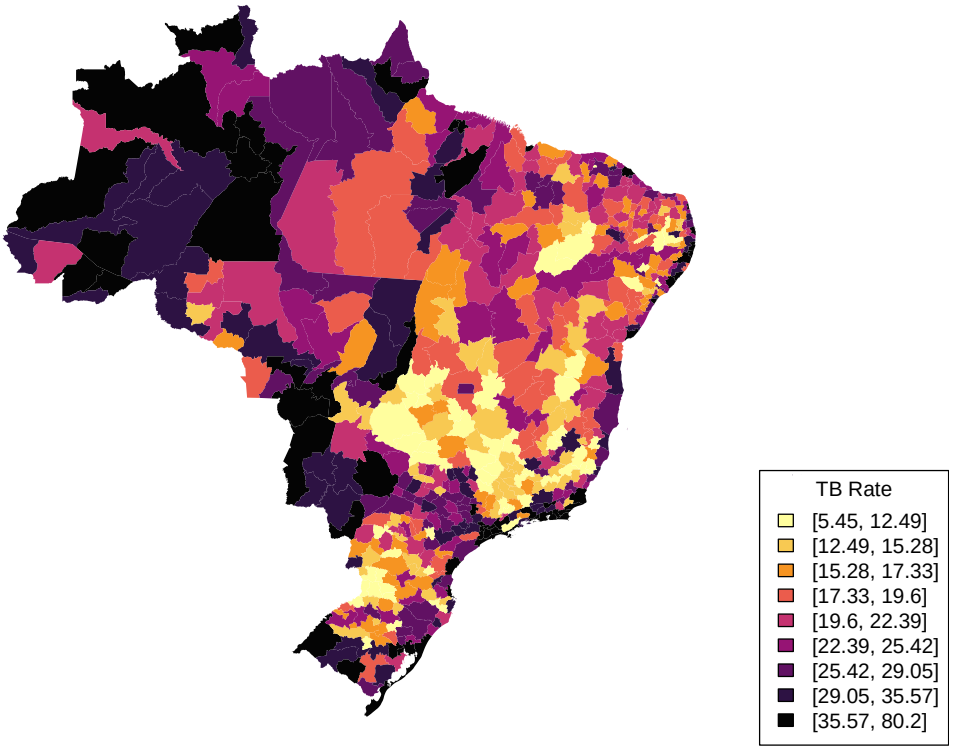
```
# Spatial visualisation of TB risk
library(sp)
```

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

```
library(colorspace)  # for hcl.colors


plot.map <- function(x, Q, main = "", cex = 1, decimals = 2,
                     palette = rev(hcl.colors(length(Q) - 1,
                                              palette = "inferno"))) {

  n.levels <- length(Q) - 1
  cols <- palette
  n <- length(x)

  # Assign colours based on quantile bins
  col <- rep(cols[1], n)
  for (i in 2:n.levels) {
    col[x >= Q[i] & x < Q[i + 1]] <- cols[i]
  }
  col[x >= Q[n.levels + 1]] <- cols[n.levels]

  # Legend labels
  legend.names <- paste0("[", round(Q[-length(Q)], decimals), ", ",
                         round(Q[-1], decimals), "]")

  # Sanity check
  stopifnot(length(col) == length(brasil_micro))

  # Plot
  plot(brasil_micro, col = col, main = main, border = NA)
  legend("bottomright", legend = legend.names, fill = cols, cex = cex,
         title = "TB Rate")
}


# Assign Region ID to brasil_micro
brasil_micro@data$Region <- brasil_micro@data$COD_MICRO

# Subset TB data for 2014
TB_2014 <- TBdata %>% filter(Year == 2014)

# Sort both datasets by Region
TB_2014 <- TB_2014[order(TB_2014$Region), ]
```

```r
brasil_micro <- brasil_micro[order(brasil_micro@data$Region), ]

# Predict using final Tweedie GAM
TB_2014$predicted_cases <- predict(gam_model_tw, newdata = TB_2014,
                                   type = "response")

# Convert to TB rate per 100,000 population
TB_2014$predicted_rate <- (TB_2014$predicted_cases / TB_2014$Population)*100000


# Create quantile bins
quantiles <- quantile(TB_2014$predicted_rate, probs =
                        seq(0, 1, length.out = 10), na.rm = TRUE)

# Create colour palette
my_palette <- rev(hcl.colors(length(quantiles) - 1, palette = "inferno"))

# Plot the map
plot.map(
  x = TB_2014$predicted_rate,
  Q = quantiles,
  main = "Figure 4: Predicted TB Risk (per 100,000)",
  cex = 1,
  decimals = 2,
  palette = my_palette
)
```

Figure 4: Predicted TB Risk (per 100,000)



6.11 Figure 5 - Temporal Changes in Effects

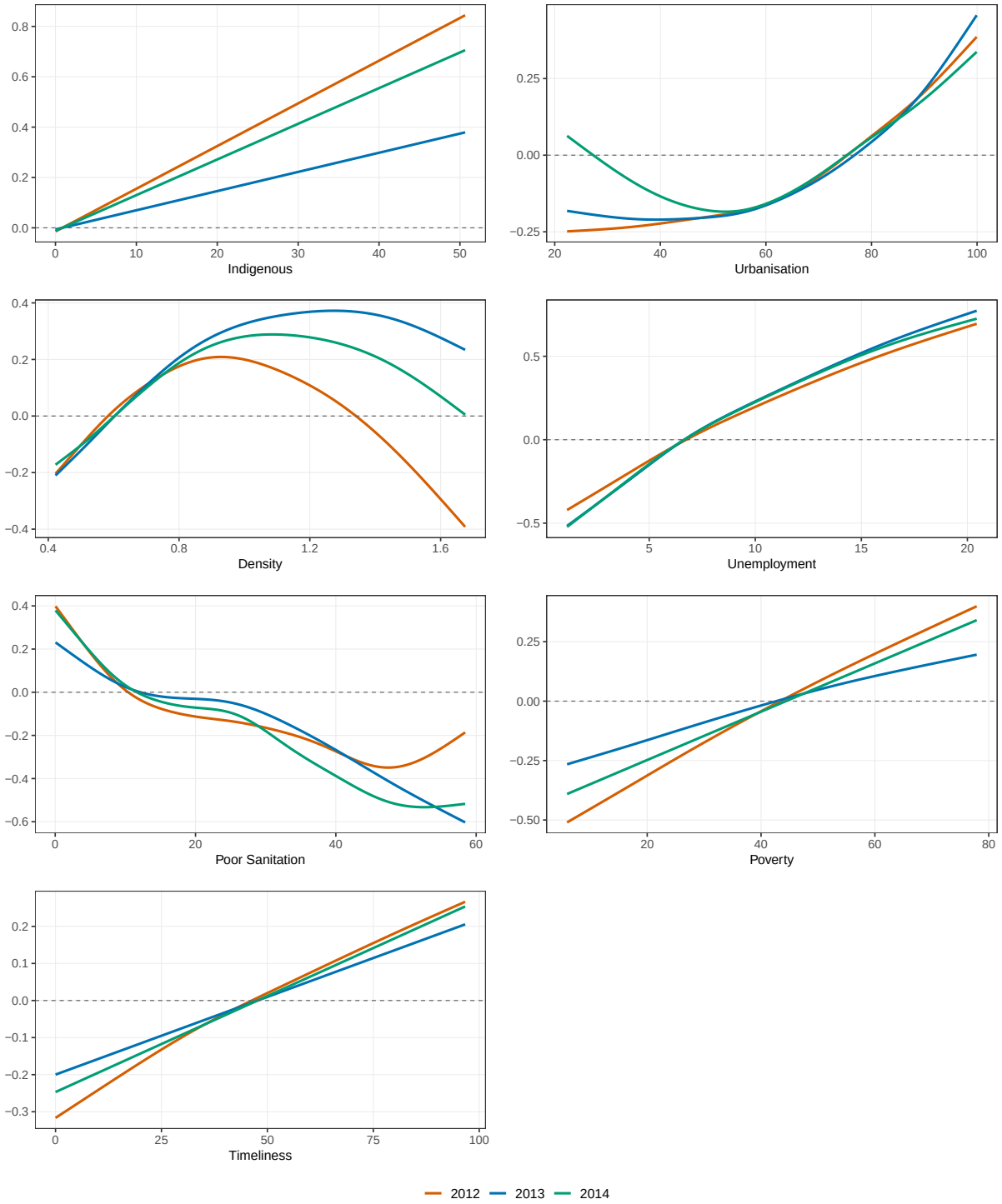
Line plots showing covariate effects for each year

```
# Temporal smooths across years using wrap_plots()


library(patchwork)
library(scales)
library(stringr)

# Colour scheme
year_colours <- c(
  "2012" = "#D55E00",  # burnt orange
  "2013" = "#0072B2",  # deep blue
  "2014" = "#009E73"   # teal green
)

# Covariate labels
label_cleaner <- function(x) {
  x %>% str_replace_all("_", " ") %>% str_to_title()
}

library(gratia)

# Extract smooth estimates from the GAM
smooth_df <- smooth_estimates(gam_model_SP)

# Filter out spatial and illiteracy terms
filtered_df <- smooth_df %>%
  filter(
    !str_detect(.smooth, "Illiteracy"),
    !str_detect(.smooth, "lon")
  ) %>%
  mutate(
    covariate = str_extract(.smooth, "(?<=s\\().+?(?=\\))"),
    # extract covariate name
    year = str_extract(.smooth, "\\d{4}")
    # extract year from label
  )


# Plot list, Using a for loop to so we cal do them all at once.
plot_list <- list()
unique_covs <- unique(filtered_df$covariate)

for (cov in unique_covs) {
  df_sub <- filtered_df %>% filter(covariate == cov)
```

```
  p <- ggplot(df_sub, aes_string(x = cov, y = ".estimate", colour = "year")) +
    geom_line(linewidth = 1) +
    geom_hline(yintercept = 0, linetype = "dashed", colour = "grey40",
               linewidth = 0.4) +
    scale_colour_manual(values = year_colours) +
    labs(
      x = label_cleaner(cov),
      y = NULL
    ) +
    theme_bw(base_size = 12) +
    theme(
      plot.title = element_blank(),
      axis.title.x = element_text(size = 11),
      axis.text = element_text(size = 10),
      panel.grid.major = element_line(size = 0.2),
      panel.grid.minor = element_blank(),
      legend.position = "none"
    )

  plot_list[[cov]] <- p
}


# Plot Layout - Shared Legend
combined_plot <- wrap_plots(plot_list, ncol = 2) +
  plot_layout(guides = "collect") &
  theme(
    legend.position = "bottom",
    legend.title = element_blank(),
    legend.text = element_text(size = 11),
    plot.margin = margin(10, 10, 10, 10)
  )
```

```
combined_plot +
  plot_annotation(
    title = "Figure 5: Temporal Variation in Socio-Economic Covariate
    Effects on TB Risk",
    theme = theme(
      plot.title = element_text(size = 16, face = "bold", hjust = 0.5)
    )
  )
```

Figure 5: Temporal Variation in Socio−Economic Covariate Effects on TB Risk

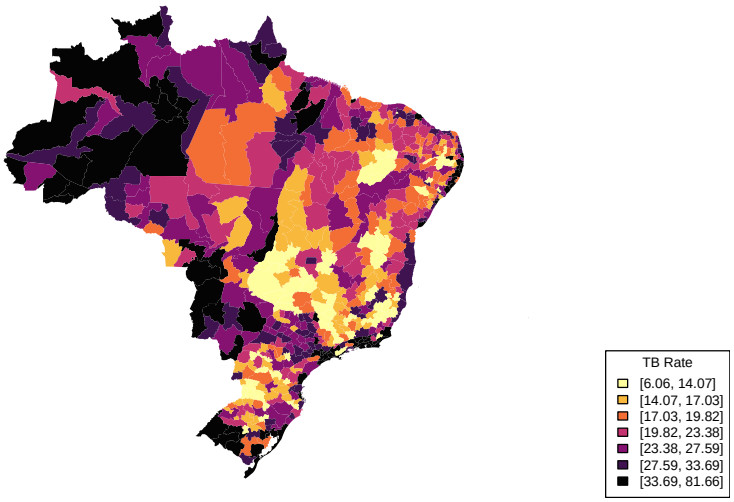


6.12 Figure 6 - Yearly Risk Maps

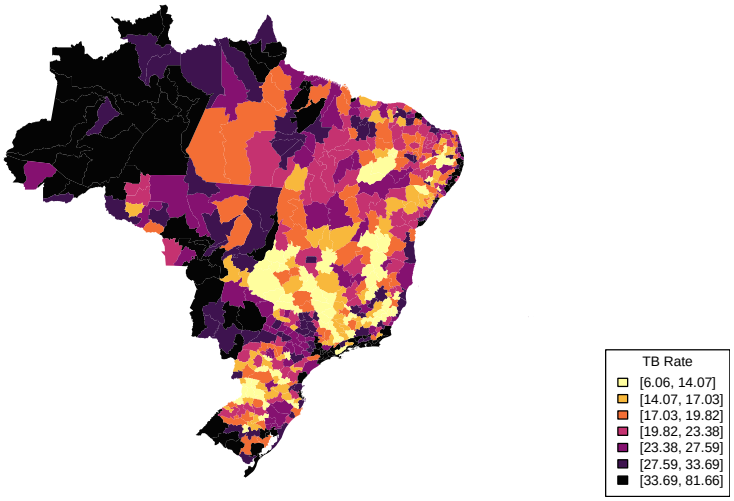
Separate maps for 2012–2014 displaying spatial risk variation.

```r
# Maps for each year
# Predict TB cases and rates using spatio-temporal GAM
TBdata$predicted_cases_sp <- predict(gam_model_SP, type = "response")
TBdata$predicted_rate_sp <- (TBdata$predicted_cases_sp/TBdata$Population)*100000

# Define consistent quantiles across all years
quantiles <- quantile(TBdata$predicted_rate_sp, probs =
                        seq(0, 1, length.out = 8), na.rm = TRUE)
my_palette <- rev(hcl.colors(length(quantiles) - 1, palette = "inferno"))

# Set up plotting layout
par(mfrow = c(3, 1), mar = c(2, 2, 2, 2))  # 3 rows, 1 column

# Loop through each year and plot map
for (year in c("2012", "2013", "2014")) {

  # Filter TBdata for that year
  TB_year <- TBdata %>% filter(Year == year)

  # Sort and align with brasil_micro
  TB_year <- TB_year[order(TB_year$Region), ]
  brasil_micro <- brasil_micro[order(brasil_micro@data$Region), ]


  # Extract predicted rate
  predicted_rate <- TB_year$predicted_rate_sp

  # Plot map for the year
  plot.map(
    x = predicted_rate,
    Q = quantiles,
    main = paste("Predicted TB Risk (per 100,000) - Year", year),
    cex = 1,
    decimals = 2,
    palette = my_palette
  )
}
```

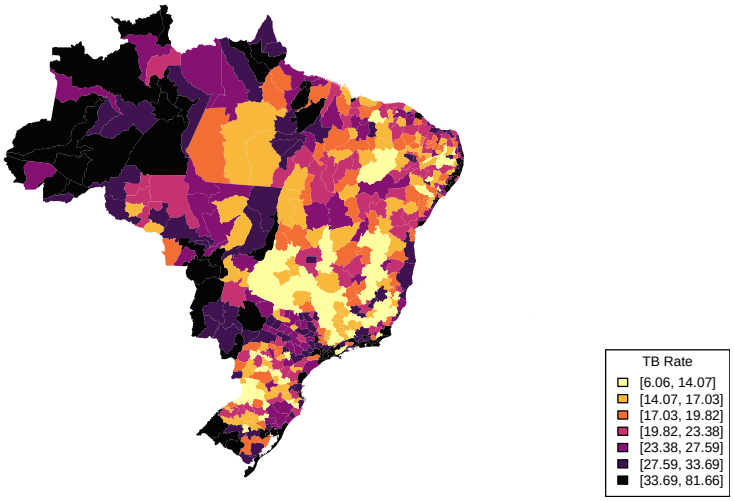
Predicted TB Risk (per 100,000) – Year 2012



Predicted TB Risk (per 100,000) – Year 2013



Predicted TB Risk (per 100,000) – Year 2014



6.12.1 Table 3 - High-Risk Region Table

Regions identified as high-risk using fixed (40 per 100k) and dynamic thresholds.

```
# Table showing high-risk regions by fixed and dynamic thresholds

# Create a lookup table from brasil_micro
region_lookup <- brasil_micro@data %>%
  select(Region = COD_MICRO, Region_Name = NM_MICRO)

TBdata$Region <- as.integer(as.character(TBdata$Region))
region_lookup$Region <- as.integer(as.character(region_lookup$Region))

# Merge region names into TBdata before table creation
TBdata_named <- TBdata %>%
  left_join(region_lookup, by = "Region")

# Create working dataset with predictions
TBdata_new <- TBdata_named %>%
  dplyr::select(Region, Region_Name, Year, Indigenous, Illiteracy, Urbanisation,
                Density, Poverty, Poor_Sanitation, Unemployment, Population,
                Timeliness, lon, lat, logPop) %>%
  mutate(Predicted_TB = predict(gam_model_tw, newdata = ., type = "response"),
         Predicted_TB_Scaled = (Predicted_TB / Population) * 100000)

# Compute dynamic threshold
high_risk_threshold <- TBdata_new %>%
  group_by(Year) %>%
  summarise(Dynamic_Threshold = quantile(Predicted_TB_Scaled, 0.90,
                                         na.rm = TRUE), .groups = "drop")

# Fixed threshold counts
fixed_threshold <- 40
high_risk_regions_fixed <- TBdata_new %>%
  filter(Predicted_TB_Scaled > fixed_threshold) %>%
  group_by(Year) %>%
  summarise(Total_High_Risk_Regions = n(), .groups = "drop")

# Top 3 high-risk regions (by NAME)
top_high_risk_regions <- TBdata_new %>%
  group_by(Year) %>%
  arrange(desc(Predicted_TB_Scaled)) %>%
  slice_head(n = 3) %>%
  summarise(Highest_Risk_Regions = paste(Region_Name, collapse = ", "),
            .groups = "drop")

# Merge all summary components
high_risk_summary <- high_risk_regions_fixed %>%
```

```
  left_join(high_risk_threshold, by = "Year") %>%
  left_join(top_high_risk_regions, by = "Year") %>%
  mutate(Dynamic_Threshold = round(Dynamic_Threshold, 2))



high_risk_summary %>%
  gt() %>%
  tab_header(
    title = md("**High-Risk TB Regions Over Time**"),
    subtitle = "Regions exceeding fixed and dynamic TB rate thresholds"
  ) %>%
  fmt_number(columns = c(Total_High_Risk_Regions, Dynamic_Threshold),
             decimals = 2) %>%
  cols_label(
    Year = "Year",
    Total_High_Risk_Regions = "Regions Above Fixed Threshold (40 per 100k)",
    Dynamic_Threshold = "Top 10% Threshold",
    Highest_Risk_Regions = "Top 3 Highest-Risk Regions"
  ) %>%
  text_transform(
    locations = cells_body(columns = Highest_Risk_Regions),
    fn = function(x) {
      stringr::str_replace_all(x, ", ", ",\n")
    }
  ) %>%
  tab_style(
    style = cell_text(weight = "bold"),
    locations = cells_column_labels(everything())
  ) %>%
  tab_options(
    table.font.size = 11,
    column_labels.font.size = 11,
    column_labels.font.weight = "bold",
    data_row.padding = px(6),
    table.width = pct(95),
    heading.align = "center",
    row.striping.include_table_body = TRUE,
    table.border.top.width = px(1),
    table.border.bottom.width = px(1),
    table.border.bottom.color = "black",
    table.border.top.color = "black"
  ) %>%
  tab_caption("Summary of High-Risk TB Regions Over Time")
```

Table 3: Summary of High-Risk TB Regions Over Time

High-Risk TB Regions Over Time

Regions exceeding fixed and dynamic TB rate thresholds

Year	Regions Above Fixed Threshold (40 per 100k)	Top 10% Threshold	Top 3 Highest-Risk Regions
2012	45.00	37.76	MANAUS, CUIABÁ, SANTOS
2013	45.00	37.75	MANAUS, CUIABÁ, SANTOS
2014	39.00	36.28	MANAUS, CUIABÁ, SANTOS