MTHM506 - Statistical Data Modelling

Joshua Harrison

2025-03-23

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

- I have used GenAI tools to proofread and correct grammar or spelling errors
- I have used GenAI to help cross-check coding/plotting

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

Contents

Quantifying spatio-temporal risk from TB in Brazil

Introduction

Tuberculosis (TB) is a prominent public health issue within Brazil which disproportionately affects vulnerable populations. The disease is caused by Mycobacterium tuberculosis, spreading through airborne transmission and is therefore closely linked to socio-economic conditions. Brazil continues to struggle with high case numbers, particularly in regions with inadequate health infrastructure, poor living conditions and limited access to timely diagnosis and treatment. Therefore, it is imperative that we understand the spatial and temporal distribution of TB cases, as well as the socio-economic factors that drive its spread, in order to effectively intervene and allocate resources efficiently.

The aim of this report is to analyse the risk of TB cases across Brazil using Generalised Additive Models (GAMs). This study aims to understand the extent to which the socio-economic covariates affect TB rate on a spatio-temporal level while providing key insights for public health policy. Using GAMs, we will examine how TB rates differ throughout Brazil's 557 microregions and how they evolve over the years 2012-2014. The key variables that we will explore include 'indigenous' (proportion of indigenous population), 'illiteracy' (a continuous measure of illiteracy levels per microregion), 'Urbanisation', 'Density', 'Poverty', 'Poor Sanitation', 'Unemployment', 'Timeliness' (the average amount of time between diagnosing a TB case and reporting it to the health system), 'Year' (2012-2014), 'TB' (TB cases in each microregion), 'Population', 'Region', 'lon', 'lat'.

Through performing an Explanatory Data Analysis (EDA) we have created a summary table of key metrics for socio-economic factors that affect TB cases within the dataset. The table describes the minimum, maximum, and median values along with the mean and standard deviation of each variable as shown in Appendix A.

Methodology - Fitting the GAM Model

In order to model tuberculosis (TB) cases across Brazil's microregions, we employ GAMs which allow for flexible, non-linear relationships between predictor variables and the response. The GAM framework is an extension of the Generalised Linear Model (GLM), replacing the linear predictor with a sum of smooth functions of the covariates.

Initially, we fit the Poisson model which assumes that the response variable follows a Poisson distribution. This models the mean number of TB cases, μ_i, using a log link function:

$$\begin{aligned}\log(\mu_i) = \beta_0 &+ s(\text{Indigenous}_i) + s(\text{Illiteracy}_i) + s(\text{Urbanisation}_i) + s(\text{Density}_i) + s(\text{Poverty}_i) \\ &+ s(\text{PoorSanitation}_i) + s(\text{Unemployment}_i) + s(\text{Timeliness}_i) + s(\text{Year}_i) + s(\text{lon}_i) + s(\text{lat}_i) \\ &+ \log(\text{Population}_i)\end{aligned}$$

However, the Poisson model assumes equidispersion where the variance equals the mean, which may not hold in this dataset due to variability in TB case counts across Brazil's microregions. To address potential overdispersion and compare model fitting, we fit a Negative Binomial GAM model. This model introduces θ as an additional dispersion parameter, allowing variance to exceed the mean.

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

$$\begin{aligned}\log(\lambda_i) = \beta_0 &+ s(\text{Indigenous}_i, k=10) + s(\text{Illiteracy}_i, k=10) + s(\text{Urbanisation}_i, k=10) \\ &+ s(\text{Density}_i, k=10) + s(\text{Poverty}_i, k=10) + s(\text{PoorSanitation}_i, k=10) + s(\text{Unemployment}_i, k=10) \\ &+ s(\text{Timeliness}_i, k=10) + s(\text{Year}_i, k=3) + s(\text{Region}_i, \text{bs} = \text{"}re\text{"}) + \log(\text{Population}_i)\end{aligned}$$

$s(x, k)$ represents the smooth functions estimated using penalised splines. $s(\text{Region}, \text{bs} = \text{"}re\text{"})$ is modelling the unobserved spatial heterogeneity as a random effect. The offset term in the model ensures that we are modelling TB rates per capita rather than TB counts specifically.

The initial k values selected were too high and led to potential overfitting of the data. For the socioeconomic variables we choose a k value of 10 to provide sufficient flexibility and capture non-linearity. For the year (temporal effect) variable, we initially fit this as a smooth function and choose k as 3. We model Region as a random effect using `bs="re"`. We also fit a negative binomial model using different k values and adding the 'Year' variable as a factor, including longitude and latitude variables. However, when comparing these two negative binomial models, the original model performed better with a greater R-Squared value of 0.854 compared to 0.849. These selections were justified using `k.check(gam_model_nb)` (see Appendix B).

In determining the optimal GAM model for the `TBdata`, we therefore opted for a negative binomial distribution over the Poisson or Gamma distributions (see Appendix B). The negative binomial model accounts for overdispersion and is more suitable for this dataset. The negative binomial model also has a lower AIC value of 14391.60 against the poisson models AIC value of 25632.08. The negative binomial model therefore ensures a better fit, providing more reliable inference in modeling TB risk across different regions and time periods.

To ensure this model fit allows for meaningful predictions of the dataset, we plot the models residuals as shown in Appendix C. The QQ plot of deviance residuals shown on the left assesses the normality of residuals. This shows that the majority of points align well with the theoretical quantities. There deviation is shown in the upper tail, suggesting the presence of higher than expected residuals. The deviance residuals vs linear predictors on the right evaluates whether residuals are randomly dispersed. The residuals here appear relatively homoscedastic, despite some dispersion seen at higher predicted values. Overall, these plots show that while the model provides a reasonable fit, there is potential overdispersion or unaccounted variability that still exists within the data.

Therefore, the final Negative Binomial GAM model accounting for regional variations is represented as:

$$\log(\lambda_i) = \beta_0 + s_1(X_i) + s_2(X_i) + \cdots + s_9(X_i) + f_j(\text{Region}_i) + \log(\text{Population}_i)$$

This is where X_i represents all socioeconomic and temporal variables and $f_j(\text{Region}_i)$ represents the spatial variation across all 557 microregions. $s_j(x_i)$ represents smooth functions estimated via the penalised splines.

Results

From our Negative Binomial GAM model (Appendix D), we can look at the smooth plots which represent the estimated TB rate per 100,000 population as a function of an individual predictor while holding other respective variables constant. The red line represents the estimated smooth function and the shaded area represents the 95% confidence interval. Noticeable factors here include the Indigenous population where higher indigenous population percentages correlate with higher TB rates. This may be due to healthcare access issues, socio-economic disadvantages or living conditions. Poor sanitation displays a strong initial effect and plateaus. This suggests that worsening sanitation rates are strongly linked to higher TB rates but improvements beyond a certain point don't significantly reduce TB incidence. The Density smooth plot displays a non-linear relationship with TB risk increasing with moderate overcrowding at 1.0-1.2 dwellers per room, before steadily declining. This may suggest that transmission rates are highest in moderately crowded areas while extreme crowded areas may have mitigating circumstances that prevent TB rates such as better healthcare access or hygiene rates.

We can also reference the effect the GAM predicted variables have on TB rates by looking at Table 2 in Appendix E. Indigenous population, density, poor sanitation, unemployment, and timeliness all have significant positive effects on TB rates ($p<0.05$). Density is seen to have the largest effect with an effect size of 74.976. Poor sanitation and timeliness have the next largest effect sizes, suggesting the importance of hygiene and a fast diagnosis in the affected TB rates. Unemployment and Indigenous population have large effect sizes of 21 and 22 respectively which suggest socioeconomic vulnerabilities drive TB risk.

To explore how TB risk varies with 'Poverty' and its interaction with key significant variables such as 'Density', 'Urbanisation' and 'Poor Sanitation', we visualise these relationships overtime in Appendix F. The Urbanisation plots indicate that as poverty increases, urbanisation decreases, meaning higher poverty regions are more rural. Although TB rates appear to decline as poverty increases, this may reflect undiagnosed cases

in more rural areas where there is limited healthcare access. Poor sanitation in the bottom row displays a strong positive correlation over the three years with poverty. Higher poverty levels directly link to poor sanitary conditions. However, TB rates as shown by the cluster of bubbles remain relatively constant, suggesting that although poor sanitation contributes to TB risk, other variables may be more dominant drivers.

The spatial plots shown in Appendix G provide a descriptive overview of TB risk across Brazil in 2012 and 2014, calculated as crude incidence rates (cases per 100,000 people). While these results were not created from the GAM model, they highlight the geographical and temporal variation in TB rates prior to formal modelling. In both years, it can be seen that TB risk appears to be clustered in specific regions, such as the North and Southeast, with certain specific microregions showing extreme high rates (shown in lighter shades). While, the temporal pattern appears minimal, it is clear that there is a persistent spatial pattern that cannot be fully explained by socio-economic variables alone.

As a result of this, we created the spatial plot shown in Appendix H which displays the estimated regional effects on TB risk across Brazil, as captured by the smooth random effect term s(Region) in the fitted GAM. The effects were extracted using the `predict()` function with `type="terms"`. This isolates the spatial structure not explained by observed covariates. The resulting values have been merged with the shapefile data (`brasil_micro`) and standardised. We can therefore infer that microregions in the South-West and Central-West show consistently higher positive regional effects, suggesting that these areas have increased TB risk even after accounting for socio-economic covariates. In contrast, Northern regions and much of the North-West show negative regional effects implying that there is relatively lower TB risk once covariates are accounted for.

We then explored any temporal effects shown by the GAM model. Table 3 in Appendix I displays the total number of high-risk TB regions per year based on a fixed threshold of 40 cases per a 100,000 population. While there is a slight decrease in the number of high-risk regions from 2012 to 2014 (42 high risk regions in 2012 to 37 high risk regions in 2014), the same regions (ID's: 35059, 35057, 35062) consistently appear as the highest risk across all 3 years. This temporal consistency is supported by the GAM model output shown in Appendix I, where the smooth term 's(Year)' has an associated p-value of 0.157, suggesting that year-to-year variation does not significantly contribute to explaining TB risk.

The spatio-temporal maps in Appendix J show GAM-predicted TB risk for each microregion from 2012 to 2014. The predictions shown in this visual reflect the combined effects of all modelled covariates, including spatial and regional smooth terms. The maps display a consistent pattern, with the absolute TB risk varying by region but the spatial distribution of high-risk areas remaining stable each year. There are several high-risk clusters persistently located in the South-East of Brazil around São Paulo, this aligns with the previously identified high risk region codes (35059, 35057, 35062). Some microregions along the North-East coast display moderate but table TB risk over the 3 years, while regions in the North and Centre West around the Amazonas, display consistently low predicted risk. These spatially anchored patterns indicate that long-term structural factors such as sanitation, urbanisation and timeliness (linked to healthcare access) are stronger drivers of TB risk than short term temporal changes.

Conclusion

Based on the model's findings, it is suggested that health authorities should focus their investment and allocation of resources initially to the identified microregions (35059, 35057 and 35062). These regions, centred close to São Paulo, consistently exhibit the highest predicted TB risk across all three years. As such, they should be prioritised for investment in TB control programmes including increased screening, early diagnosis or community-based treatment strategies. The clear spatial patterns shown by the GAM model reinforce the importance of using spatially informed risk predictions to guide funding and health infrastructure decisions.

While the GAM framework provided a flexible and interpretable approach to modelling TB risk in our use case scenario, there were several limitations. The final model used in this analysis did not include an explicit spatial smoother over geographic coordinates (s(lon), s(lat)), which could have captured additional unexplained spatial correlation. These variables were chosen to be excluded after analysis of their significance on TB risk. Instead, spatial variation was modelled through a region-level random effect (s(Region, bs

= “re”)), which assumes independence between non-neighbouring regions. Additionally, while covariates were treated as smooth functions, the model may under perform in the presence of sharp discontinuities or thresholds in the data. Key interaction effects between covariates, though visually explored (e.g. Poverty and Sanitation), were not formally included in the model. Using tensor product smooths (e.g. te(Poverty, Sanitation)), may have improved the predictive accuracy of the model.

Despite these limitations, the model addressed the significant underlying socio-economic variables such as unemployment, poverty, and indigenous population proportion. These variables strongly suggest that there are structural socio-economic inequalities that increase TB risk. The strong relationship between poor sanitation and TB risk, especially in poverty-stricken areas, indicate that improving basic infrastructure such as water quality, waste management and housing could have a meaningful impact on reducing TB transmission.

Appendices (code)

Appendix A: EDA of the dataset and its key variables

```
# Summarising key socio-economic indicators and TB cases
# Summarising key socio-economic indicators and TB rate per 100,000 people
summary_table <- data.frame(
  Variable = c("Indigenous", "Urbanisation", "Poverty", "Poor_Sanitation", "Unemployment",
               "Timeliness", "TB Rate (per 100,000)"),
  Mean = round(c(mean(TBdata$Indigenous, na.rm = TRUE),
                 mean(TBdata$Urbanisation, na.rm = TRUE),
                 mean(TBdata$Poverty, na.rm = TRUE),
                 mean(TBdata$Poor_Sanitation, na.rm = TRUE),
                 mean(TBdata$Unemployment, na.rm = TRUE),
                 mean(TBdata$Timeliness, na.rm = TRUE),
                 mean((TBdata$TB / TBdata$Population) * 100000, na.rm = TRUE)), 3),
                                              # TB Rate per 100,000 people
  SD = round(c(sd(TBdata$Indigenous, na.rm = TRUE),
               sd(TBdata$Urbanisation, na.rm = TRUE),
               sd(TBdata$Poverty, na.rm = TRUE),
               sd(TBdata$Poor_Sanitation, na.rm = TRUE),
               sd(TBdata$Unemployment, na.rm = TRUE),
               sd(TBdata$Timeliness, na.rm = TRUE),
               sd((TBdata$TB / TBdata$Population) * 100000, na.rm = TRUE)), 3),
  Min = round(c(min(TBdata$Indigenous, na.rm = TRUE),
                min(TBdata$Urbanisation, na.rm = TRUE),
                min(TBdata$Poverty, na.rm = TRUE),
                min(TBdata$Poor_Sanitation, na.rm = TRUE),
                min(TBdata$Unemployment, na.rm = TRUE),
                min(TBdata$Timeliness, na.rm = TRUE),
                min((TBdata$TB / TBdata$Population) * 100000, na.rm = TRUE)), 3),
  Max = round(c(max(TBdata$Indigenous, na.rm = TRUE),
                max(TBdata$Urbanisation, na.rm = TRUE),
                max(TBdata$Poverty, na.rm = TRUE),
                max(TBdata$Poor_Sanitation, na.rm = TRUE),
                max(TBdata$Unemployment, na.rm = TRUE),
                max(TBdata$Timeliness, na.rm = TRUE),
                max((TBdata$TB / TBdata$Population) * 100000, na.rm = TRUE)), 3),
  Median = round(c(median(TBdata$Indigenous, na.rm = TRUE),
                   median(TBdata$Urbanisation, na.rm = TRUE),
                   median(TBdata$Poverty, na.rm = TRUE),
                   median(TBdata$Poor_Sanitation, na.rm = TRUE),
                   median(TBdata$Unemployment, na.rm = TRUE),
                   median(TBdata$Timeliness, na.rm = TRUE),
                   median((TBdata$TB / TBdata$Population) * 100000, na.rm = TRUE)), 3)
)

# Display formatted table with kableExtra
# Display summary table in LaTeX for PDF
kable(summary_table, format="latex", booktabs=TRUE,
      caption="Summary Statistics of Socio-Economic Indicators and TB Rate per 100,000 People")
```

Table 1: Summary Statistics of Socio-Economic Indicators and TB Rate per 100,000 People

Variable	Mean	SD	Min	Max	Median
Indigenous	0.843	3.530	0.010	50.646	0.106
Urbanisation	71.961	16.528	22.336	99.927	72.655
Poverty	44.371	19.368	5.923	77.883	42.603
Poor_Sanitation	16.449	12.510	0.047	58.433	13.913
Unemployment	6.930	2.561	1.128	20.438	6.782
Timeliness	47.668	21.475	0.000	96.685	48.361
TB Rate (per 100,000)	23.540	15.419	0.000	117.726	20.106

Appendix B: Fitting GAM Models - Poisson vs Negative Binomial

```
# Fit a GAM model using poisson distribution
gam_model_p <- gam(TB ~ s(Indigenous) + s(Illiteracy) + s(Urbanisation) + s(Density) +
                  s(Poverty) + s(Poor_Sanitation) + s(Unemployment) + s(Timeliness) + s(lon) + s(lat) +
                  s(Year, bs="re") + offset(log(Population)),
                  family=poisson, data=TBdata)
```

```
# Fitting a Negative Binomial GAM with offsetting population
gam_model_nb <- gam(TB ~
                      s(Indigenous, k=10) +
                      s(Illiteracy, k=10) +
                      s(Urbanisation, k=10) +
                      s(Density, k=10) +
                      s(Poverty, k=10) +
                      s(Poor_Sanitation, k=10) +
                      s(Unemployment, k=10) +
                      s(Timeliness, k=10) +
                      s(Year, k=3) +
                       s(Region, bs="re"),
                    offset=log(Population),
                    family=nb(),
                    data=TBdata)


# Summary of the model
summary(gam_model_nb)
```

```
##
## Family: Negative Binomial(6.201)
## Link function: log
##
## Formula:
## TB ~ s(Indigenous, k = 10) + s(Illiteracy, k = 10) + s(Urbanisation,
##     k = 10) + s(Density, k = 10) + s(Poverty, k = 10) + s(Poor_Sanitation,
##     k = 10) + s(Unemployment, k = 10) + s(Timeliness, k = 10) +
##     s(Year, k = 3) + s(Region, bs = "re")
##
## Parametric coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -8.6758     0.0534  -162.5   <2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                       edf Ref.df  Chi.sq  p-value
## s(Indigenous)      1.0162  1.032  22.646 2.42e-06 ***
## s(Illiteracy)      1.0032  1.006   0.353 0.555114
## s(Urbanisation)    6.8247  7.938  28.184 0.000342 ***
## s(Density)         4.4468  5.528 158.106  < 2e-16 ***
## s(Poverty)         3.9528  4.948   7.230 0.173580
## s(Poor_Sanitation) 6.1503  7.325  83.115  < 2e-16 ***
## s(Unemployment)    5.7413  6.944  51.678  < 2e-16 ***
## s(Timeliness)      3.9000  4.861  72.115  < 2e-16 ***
## s(Year)            1.0316  1.062   2.005 0.157569
## s(Region)          0.9571  1.000  22.293 1.44e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.854   Deviance explained = 44.4%
## -REML = 7231.2  Scale est. = 1         n = 1671
```

```
k.check(gam_model_nb)
```

```
##                     k'       edf   k-index p-value
## s(Indigenous)        9 1.0162344 0.4904663       0
## s(Illiteracy)        9 1.0031661 0.4837377       0
## s(Urbanisation)      9 6.8246655 0.5003842       0
## s(Density)           9 4.4468312 0.4974882       0
## s(Poverty)           9 3.9527595 0.4885257       0
## s(Poor_Sanitation)   9 6.1503439 0.4968098       0
## s(Unemployment)      9 5.7412575 0.4985584       0
## s(Timeliness)        9 3.9000106 0.5564241       0
## s(Year)              2 1.0315907 0.6550155       0
## s(Region)            1 0.9570863 0.3842984       0
```

```
# Compare AIC values
AIC(gam_model_p, gam_model_nb)
```

```
##                     df      AIC
## gam_model_p  89.51426 25632.08
## gam_model_nb 43.64079 14391.60
```

```r
# Treating Year as a factor using $
TBdata$YearFactor <- as.factor(TBdata$Year)

# Fitting the Negative Binomial GAM with Year as a facto and including lon, lat terms
gam_model_nb2 <- gam(TB ~
                       s(Indigenous, k=10) +
                       s(Illiteracy, k=10) +
                       s(Urbanisation, k=10) +
                       s(Density, k=10) +
                       s(Poverty, k=10) +
                       s(Poor_Sanitation, k=10) +
                       s(Unemployment, k=10) +
                       YearFactor +                    # Factor variable (linear effect)
                       s(Region, bs="re") +            # Random effect for regions
                       s(lon, lat, k=20),              # Spatial smooth
                       offset = log(Population),
                       family = nb(),
                       data = TBdata)

# Summary of the model
summary(gam_model_nb2)
```

```
##
## Family: Negative Binomial(7.502)
## Link function: log
##
## Formula:
## TB ~ s(Indigenous, k = 10) + s(Illiteracy, k = 10) + s(Urbanisation,
##     k = 10) + s(Density, k = 10) + s(Poverty, k = 10) + s(Poor_Sanitation,
##     k = 10) + s(Unemployment, k = 10) + YearFactor + s(Region,
##     bs = "re") + s(lon, lat, k = 20)
##
## Parametric coefficients:
##                 Estimate Std. Error  z value Pr(>|z|)
## (Intercept)    -8.360501   0.060441 -138.326   <2e-16 ***
## YearFactor2013  0.001907   0.024431    0.078    0.938
## YearFactor2014 -0.040060   0.024468   -1.637    0.102
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                       edf Ref.df  Chi.sq  p-value
## s(Indigenous)       1.0029  1.006  14.837  0.00012 ***
## s(Illiteracy)       4.0262  5.058  12.208  0.03394 *
## s(Urbanisation)     4.5238  5.648  34.878 3.34e-06 ***
## s(Density)          4.1698  5.240  63.590  < 2e-16 ***
## s(Poverty)          1.0044  1.008   2.002  0.15733
## s(Poor_Sanitation)  6.5872  7.711  73.267  < 2e-16 ***
## s(Unemployment)     6.3165  7.497 103.406  < 2e-16 ***
## s(Region)           0.5683  1.000   1.244  0.12190
## s(lon,lat)         17.5203 18.758 377.948  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## R-sq.(adj) =  0.849   Deviance explained = 52.5%
## -REML =   7131  Scale est. = 1         n = 1671
```

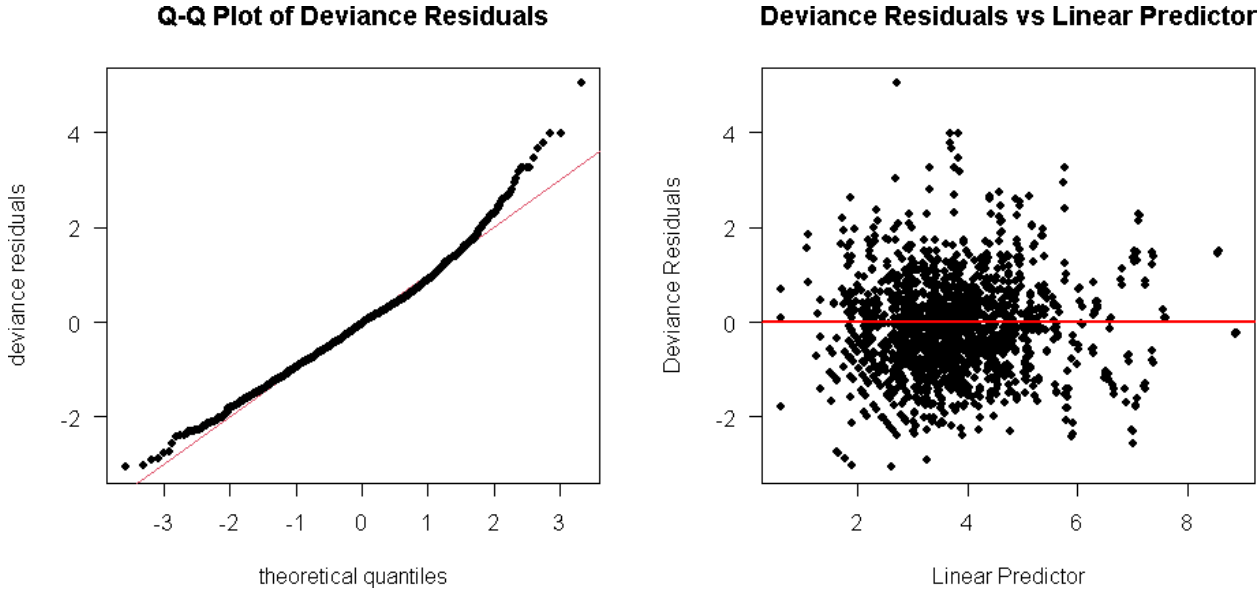
```
# Check smoothing basis
k.check(gam_model_nb2)
```

```
##                     k'        edf   k-index p-value
## s(Indigenous)        9  1.0029376 0.5175600       0
## s(Illiteracy)        9  4.0261799 0.5049460       0
## s(Urbanisation)      9  4.5237563 0.5160126       0
## s(Density)           9  4.1697707 0.5157949       0
## s(Poverty)           9  1.0044273 0.5289279       0
## s(Poor_Sanitation)   9  6.5872315 0.5157335       0
## s(Unemployment)      9  6.3164584 0.5214862       0
## s(Region)            1  0.5683042 0.4477090       0
## s(lon,lat)          19 17.5202745 0.4659458       0
```

The R squared value in this second revised model is lower than the original where lon and lat variables are not included. Therefore, we will begin the model fitting process with the original negative binomial model. It is also worth noting that Year as a factor does not produce significant results as shown in the summary table above.

Appendix C: QQ Plot of Deviance Residuals and Deviance Residuals vs Linear Predictors

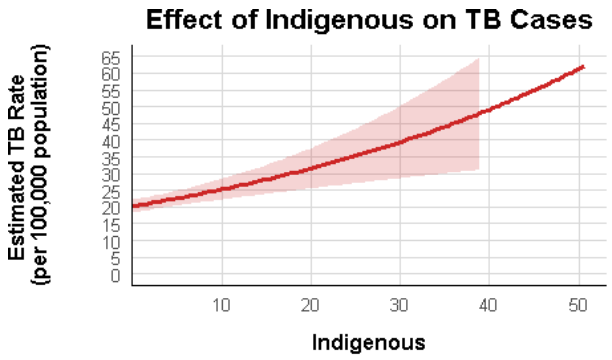
```r
layout(matrix(1:2, nrow=1))

# QQ plot
qq.gam(gam_model_nb, pch=20, cex=1.2, las=1, main="Q-Q Plot of Deviance Residuals")

# Deviance residuals vs linear predictor
xx <- gam_model_nb$linear.predictors
yy <- residuals(gam_model_nb, type="deviance")

plot(xx, yy, pch=20, cex=1.2, xlab="Linear Predictor", ylab="Deviance Residuals",
     main="Deviance Residuals vs Linear Predictor", las=1)
abline(h=0, col="red", lwd=2)
```

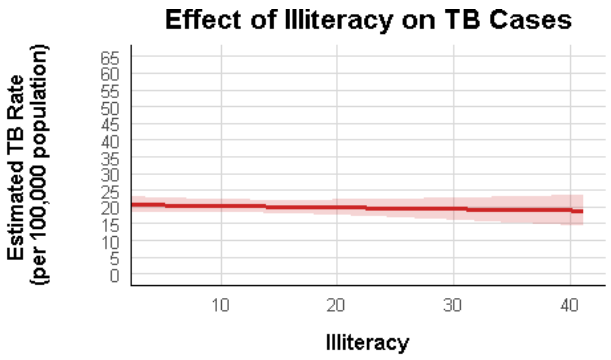


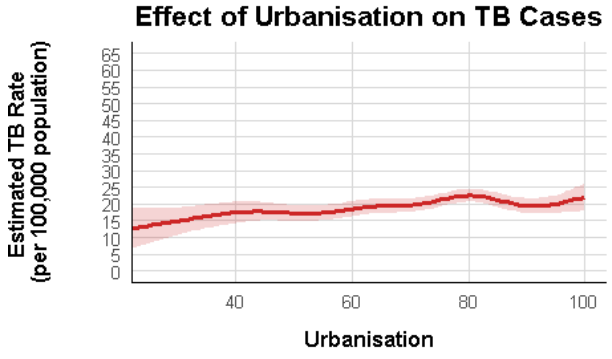
Appendix D: Creating plots to visualise the socio-economic factors affecting TB using the negative binomial GAM model.

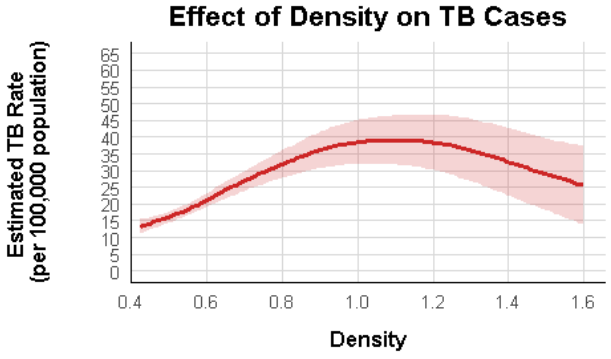
```
# Define global variables
variables <- c("Indigenous", "Illiteracy", "Urbanisation", "Density",
               "Poverty", "Poor_Sanitation", "Unemployment", "Timeliness")

scaling_factor <- 100000  # per 100,000 people

# Fixed y-axis limits and breaks
y_limits <- c(0, 65)
y_breaks <- seq(0, 65, by = 5)

# Helper to clean variable names
clean_variable_name <- function(variable) {
  variable <- str_replace_all(variable, "_", " ")
  variable <- str_to_title(variable)
  return(variable)
}

# Function to generate smooth plots with correct X-Axis scale
plot_smooth <- function(model, variable, data, y_limits, y_breaks, scaling_factor) {
  if (!(variable %in% names(data))) {
    stop(paste("Variable", variable, "not found in dataset"))
  }

  # Construct new data
  new_data <- data.frame(lapply(data,
                                function(x) if(is.numeric(x)) median(x, na.rm=TRUE) else x[1]))
  new_data <- new_data[rep(1, 100), ]
  new_data[[variable]] <- seq(min(data[[variable]], na.rm=TRUE),
                              max(data[[variable]], na.rm=TRUE), length.out=100)
  pred <- predict(model, newdata=new_data, type="response", se.fit=TRUE)

  # Scale predictions
  plot_data <- data.frame(
    variable = new_data[[variable]],
    fit = pred$fit * scaling_factor,
    se = pred$se.fit * scaling_factor
  )

  # **Fix X-axis scale for "Density"**
  x_limits <- if (variable == "Density") c(0.4, 1.6) else range(plot_data$variable)
  x_breaks <- if (variable == "Density") seq(0.4, 1.6, by=0.2) else pretty(plot_data$variable, n=5)

  # Create plot with **correct X-Axis & Aspect Ratio**
  ggplot(plot_data, aes(x=variable, y=fit)) +
    geom_line(color="firebrick3", size=1.2) +  # Red line
    geom_ribbon(aes(ymin=fit-1.96*se, ymax=fit+1.96*se), alpha=0.2, fill="firebrick3") +
    scale_x_continuous(limits=x_limits, breaks=x_breaks, expand = expansion(mult = c(0, 0.05))) +
    scale_y_continuous(labels = scales::comma, limits=y_limits, breaks=y_breaks) +
    labs(
      title = paste("Effect of", clean_variable_name(variable), "on TB Cases"),
      x = clean_variable_name(variable),
```

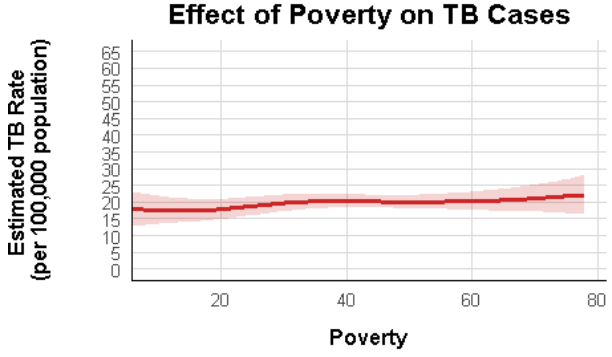
```
      y = "Estimated TB Rate \n(per 100,000 population)"
    ) +
    theme_minimal(base_size = 14) +
    theme(
      plot.title = element_text(hjust = 0.5, size=16, face="bold"),
      axis.title.x = element_text(size=12, face="bold", margin=margin(t=10)),
      axis.title.y = element_text(size=12, face="bold", margin=margin(r=30)),
      axis.text = element_text(size=10),
      panel.grid.major = element_line(color="gray85"),
      panel.grid.minor = element_blank(),
      panel.border = element_blank(),
      axis.line = element_line(color="black"),
      plot.margin = margin(10, 10, 10, 10)
    )
}

# Generate all smooth plots
plots <- lapply(variables, function(var) plot_smooth(gam_model_nb, var, TBdata,
                                                     y_limits, y_breaks, scaling_factor))

# Arrange plots in grid with 2 columns
grid.arrange(grobs = plots, ncol = 2)
```

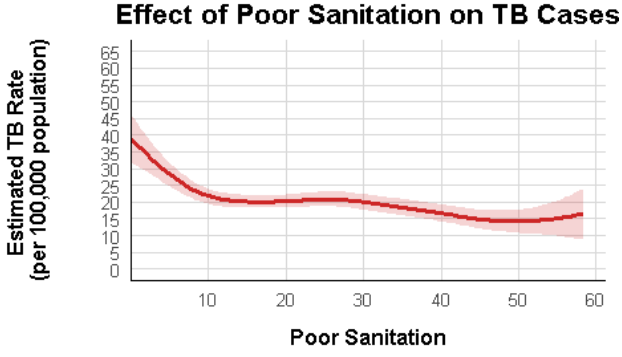

Effect of Indigenous on TB Cases
Estimated TB Rate (per 100,000 population)
0 5 10 15 20 25 30 35 40 45 50 55 60 65
10 20 30 40 50
Indigenous

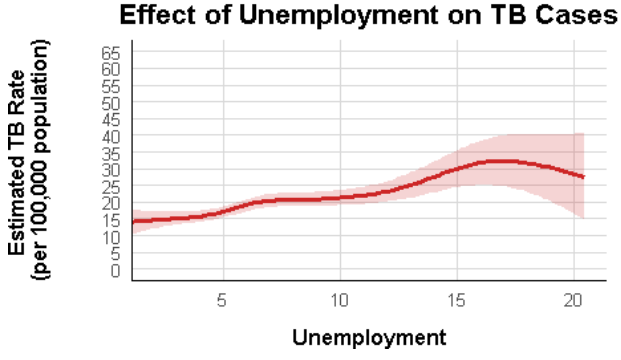

Effect of Illiteracy on TB Cases
Estimated TB Rate (per 100,000 population)
0 5 10 15 20 25 30 35 40 45 50 55 60 65
10 20 30 40
Illiteracy

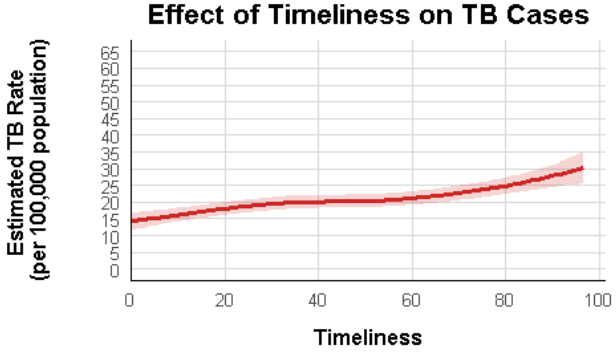

Effect of Urbanisation on TB Cases
Estimated TB Rate (per 100,000 population)
0 5 10 15 20 25 30 35 40 45 50 55 60 65
40 60 80 100
Urbanisation


Effect of Density on TB Cases
Estimated TB Rate (per 100,000 population)
0 5 10 15 20 25 30 35 40 45 50 55 60 65
0.4 0.6 0.8 1.0 1.2 1.4 1.6
Density


Effect of Poverty on TB Cases
Estimated TB Rate (per 100,000 population)
0 5 10 15 20 25 30 35 40 45 50 55 60 65
20 40 60 80
Poverty


Effect of Poor Sanitation on TB Cases
Estimated TB Rate (per 100,000 population)
0 5 10 15 20 25 30 35 40 45 50 55 60 65
10 20 30 40 50 60
Poor Sanitation


Effect of Unemployment on TB Cases
Estimated TB Rate (per 100,000 population)
0 5 10 15 20 25 30 35 40 45 50 55 60 65
5 10 15 20
Unemployment


Effect of Timeliness on TB Cases
Estimated TB Rate (per 100,000 population)
0 5 10 15 20 25 30 35 40 45 50 55 60 65
0 20 40 60 80 100
Timeliness

Appendix E: Table showing the effect that the predicted variables have on TB rates

```
coeffs <- broom::tidy(gam_model_nb) %>%
  filter(!term %in% "(Intercept)") %>%
  mutate(
    Effect_Size = statistic / sqrt(edf),
    term = gsub("s\\(|\\)", "", term),
    `P-Value` = ifelse(p.value < 0.05, "p < 0.05", sprintf("%.3f", p.value))
  )

# Ensure coeffs is a data frame
coeffs <- as.data.frame(coeffs)

coeffs %>%
  dplyr::select(term, Effect_Size, edf, statistic, `P-Value`) %>%
  rename(
    Predictor = term,
    `Effect Size` = Effect_Size,
    EDF = edf,
    Statistic = statistic
  ) %>%
  kable("latex", digits = 3, caption = "Effect Size of Predictors on TB Rate", booktabs = TRUE) %>%
  kable_styling(latex_options = c("striped", "hold_position")) %>%
  row_spec(0, bold = TRUE) %>%
  footnote(general = "Based on Generalized Additive Model (GAM) Results",
           general_title = "Note:", footnote_as_chunk = TRUE)
```

Table 2: Effect Size of Predictors on TB Rate

Predictor	Effect Size	EDF	Statistic	P-Value
Indigenous	22.464	1.016	22.646	$p < 0.05$
Illiteracy	0.353	1.003	0.353	0.555
Urbanisation	10.788	6.825	28.184	$p < 0.05$
Density	74.976	4.447	158.106	$p < 0.05$
Poverty	3.637	3.953	7.230	0.174
Poor_Sanitation	33.514	6.150	83.115	$p < 0.05$
Unemployment	21.568	5.741	51.678	$p < 0.05$
Timeliness	36.517	3.900	72.115	$p < 0.05$
Year	1.974	1.032	2.005	0.158
Region	22.788	0.957	22.293	$p < 0.05$

Note: Based on Generalized Additive Model (GAM) Results

Appendix F: Relationship of Poverty against Significant Variables on the Rate of TB Cases

```
# Computing the predicted rate again (ensuring TB rate is per 100,000)
predicted_rate <- predict(gam_model_nb, type = "response", newdata = TBdata) * 100000


# Urbanisation Plot
p1 <- ggplot(TBdata, aes(x = Poverty, y = Urbanisation)) +
  geom_jitter(aes(size = TB, color = predicted_rate), alpha = 0.9, width = 1, height = 1) +
  geom_smooth(color = "black", method = "loess", se = FALSE, linewidth = 0.7) +

  scale_color_gradientn(
    colors = c("#4A1486", "#DF65B0", "#FED976"),
    name = "TB Predicted Rate",
    labels = scales::comma_format(accuracy = 1)
  ) +

  scale_size_continuous(
    range = c(2, 8),
    name = "TB Cases"
  ) +

  facet_wrap(~Year, nrow = 1) +
  theme_minimal() +
  theme(
    strip.text = element_text(size = 10, face = "bold"),
    axis.title.y = element_text(size = 10),
    axis.text = element_text(size = 8),
    legend.position = "bottom"
  ) +
  labs(
    x = NULL,
    y = "Urbanisation"
  )

# Density Plot
p2 <- ggplot(TBdata, aes(x = Poverty, y = Density)) +
  geom_jitter(aes(size = TB, color = predicted_rate), alpha = 0.9, width = 1, height = 1) +
  geom_smooth(color = "black", method = "loess", se = FALSE, linewidth = 0.7) +

  scale_color_gradientn(
    colors = c("#4A1486", "#DF65B0", "#FED976"),,
    name = "TB Predicted Rate",
    labels = function(x) round(x * mean(TBdata$Population), 0)
  ) +

  scale_size_continuous(
    range = c(2, 8),
    name = "TB Cases"
  ) +

  facet_wrap(~Year, nrow = 1) +
  theme_minimal() +
```

```
  theme(
    strip.text = element_blank(),
    axis.title.y = element_text(size = 10),
    axis.text = element_text(size = 8),
    legend.position = "none"
  ) +
  labs(
    x = "Poverty Level",
    y = "Density"
  )

# Poor Sanitation Plot
p3 <- ggplot(TBdata, aes(x = Poverty, y = Poor_Sanitation)) +
  geom_jitter(aes(size = TB, color = predicted_rate), alpha = 0.9, width = 1, height = 1) +
  geom_smooth(color = "black", method = "loess", se = FALSE, linewidth = 0.7) +

  scale_color_gradientn(
    colors = c("#4A1486", "#DF65B0", "#FED976"),
    name = "TB Predicted Rate",
    labels = function(x) round(x * mean(TBdata$Population), 0)
  ) +

  scale_size_continuous(
    range = c(2, 8),
    name = "TB Cases"
  ) +

  facet_wrap(~Year, nrow = 1) +
  theme_minimal() +
  theme(
    strip.text = element_blank(),
    axis.title.y = element_text(size = 10),
    axis.text = element_text(size = 8),
    legend.position = "none"
  ) +
  labs(
    x = NULL,
    y = "Poor Sanitation"
  )


# Combine Plots, Collecting Legend
combined_plot <- (p1 / p2 / p3) +
  plot_layout(guides = "collect") +
  plot_annotation(
    title = "Relationship Between Poverty and Socio-Economic Factors",
    theme = theme(
      plot.title = element_text(size = 14, face = "bold", hjust = 0.5),
      plot.caption = element_text(size = 10),
      legend.position = "bottom",
      legend.margin = margin(t = 10)
    )
  )
```

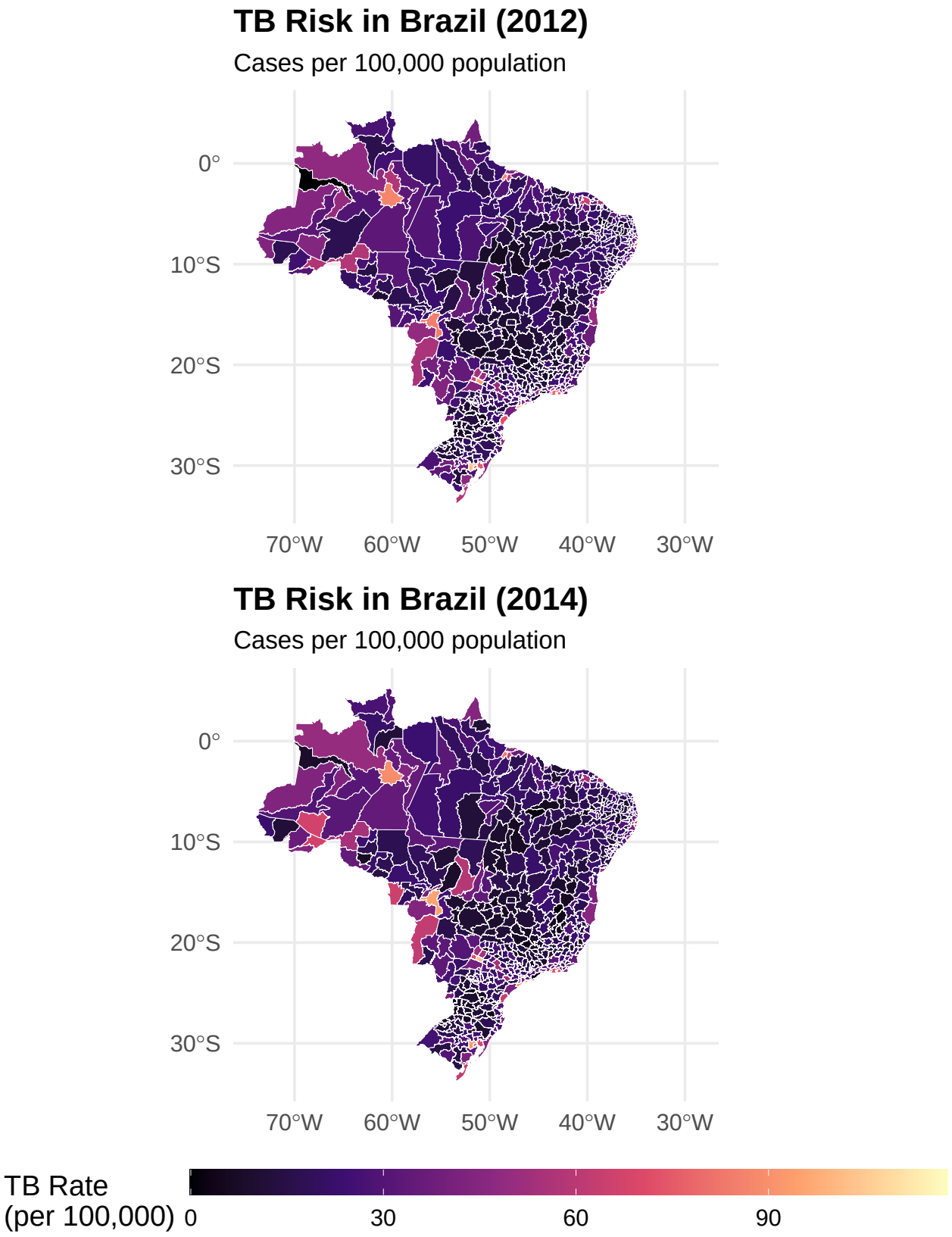
```r
# Improve Legend Styling
combined_plot <- combined_plot +
  guides(
    color = guide_colorbar(
      barwidth = 8,
      barheight = 0.5,
      title.position = "top",
      title.theme = element_text(size = 10)
    ),
    size = guide_legend(
      keywidth = 0.5,
      keyheight = 0.5,
      title.position = "top",
      title.theme = element_text(size = 10)
    )
  )

# Display the Plot
print(combined_plot)
```

Appendix G: Creating Spatial-Temporal Plot of TB risk

```
# Convert to sf format
brasil_sf <- st_as_sf(brasil_micro)

# Computing TB rate per 100,000 population for 2012 & 2014
brasil_sf <- brasil_sf %>%
  mutate(
    TBcases.2012 = ifelse(is.na(TBcases.2012), 0, TBcases.2012),
    Pop2012 = ifelse(is.na(Pop2012) | Pop2012 == 0, median(Pop2012, na.rm=TRUE), Pop2012),
    TB_rate_2012 = (TBcases.2012 / Pop2012) * 100000,

    TBcases.2014 = ifelse(is.na(TBcases.2014), 0, TBcases.2014),
    Pop2014 = ifelse(is.na(Pop2014) | Pop2014 == 0, median(Pop2014, na.rm=TRUE), Pop2014),
    TB_rate_2014 = (TBcases.2014 / Pop2014) * 100000
  )
```

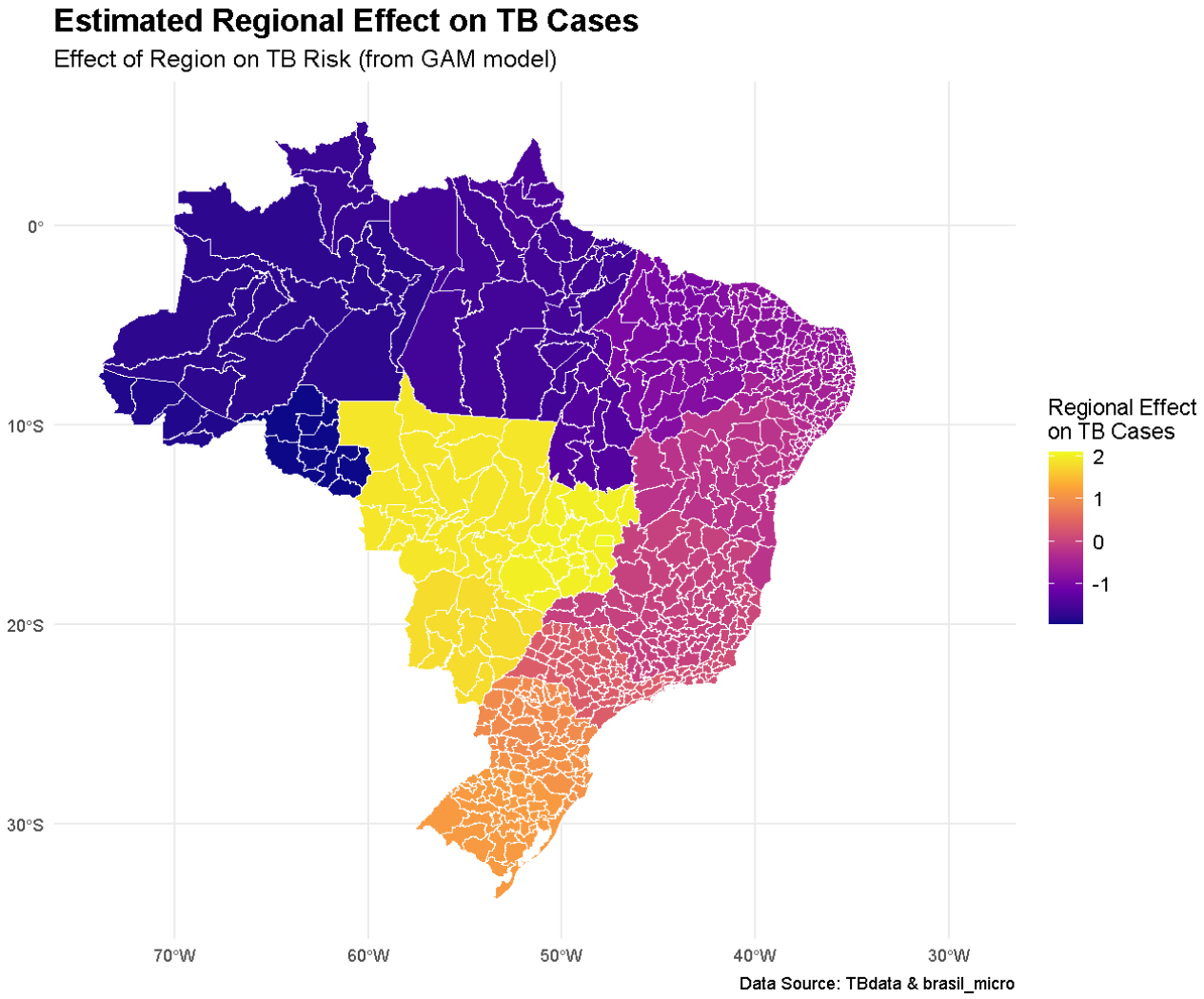
```
# Scale limits to keep both maps consistent
tb_min <- min(c(brasil_sf$TB_rate_2012, brasil_sf$TB_rate_2014), na.rm=TRUE)
tb_max <- max(c(brasil_sf$TB_rate_2012, brasil_sf$TB_rate_2014), na.rm=TRUE)

# Create TB map for 2012
plot_2012 <- ggplot(brasil_sf) +
  geom_sf(aes(fill=TB_rate_2012), color="white", size=0.1) +
  scale_fill_viridis_c(option="magma", name="TB Rate\n(per 100,000)", na.value="grey90",
                       limits=c(tb_min, tb_max)) +
  labs(title="TB Risk in Brazil (2012)", subtitle="Cases per 100,000 population") +
  theme_minimal(base_size=16) +
  theme(
    legend.position="bottom",
    legend.key.width = unit(3, "cm"),
    legend.key.height = unit(0.5, "cm"),
    plot.title = element_text(size=18, face="bold"),
    plot.subtitle = element_text(size=14)
  )

# Create TB map for 2014
plot_2014 <- ggplot(brasil_sf) +
  geom_sf(aes(fill=TB_rate_2014), color="white", size=0.1) +
  scale_fill_viridis_c(option="magma", name="TB Rate\n(per 100,000)", na.value="grey90",
                       limits=c(tb_min, tb_max), guide = "none") +
  labs(title="TB Risk in Brazil (2014)", subtitle="Cases per 100,000 population") +
  theme_minimal(base_size=16) +
  theme(
    plot.title = element_text(size=18, face="bold"),
    plot.subtitle = element_text(size=14)
  )

# Combine plots vertically
combined_plot <- (plot_2012 / plot_2014) +
  plot_layout(guides = "collect") &
  theme(legend.position = "bottom")

# Print the final combined plot
```

```
print(combined_plot)
```



Appendix H: Spatial Plot displaying Regional Effects shown by the GAM model

```
# Extract smooth function estimates for the region effect
region_effects <- data.frame(
  COD_MICRO = unique(TBdata$Region),
  # Extracting region-specific effect
  Effect = predict(gam_model_nb, type="terms", terms="s(Region)", bs="re")
)

# Merging regional effects into the spatial dataset
brasil_sf <- brasil_sf %>%
  left_join(region_effects, by = "COD_MICRO")

brasil_sf <- brasil_sf %>%
  rename(Effect = s.Region.)

brasil_sf$Effect <- scale(brasil_sf$Effect) # Standardising the effect by normalising the scale



gam_spatial_plot <- ggplot(brasil_sf) +
  geom_sf(aes(fill = Effect), color = "white", size = 0.1) +  # Plotting spatial variation
  scale_fill_viridis_c(option="plasma", name="Regional Effect\non TB Cases") +
  labs(title="Estimated Regional Effect on TB Cases",
       subtitle="Effect of Region on TB Risk (from GAM model)",
       caption="Data Source: TBdata & brasil_micro") +
  theme_minimal() +
  theme(legend.position="right",
        legend.text = element_text(size=10),
        plot.title = element_text(size=16, face="bold"),
        plot.subtitle = element_text(size=12))

gam_spatial_plot
```


Estimated Regional Effect on TB Cases
Effect of Region on TB Risk (from GAM model)
0°
10°S
20°S
30°S
70°W
60°W
50°W
40°W
30°W
Regional Effect on TB Cases
2
1
0
-1
Data Source: TBdata & brasil_micro

Appendix I: Temporal Effect of the GAM Model

```
summary(gam_model_nb) # Checking the EDF of 'Year' variable and its significance.
```

```
##
## Family: Negative Binomial(6.201)
## Link function: log
##
## Formula:
## TB ~ s(Indigenous, k = 10) + s(Illiteracy, k = 10) + s(Urbanisation,
##     k = 10) + s(Density, k = 10) + s(Poverty, k = 10) + s(Poor_Sanitation,
##     k = 10) + s(Unemployment, k = 10) + s(Timeliness, k = 10) +
##     s(Year, k = 3) + s(Region, bs = "re")
##
## Parametric coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -8.6758     0.0534  -162.5   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                       edf Ref.df  Chi.sq  p-value
## s(Indigenous)      1.0162  1.032  22.646 2.42e-06 ***
## s(Illiteracy)      1.0032  1.006   0.353 0.555114
## s(Urbanisation)    6.8247  7.938  28.184 0.000342 ***
## s(Density)         4.4468  5.528 158.106  < 2e-16 ***
## s(Poverty)         3.9528  4.948   7.230 0.173580
## s(Poor_Sanitation) 6.1503  7.325  83.115  < 2e-16 ***
## s(Unemployment)    5.7413  6.944  51.678  < 2e-16 ***
## s(Timeliness)      3.9000  4.861  72.115  < 2e-16 ***
## s(Year)            1.0316  1.062   2.005 0.157569
## s(Region)          0.9571  1.000  22.293 1.44e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.854   Deviance explained = 44.4%
## -REML = 7231.2  Scale est. = 1         n = 1671
```

s(Year) alone is shown to not be a significant variable as shown from the summary of the negative binomial GAM model. Removing s(Year) could therefore improve the model with not much loss to predictive power. However, if we relate rate of TB cases per year predicted by the GAM model we can create meaningful inferences.

```
TBdata_new <- TBdata %>%
  dplyr::select(`Region`, `Year`, `Indigenous`, `Illiteracy`, `Urbanisation`,
                `Density`, `Poverty`, `Poor_Sanitation`, `Unemployment`, `Population`, `Timeliness`)

# Predict TB rates for all years at once
TBdata_new$Predicted_TB <- predict(gam_model_nb, newdata = TBdata_new, type = "response")

# Rescale the predicted TB rates
TBdata_new <- TBdata_new %>%
  mutate(Predicted_TB_Scaled = Predicted_TB * 100000) # Scale to per 100,000 population
```

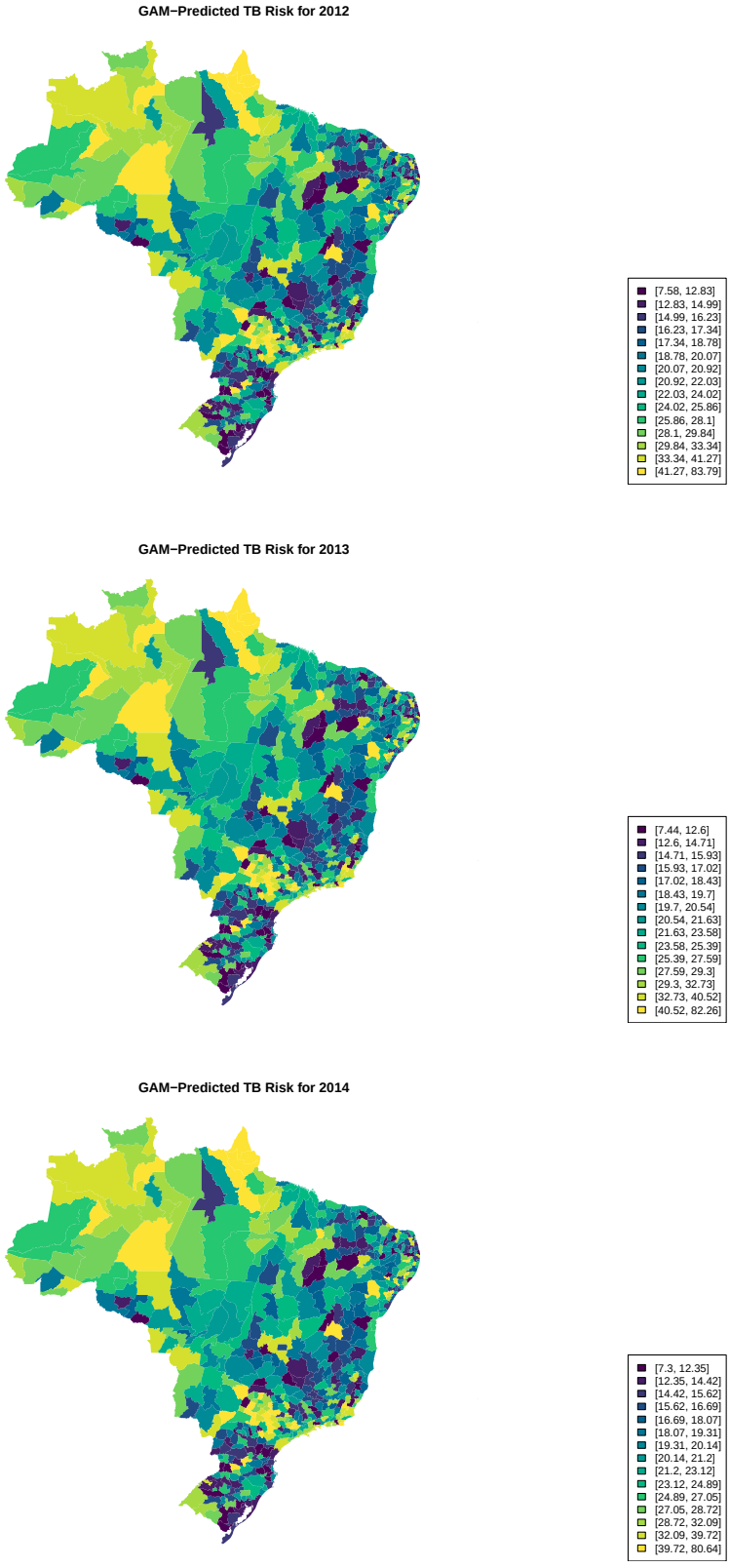
```r
# Creating Table to Indicate the change in High-Risk regions for TB across Brazil each year

# Defining high-risk threshold (top 10%)
high_risk_threshold <- TBdata_new %>%
  group_by(Year) %>%
  summarise(threshold = quantile(Predicted_TB_Scaled, 0.90, na.rm = TRUE))

# Defining a fixed high-risk threshold (40 cases per 100,000)
fixed_threshold <- 40

# Count high-risk regions exceeding the fixed threshold
high_risk_regions_fixed <- TBdata_new %>%
  filter(Predicted_TB_Scaled > fixed_threshold) %>%
  group_by(Year) %>%
  summarise(Total_High_Risk_Regions = n())

# Identify highest-risk regions per year
top_high_risk_regions <- TBdata_new %>%
  group_by(Year) %>%
  arrange(desc(Predicted_TB_Scaled)) %>%
  slice_head(n = 3) %>%
  summarise(Highest_Risk_Regions = paste(Region, collapse = ", "))

# Merge data to include highest-risk regions
high_risk_regions_fixed <- left_join(high_risk_regions_fixed, top_high_risk_regions, by = "Year")



# Remove underscores from column names
colnames(high_risk_regions_fixed) <- str_replace_all(colnames(high_risk_regions_fixed), "_", " ")

# Display table in LaTeX format
kable(high_risk_regions_fixed, format = "latex", booktabs = TRUE,
      caption = "Total High-Risk TB Regions Per Year in Brazil (Fixed Threshold)",
      align = "c") %>%
  kable_styling(latex_options = c("striped", "hold_position", "scale_down")) %>%
  column_spec(1, width = "10em") %>%
  column_spec(2, width = "10em") %>%
  column_spec(3, width = "15em")
```

Table 3: Total High-Risk TB Regions Per Year in Brazil (Fixed Threshold)

Year	Total High Risk Regions	Highest Risk Regions
2012	42	35059, 35057, 35062
2013	38	35059, 35057, 35062
2014	37	35059, 35057, 35062

Appendix J: Spatio-Temporal Plot of TB cases in Brazil

```r
# Define a common color scale
tb_min <- min(TBdata_new$Predicted_TB_Scaled, na.rm = TRUE)
tb_max <- max(TBdata_new$Predicted_TB_Scaled, na.rm = TRUE)

par(mfcol = c(3, 1), mar = c(4, 4, 2, 2), oma = c(6, 5, 5, 5))

# Looping through each year and plot
for (yr in c(2012, 2013, 2014)) {

  TB_subset <- TBdata_new$Predicted_TB_Scaled[TBdata_new$Year == yr]

  if (length(TB_subset) == 0) {
    warning(paste("No data found for year", yr))
    next  # Skip this iteration if no data
  }

  plot.map(
    TB_subset,  # Subset data for the specific year
    n.levels = 15,
    main = paste("GAM-Predicted TB Risk for", yr)
  )
}

# Add multi-line caption below all plots
mtext(
  "Legend: Predicted TB incidence rates per 100,000 population.",
  side = 1, line = 3, outer = TRUE, cex = 0.9
)
mtext(
  "Darker shades indicate lower predicted TB risk; lighter shades indicate higher risk.",
  side = 1, line = 4, outer = TRUE, cex = 0.9
)
mtext(
  "Values shown are scaled predictions from the GAM model, split into 15 quantile-based levels.",
  side = 1, line = 5, outer = TRUE, cex = 0.9
)
```



Legend: Predicted TB incidence rates per 100,000 population.
Darker shades indicate lower predicted TB risk; lighter shades indicate higher risk.
Values shown are scaled predictions from the GAM model, split into 15 quantile−based levels.