The effects of Water and Sanitation on Tuberculosis Incidence in South East Asian and African countries

Warning: package 'knitr' was built under R version 4.4.3

Warning: package 'dplyr' was built under R version 4.4.3

**Authors**

* Muhammad Nasir (ORCID: 0000-0000-1234-5678)

**Author affiliations**

1. College of Public Health, University of Georgia, Athens, GA, USA.
2. Faculty of Public Health, Universitas Tadulako, Palu, Indonesia.

These authors contributed equally to this work.

Corresponding author: some@muhammad.nasir@uga.edu

Disclaimer: The opinions expressed in this article are the author’s own and don’t reflect their employer.

# 1. Abstract

Tuberculosis (TB) remains a significant global health concern, particularly in low- and middle-income countries. In 2023, an estimated 10.8 million people contracted TB, with nearly 70% of cases reported in South-East Asia and Africa. While TB is primarily airborne, environmental and socioeconomic factors, including water and sanitation access, may influence transmission by affecting hygiene and immune function. This study aims to investigate the impact of access to water and sanitation services on TB incidence across South-East Asia and Africa, incorporating the Human Development Index (HDI) as a control for socioeconomic context. A Generalized Additive Mixed Model (GAMM) was applied to longitudinal data from 2000 to 2022. The model included smooth terms for year, water and sanitation indicators, and HDI, with country as a random effect.The model showed excellent fit (adjusted R² = 0.997), with significant non-linear effects observed for year, access to basic water (p = 0.013), safe water (p < 0.001), and HDI (p < 0.001). Sanitation variables were not independently significant. However, key interactions—such as between water access and HDI—were highly significant (p < 0.001), suggesting complex environmental-development dynamics influencing TB risk.In conclusion, expanding water access and addressing development disparities may be critical in reducing TB burden in high-incidence regions.

# 2. Introduction

## 2.1 General Background Information

Tuberculosis is an old infectious disease which still exists and become major public health in the world, particularly in developing countries. In 2023, World Health Organization (WHO) estimated 10.8 million people infected by TB globally with (95% uncertainty interval 10.1-11.7 million), which increased from 10.7 million in 2022. The number is equivalent to 134 incident cases per 100 000 population. 45% and 24% of the cases were detected in South-East Asia and Africa repectively.(1) Tuberculosis has been associated with socioeconomic status, housing conditions, nutrition and multiple environmental factors. Climate Change Phenomena also is suspected to affect tubercolusos both direct and indirect such as temperature, precipitation and humidity change (2) and worseing air polution (3), agent survival(4,5), water resources destruption, crop destruption, and economic destruption, and nutrition (6–10). In term of Environmental factors on Tuberculosis, most studies examined climatic indicators and air poluttion. There is alimited studies examine the effect of Water and Sanitation on tubercolusis. Althought Tuberculosis is an airborne diseases, lack of access to water and sanitation probably can affect the transmission of tubercolosis because it relates to personal hygiene, home cleaningness, and other infectious diseases and immune system. Therefore, the study aims to explore the effect of water and sanitation on Tuberculosis Incidence Rate in regions, which are major contributors on tuberculosis cases globally namely South-East Asia and Africa. The study also include Human Development Index in the analysis as a control variable.

## 2.2 Description of data and data source

The study will use multi sources of data including World Bank, World Health Organization (WHO), United Nation for Children Fund (Unicef). Different dataset contains different variable, however all dataset have same year, country name and country codes that can be used to combine the dataset. The data also contain information of countries globally from 2000 to 2020. The variables will include percentage of population using safely managed sanitation services, percentage of population using at least basic drinking water service, percentage of population using well-managed water, percentage of population using basic sanitation, percentage of population using well-managed sanitation, and human development index.

## 2.3 Questions/Hypotheses to be addressed

*State the research questions you plan to answer with this analysis.*

The main aim of this study is to look at the effect of Water and Sanitation on on tuberculosis incidence.

RQ: Does access to water and sanitation effect Tuberculosis incidence in SOuth-East Asia and African Regions?

Null Hypotheisis

H0: thee is no effect of access to basic sanitation on tuberculosis incidence

H0: thee is no effect of access to safely managed sanitation and on tuberculosis incidence

H0: there is no effect of access to access basic drinking water and tuberculosis incidence

H0: there is no effect of access to safely managed drinking water on tuberculosis incidence

Alternative Hypothesis Ha: thee is an effect of access to basic sanitation on tuberculosis incidence

Ha: thee is an effect of access to safely managed sanitation and on tuberculosis incidence

Ha: there is an effect of access to access basic drinking water and tuberculosis incidence

Ha: there is an effect of access to safely managed drinking water on tuberculosis incidence

# 3. Methods

The datasets were loaded one by one, then drop all unused variable before combined those datasets using reduce() function. Countries, which have at least one missing years from 2000 to 2020, were dropped from the study. To make sure that the data is clean and ready to analyse, all missing values were dropped. This study focuses on South East Asia and Africa Regions, where the major proportion of tuberculosis cases came. Therefore, a newsubset for South East Asia and African countries was created.

After data cleaning, multiple exploratory data analysis was performed such us summary and visualizing the data. A bloxplot was created to show Tuberculosis Incidence in each country during the time frame. Spaghetti plot was created to visualize the trend of tuberculosis incidence and other variables over the time. Lastly, line charts showing the trend of all variables in each country.

The exploratory data analysis showed that the data were not normally distributed and there was no linear relationship between tuberculosis incidence and other predictors. Therefore, Spearman’s correlation matrix was performed to see the correlations. To have better understanding, a heat map matrix plot was created to show the correlation.

The main statistical analysis in the study is longitudinal data analysis (either fixed effect or mixed effect method). To see which one is better, Hausman Test was performed to determine which model is better.

NOTE: I still need to discuss and need more advice about the data analysis from Dr. Handel

## 3.1 Data acquisition

The data was downloaded from Worlbank Dataset. Those data are originally from many different courses/ organizations, such as World Bank, UNICEF, and UN-HABITAT

## 3.2 Data import and cleaning

Data Import and data cleaning process have been explained in the method part. To see the code for data import and data cleaning , please visit working directory (“code”, “processing-code”, “processingfile.qdm”). Processed and clean data are saved at a folder (“data”, “processed-data”). There are many .rds files in the folder indulging file for each variables, and file after combined all variables.

Dataset are either in excel or CSV files. Several steps from loading to data analysis:

* to load the data set in this part using read\_excel or read\_csv.
* to combine them base on country name/ code and years.
* to drop unwanted variables
* to select Asian and African Countries only
* to drop N/A or missing value
* to conduct data exploration
* to decide suitable analysis method
* to perform data analysis

## 3.3 Statistical analysis

*Explain anything related to your statistical analyses.*

The study applied Generalized Additive Mixed Model (GAMM). The model can capture mixed effect of countries on Tuberculosis Incidence with non linear relationship. The data analysis started by looking at the relationship between Tuberculosis Incidence and Predictors. Non linear relationship was discovered. Human Development Index (HDI), Access to basic water and well-managed water show curvature relationship, while access to basic sanitation and well-managed sanitation show polynomial. There are several models fitted in the study, we start with linear model with all predictors included, followed by Generalized Addictive Models (GAM), Generalized Mixed Addictive Models (GAMM) without interaction, Generalized Mixed Addictive Models (GAMM) with two factor interaction, and Generalized Mixed Addictive Models (GAMM) with two factor interaction after drop interaction which are not significant. After fitting the models, RMSE and AIC were printed to chose the best models. Based on the RMSEs and AICs, GAMM with interaction after deleting unsigniuficant interactions performed best. The RMSE and AIC were visulized using plot for better understanding.

To maximize test the performance of the model, machine learning was performed. We created splitted data with 75% train data and 25 % test data and setseed (2345). RMSEs and AICs were printed to see the performance of the models and plotted. The final model was fitted into a machine learning, XGBoost. The method is very plexible and can deal with the data conditions (non-linear relationship, some multicolinearity, and binomial distribution). Finally, the model was tunned.

# 4. Results

## 4.1 Exploratory/Descriptive analysis

Firs of all, the pattern of the data was explored. The trend of Tuberculosis Incidence rate for each country during the period was assesed using spagethi plots. The ( **?@fig-1** ) shows that most countries shows a dicrease of tuberculosis incidence over the period of time.

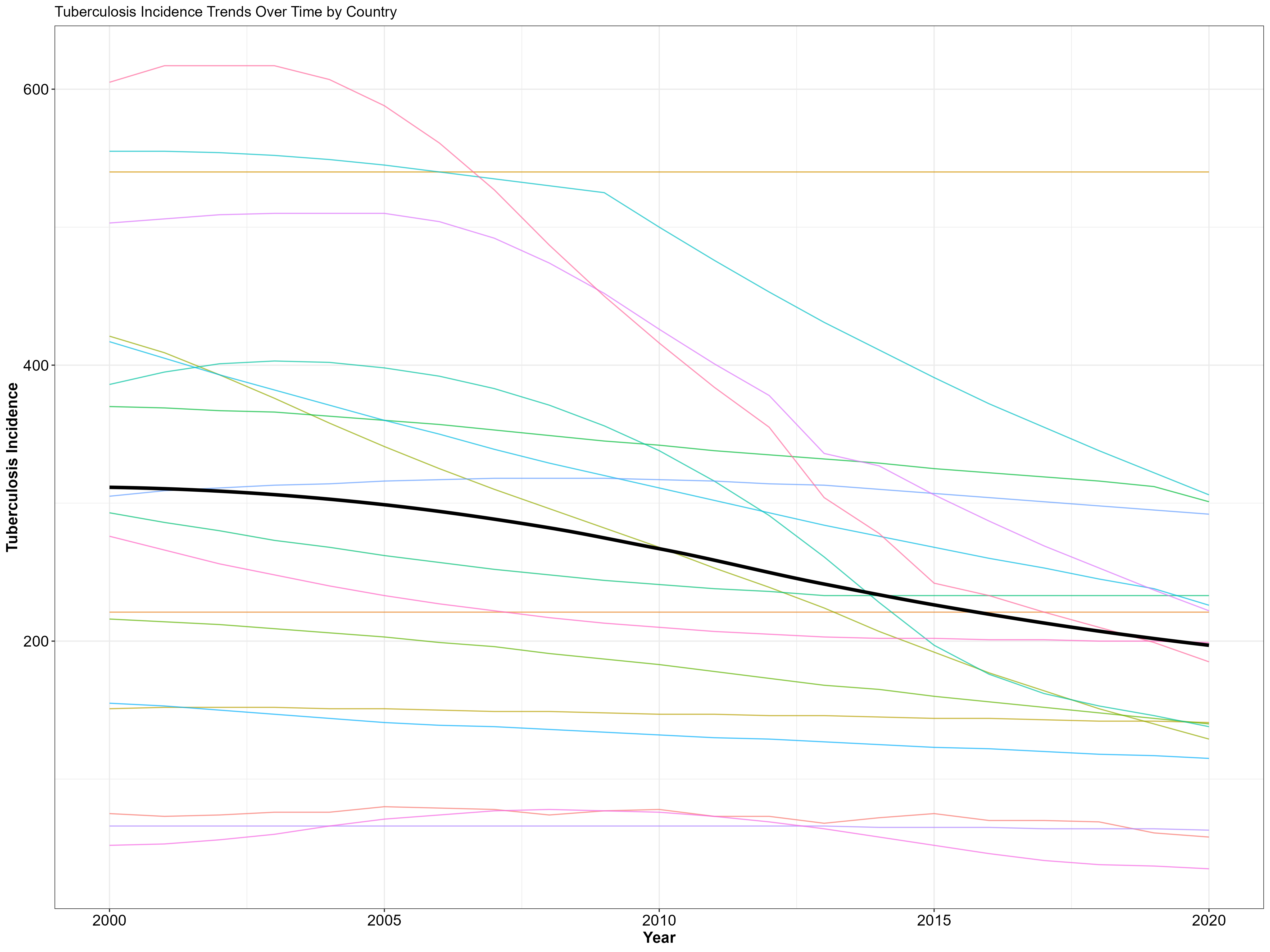
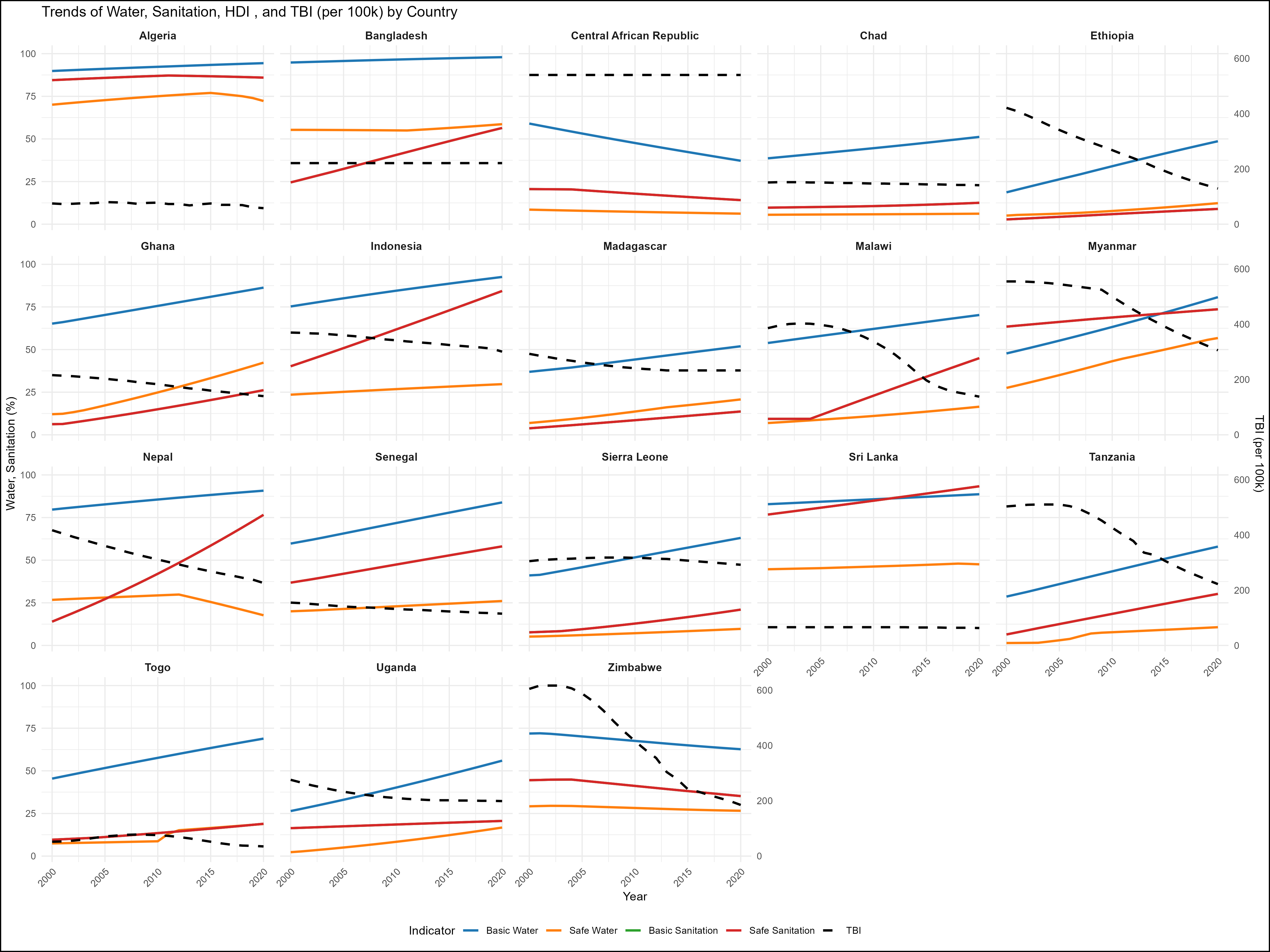
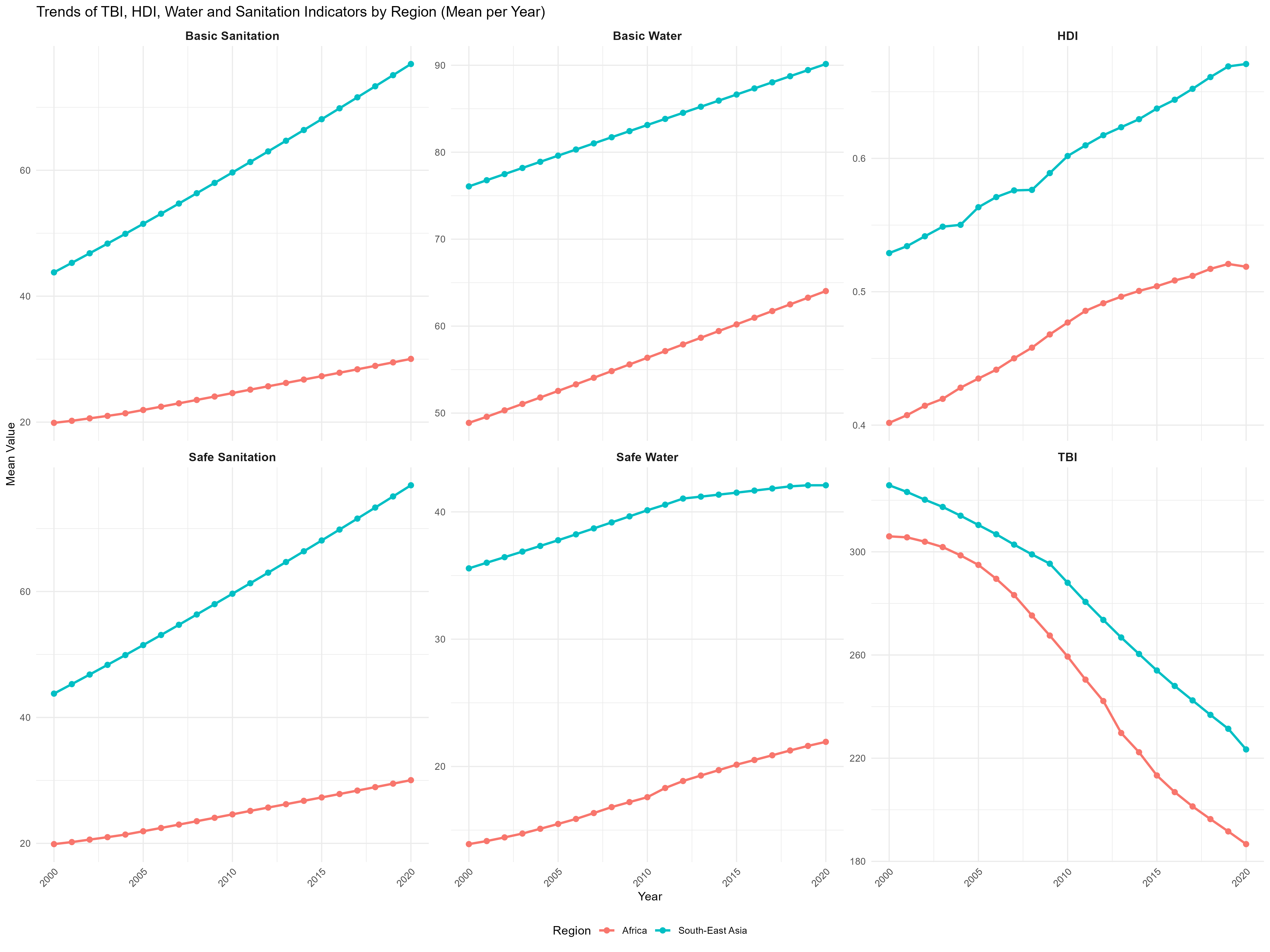


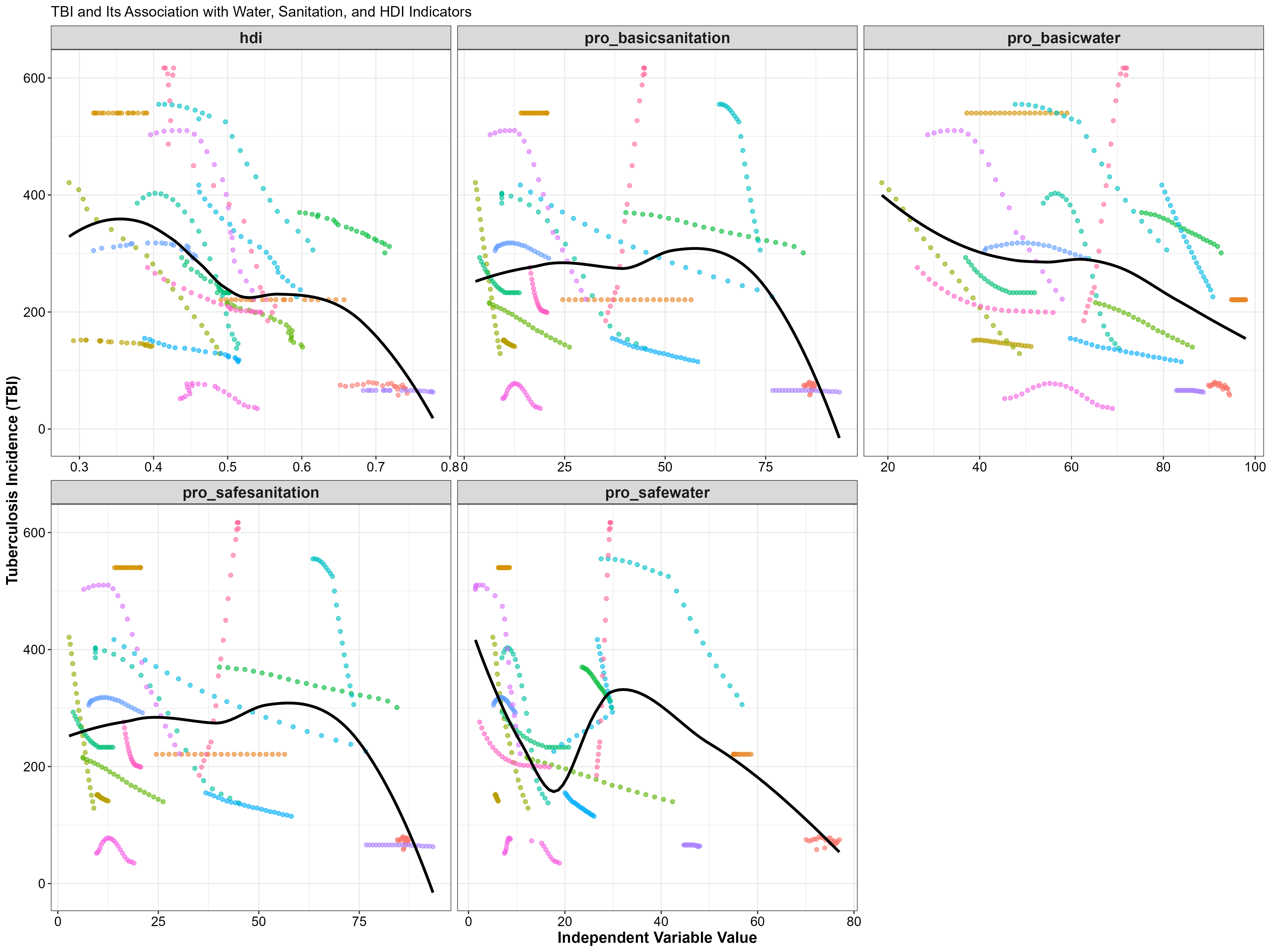
Figure ( **?@fig-2** ) shows the trend of percentage of population using safe water, basic water, safe sanitation, basic sanitation, tuberculosis incidence, and human development index over the prerion for each country. It can be seen that the percentage of population having access to water and sanitation and human development index increase over the time, while the tuberculosis incidence decrease. Figure ( **?@fig-3** ) show the trends of all variables based on region.

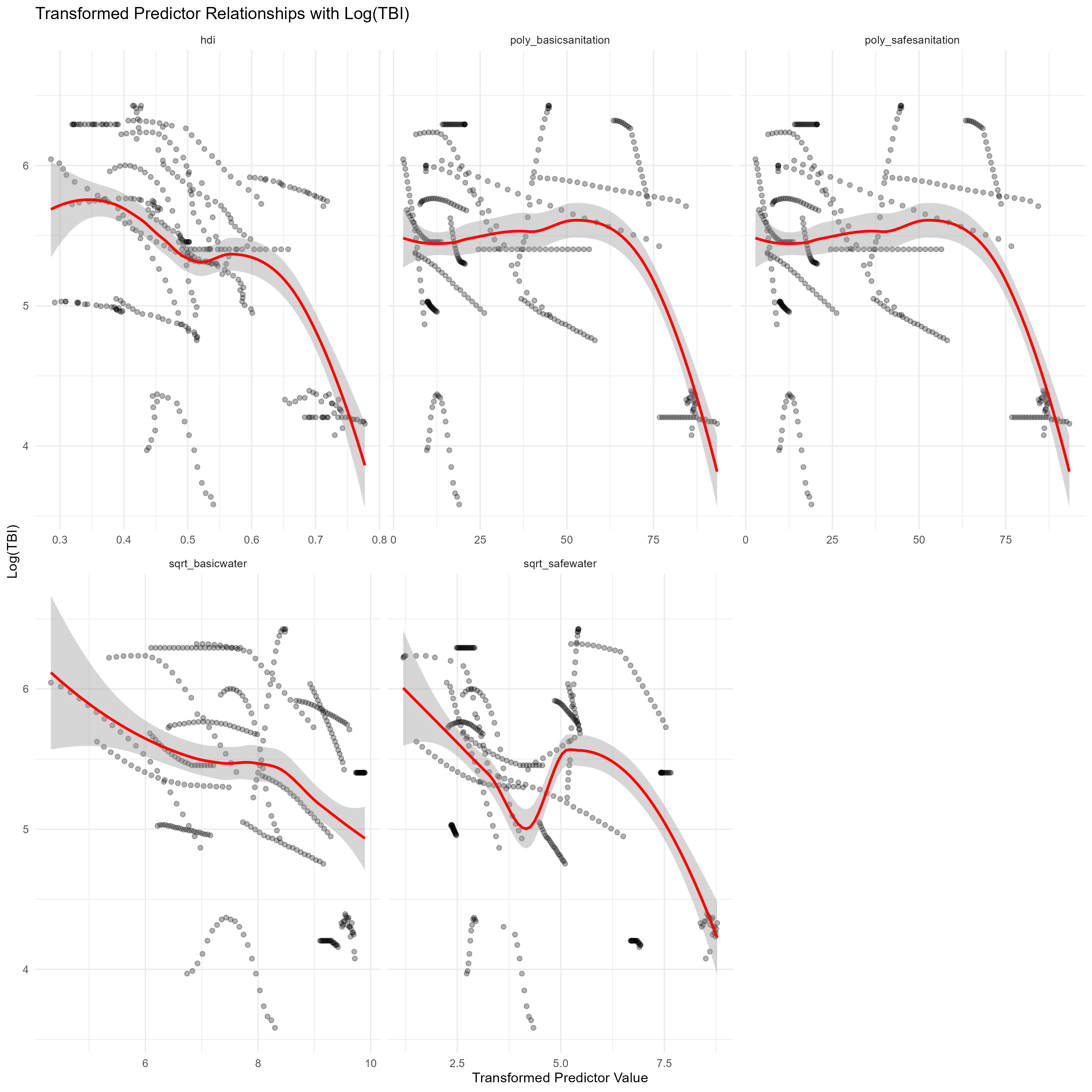




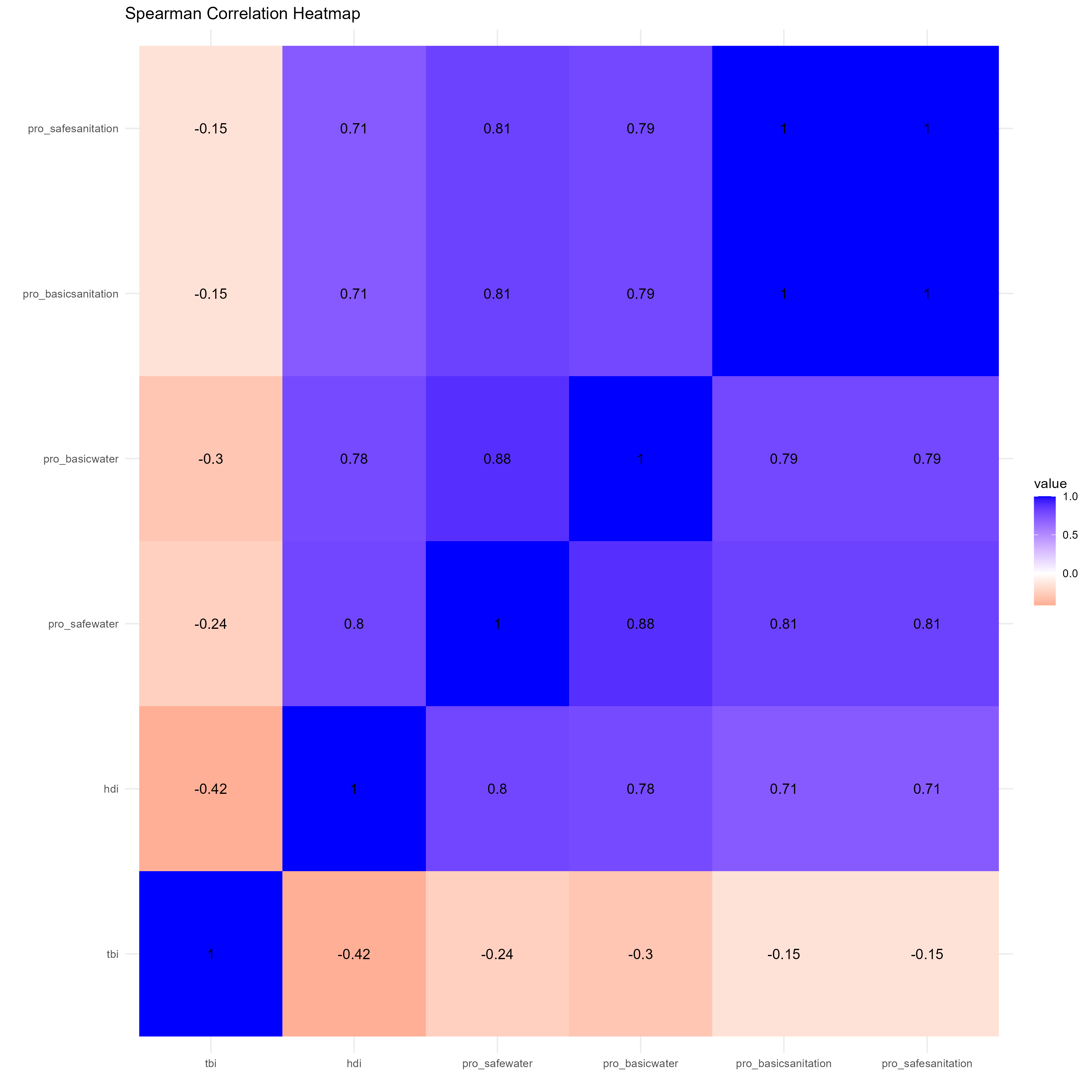
## 4.2 Basic statistical analysis

As first step of statistical analysis. linearity test was performed, the result shows that there is non-linear relationship between tuberculosis incidence rate and other predictors. Loess method is used to show the non-linear relationship betwen tuberculosis incidence rate and predictors. ( **?@fig-4** ) shows the relationship between tuberculosis incidence and predictors and ( **?@fig-5** ) show the relationhsip after data transformation. It can be seen that after data transformed, non-linear relationship stil exists. Therefore, Generalized Addittive Mixed Model is used in this study.



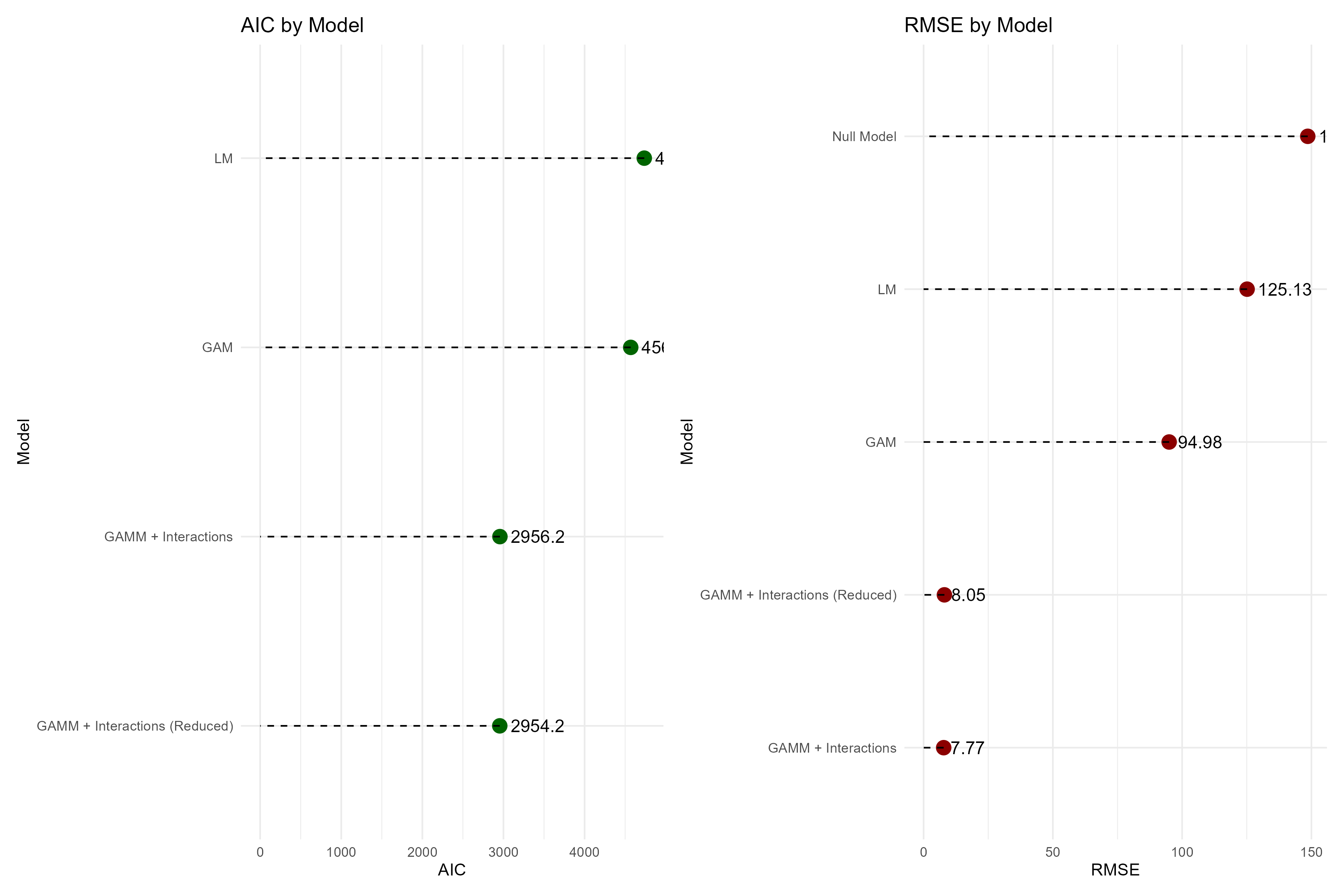


Before going further, we need to look at correlation between tuberculosis Incidence and all predictors. Since they show non linearity, spearman’s correlation performed and showing in the following correlation matrix. See ( **?@fig-6** ) shows that some variables have multi collinearity. The matrix shows that water and sanitation have low correlation with tuberculosis incidence.



## 4.3 Full analysis

We start with linear model with all predictors included, followed by Generalized Addictive Models (GAM), Generalized Mixed Addictive Models (GAMM) without interaction, Generalized Mixed Addictive Models (GAMM) with two factor interaction, and Generalized Mixed Addictive Models (GAMM) with two factor interaction after drop interaction which are not significant. AIC and RMSE are used to select the best model. ( **?@fig-7**) show that and Generalized Mixed Addictive Models (GAMM) with two factor interaction after dropin insignificant interaction performed better. GAMM with full interactions has low AIC and RMSE, but the degree of freedom shows negarive values, indicating inconsistency of the model.



***Result of Final Model***

We fitted a generalized additive mixed model (GAMM) using a negative binomial distribution with a log link to model tuberculosis incidence (TBI) rate across countries in South East Asia and Africa. The model incorporated smooth functions of year, access to basic and safe water and sanitation services, and the Human Development Index (HDI), as well as interaction terms between key environmental and socioeconomic indicators. A random effect was included for country to account for unobserved heterogeneity. (**?@tbl-1** ) shows the result of the final model.

Table 1: Summary of GAMM Estimates

Parametric and Smooth Terms

| Term | Estimate | Std. Error | EDF | Chi-square | p-value |
| --- | --- | --- | --- | --- | --- |
| Parametric | | | | | |
| (Intercept) | 5.549 | 0.185 | - | - | 0.000 |
| Smooth | | | | | |
| s(year) | - | - | 3.380 | 52.775 | 0.000 |
| s(pro\_basicwater) | - | - | 6.518 | 18.226 | 0.013 |
| s(pro\_safewater) | - | - | 1.000 | 15.868 | 0.000 |
| s(pro\_basicsanitation) | - | - | 2.674 | 2.572 | 0.462 |
| s(pro\_safesanitation) | - | - | 3.674 | 3.283 | 0.598 |
| s(hdi) | - | - | 1.001 | 23.359 | 0.000 |
| ti(pro\_basicwater,hdi) | - | - | 7.522 | 37.385 | 0.000 |
| ti(pro\_safewater,hdi) | - | - | 8.001 | 59.489 | 0.000 |
| ti(pro\_safesanitation,pro\_basicwater) | - | - | 3.386 | 55.697 | 0.000 |
| ti(pro\_safewater,pro\_basicwater) | - | - | 5.138 | 43.652 | 0.000 |
| s(country) | - | - | 16.061 | 5,340.721 | 0.000 |

The model demonstrated excellent fit, explaining 99.7% of the deviance (adjusted R² = 0.997). Among the smooth terms, year (edf = 3.38, p < 0.001), access to basic water (edf = 6.52, p = 0.013), access to safe water (edf = 1.00, p < 0.001), and HDI (edf = 1.00, p < 0.001) showed significant non-linear associations with TBI. In contrast, basic and safe sanitation services were not significantly associated with TBI rate after accounting for other variables.

Notably, several two-dimensional interaction smooths were also statistically significant, including between basic water and HDI (p < 0.001), safe water and HDI (p < 0.001), safe sanitation and basic water (p < 0.001), and safe water and basic water (p < 0.001), suggesting complex joint effects of environmental infrastructure and development on TBI patterns. The random effect for country was highly significant, capturing substantial variation across national contexts.

These findings underscore the importance of improving access to clean water and addressing socioeconomic disparities to reduce tuberculosis burden, while also highlighting the complex interplay between infrastructure and development indicators.

### 4.3.1 Machine Learning

To test the consistency of the model, machine learning was used.

Using Splitted / generated data shows that the model has a consistent result. ([Figure 1](#fig-8) ) shows the observed and predicted values both original and generated data. It can be seen that both original data and generated data (both test and training data) have similiar pattern and the plots lay very closed or around the line.

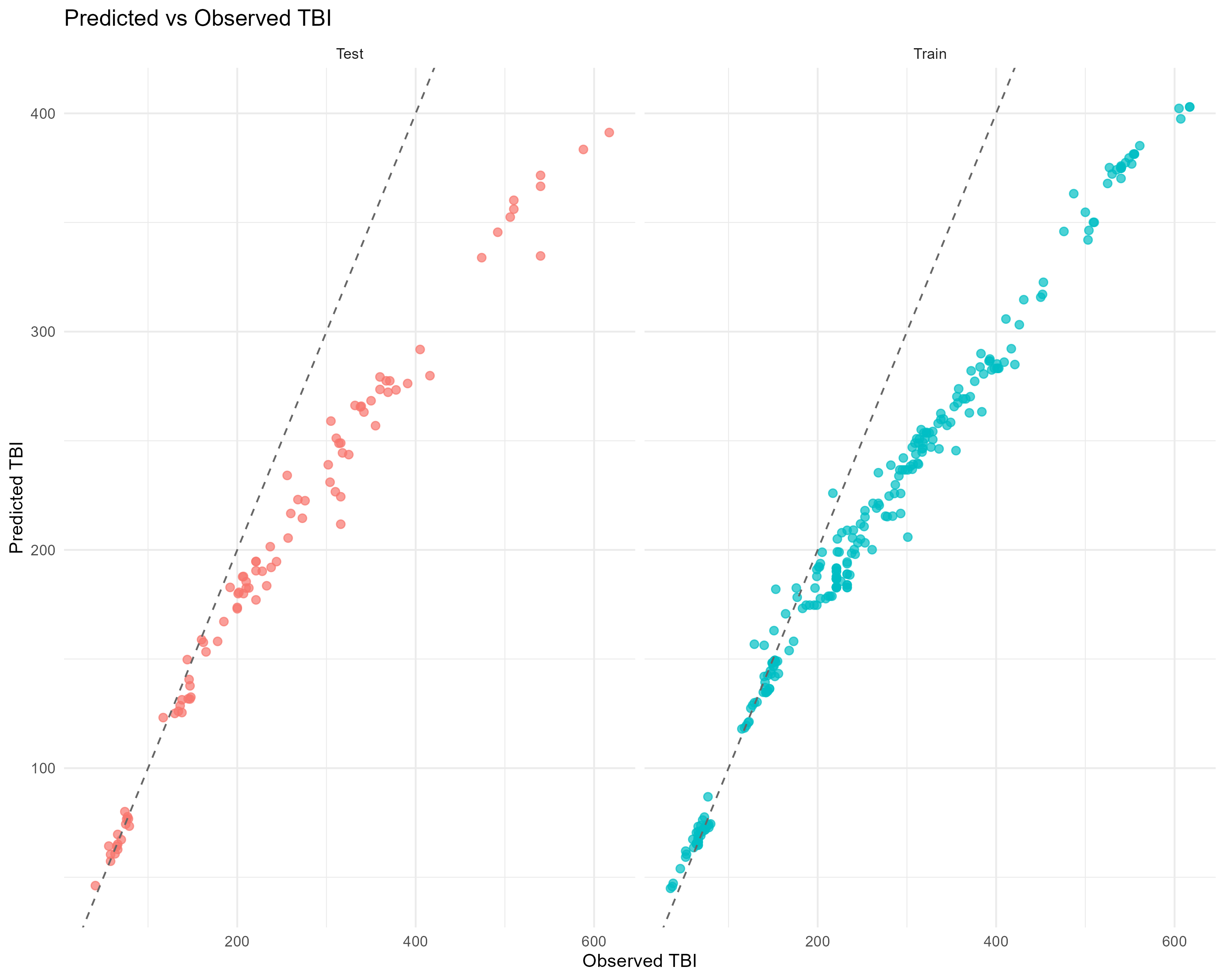
|  |
| --- |
| Figure 1: Predicted vs Observed TBI originalvs generated data |

Figure (**?@fig-9**) shows the performance of machine learning (XGboost). The figure shows how well the model predicted tuberculosis incidence (TBI) by comparing the predicted values with the actual reported values for both the training and test data. The dotted line represents perfect prediction—where predicted and actual values would match exactly.

For the training data (right side), the dots are closely lined up along the diagonal line, which means the model predicted the TBI values very accurately for the data it was trained on.

For the test data (left side), most of the points are still close to the line, but some are spread out more, especially at higher TBI values. This means that while the model still performs well on new data, it slightly underestimates TBI in some cases when the actual rates are very high.

Overall, the plot shows that the model is reliable and provides accurate predictions for TBI, even on data it hasn’t seen before. This suggests the model can be useful for forecasting TB burden in similar settings.



***Tunning for XGBoost***

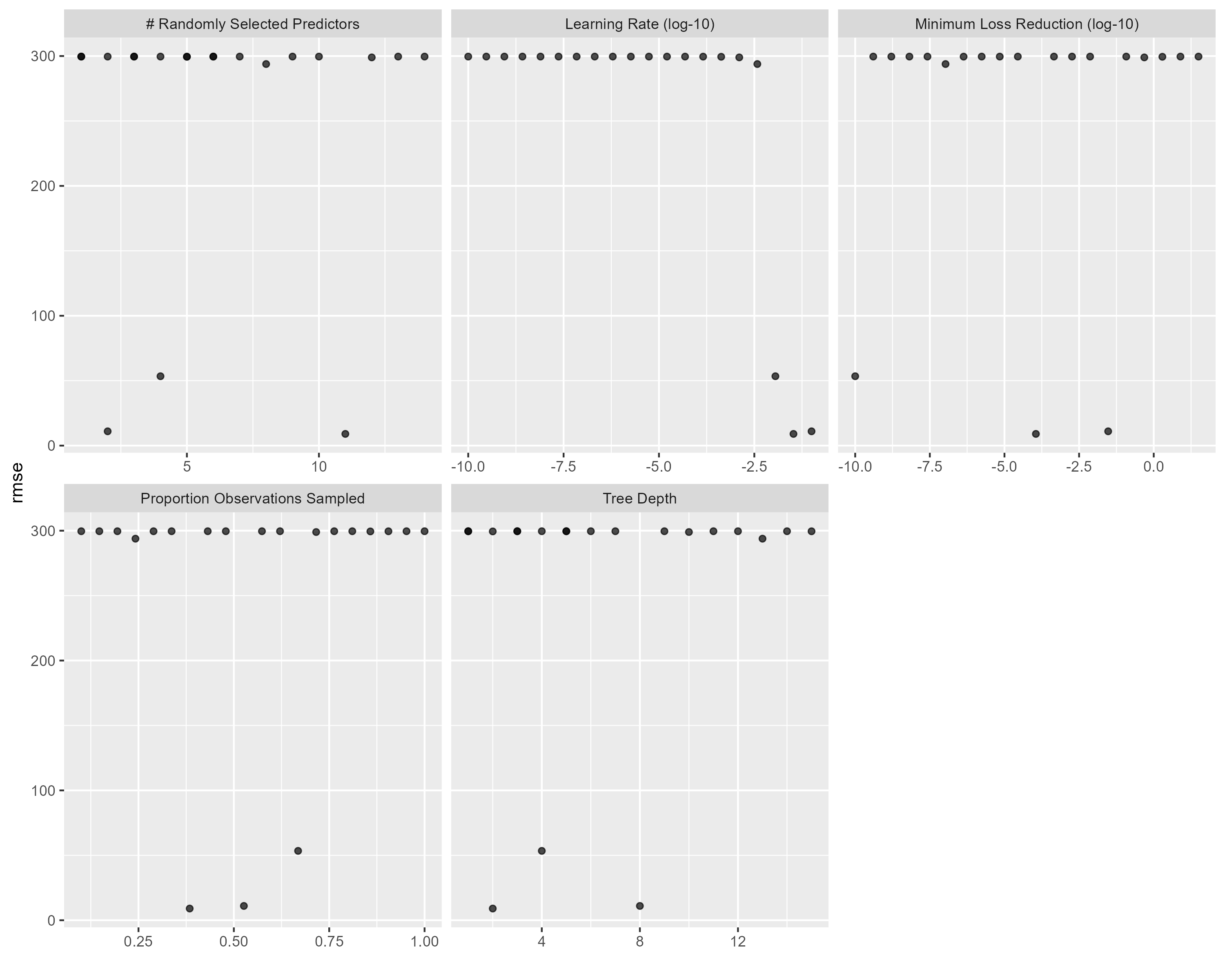


Figure (**?@fig-10**) shows the results of hyperparameter tuning for the XGBoost model, showing how different parameter values affected the model’s performance, measured by root mean square error (RMSE). Each dot represents one combination of hyperparameters tested during cross-validation. These results highlight that careful tuning of parameters, especially learning rate, tree depth, and sampling rate, is essential to significantly improve model accuracy. The tuning process allowed us to identify a combination of settings that minimized prediction error and led to a more reliable model.

# 5. Discussion

## 5.1 Summary and Interpretation

*Summarize what you did, what you found and what it means.*

This study highlights the significant impact of environmental infrastructure and socioeconomic development on tuberculosis incidence (TBI) in South East Asia and Africa. The use of a generalized additive mixed model (GAMM) with a negative binomial distribution and log link allowed us to flexibly assess both linear and non-linear associations between predictors and TBI rates, accounting for random variation between countries. Our model demonstrated excellent fit, explaining 99.7% of the deviance, and identified multiple significant associations between water access, development indices, and TBI incidence, consistent with and extending prior literature.

The significant non-linear effect of year on TBI incidence underscores the ongoing temporal trends in tuberculosis control, shaped by global health policies and national-level interventions over time. Previous studies have documented a steady decline in TB incidence rates globally, albeit at a slower rate than needed to meet elimination targets, particularly in high-burden settings such as sub-Saharan Africa and Southeast Asia (11,12). Our findings align with these trends and further demonstrate the utility of flexible modeling approaches in capturing such non-linear patterns over extended time frames.

The significant associations of access to basic and safe water with TBI reinforce the importance of environmental determinants in TB control. Limited access to clean water can contribute to a cycle of poor hygiene, malnutrition, and vulnerability to infectious diseases, including TB (13,14). Our identification of complex non-linear relationships suggests that the effects of water access on TB are not constant but vary depending on context, infrastructure level, and interaction with other factors.

Moreover, the Human Development Index (HDI) was significantly associated with TBI rates in a non-linear fashion. This is consistent with research indicating that higher levels of human development—including education, life expectancy, and income—are linked to reduced susceptibility to TB through better living conditions, nutrition, and access to healthcare services (15,16). These findings reaffirm the critical role of social determinants in TB epidemiology and provide quantitative support for global policy frameworks that integrate development with infectious disease control.

Notably, basic and safe sanitation access did not show a significant independent association with TB in our adjusted model. While sanitation is undeniably important for overall public health, its isolated effect on TB incidence may be attenuated once other covariates such as water access and HDI are considered. This finding echoes studies showing that while sanitation is crucial for preventing diseases like diarrhea and helminth infections, its direct effect on TB—which is primarily airborne—may be more indirect and mediated by broader health and development pathways (17,18).

The interaction terms in our model revealed statistically significant joint effects between several pairs of variables, suggesting that infrastructure and development indicators do not act independently. For instance, the interaction between basic water and HDI supports prior findings that the protective effects of environmental infrastructure are magnified in settings with higher socioeconomic development (19,20). Likewise, the interaction between safe water and HDI emphasizes the synergistic value of combining access to safe water with broader improvements in human development.

Interactions between safe sanitation and basic water, as well as between safe water and basic water, also showed significant effects on TBI, implying that layered improvements in environmental infrastructure produce more than additive benefits. These patterns align with ecological frameworks emphasizing cumulative and contextual effects in public health interventions (21,22).

Importantly, the random effects for country were highly significant, indicating unmeasured country-level heterogeneity in TB patterns. These may include factors such as political stability, health system capacity, HIV prevalence, urbanization patterns, or population density—all known modifiers of TB burden (23,24).

Our findings have several policy implications. First, they highlight the need for multisectoral strategies that address both environmental conditions and broader socioeconomic development to combat TB effectively. Investments in water infrastructure should be accompanied by programs targeting education, nutrition, housing, and healthcare access (25). Second, the complex interactions observed suggest that context-specific interventions are essential. A one-size-fits-all approach may fail to capture the unique synergy between environmental and developmental factors within each country.

## 5.2 Strengths and Limitations

The strength of our study lies in the use of a flexible GAMM framework that accommodates non-linear relationships and interaction effects, offering a nuanced view of TB determinants. By including a random effect for country, we account for latent national characteristics that may bias fixed-effect estimates. However, our study has limitations. The ecological nature of the data precludes individual-level inferences, and potential measurement error in national statistics could affect the reliability of some estimates.

## 5.3 Conclusions

This study reinforces that tuberculosis incidence in South East Asia and Africa is not solely a biomedical issue but is deeply influenced by environmental infrastructure and socioeconomic development. Our findings demonstrate that access to clean water and improvements in human development significantly reduce TB burden, with their effects magnified when considered jointly. The non-linear and interactive associations revealed by our model highlight the importance of tailored, multisectoral approaches that integrate public health with infrastructure and development planning. To achieve meaningful and sustained reductions in TB, especially in high-burden regions, global health strategies must move beyond treatment alone and address the root structural determinants of disease.

# 6. References

1. National tuberculosis prevalence surveys 2007-2016. Genève, Switzerland: World Health Organization; 2021.

2. Tosepu R, Sani A, Effendy DS, et al. The association between climate variables and tuberculosis in kolaka district, southeast sulawesi province, indonesia, 2013–2020: A bayesian autoregressive model. *F1000Research* [electronic article]. 2024;12:1507. (<http://dx.doi.org/10.12688/f1000research.138859.2>)

3. Andersen ZJ, Vicedo-Cabrera AM, Hoffmann B, et al. Climate change and respiratory disease: Clinical guidance for healthcare professionals. *Breathe* [electronic article]. 2023;19(2):220222. (<http://dx.doi.org/10.1183/20734735.0222-2022>)

4. Uwishema O, Masunga DS, Naisikye KM, et al. Impacts of environmental and climatic changes on future infectious diseases. *International Journal of Surgery* [electronic article]. 2023;109(2):167–170. (<http://dx.doi.org/10.1097/JS9.0000000000000160>)

5. El-Sayed A, Kamel M. Climatic changes and their role in emergence and re-emergence of diseases. *Environmental Science and Pollution Research* [electronic article]. 2020;27(18):22336–22352. (<http://dx.doi.org/10.1007/s11356-020-08896-w>)

6. Kharwadkar S, Attanayake V, Duncan J, et al. The impact of climate change on the risk factors for tuberculosis: A systematic review. *Environmental Research* [electronic article]. 2022;212:113436. (<http://dx.doi.org/10.1016/j.envres.2022.113436>)

7. Hong C, Zhong R, Xu M, et al. Interactions among food systems, climate change, and air pollution: A review. *Engineering* [electronic article]. 2025;44:215–233. (<http://dx.doi.org/10.1016/j.eng.2024.12.021>)

8. Mirzabaev A, Bezner Kerr R, Hasegawa T, et al. Severe climate change risks to food security and nutrition. *Climate Risk Management* [electronic article]. 2023;39:100473. (<http://dx.doi.org/10.1016/j.crm.2022.100473>)

9. Microbes and climate change - science, people &amp; impacts. American Society for Microbiology; 2022.(<http://dx.doi.org/10.1128/AAMCol.Nov.2021>)

10. Cai Y, Zhao S, Niu Y, et al. Modelling the effects of the contaminated environments on tuberculosis in jiangsu, china. *Journal of Theoretical Biology* [electronic article]. 2021;508:110453. (<http://dx.doi.org/10.1016/j.jtbi.2020.110453>)

11. Pai M, Behr MA, Dowdy D, et al. Tuberculosis. *Nature Reviews Disease Primers*. 2016;2:16076.

12. Organization WH. Global tuberculosis report 2023. World Health Organization; 2023.(<https://iris.who.int/handle/10665/373828>)

13. Esrey SA, Potash JB, Roberts L, et al. Effects of improved water supply and sanitation on ascariasis, diarrhoea, dracunculiasis, hookworm infection, schistosomiasis, and trachoma. *Bulletin of the World Health Organization*. 1991;69(5):609.

14. Pruitt S, Schwab KJ. Water, sanitation, and hygiene (WASH) and the risk of respiratory infections in children under 5 years old in low- and middle-income countries: A systematic review and meta-analysis. *International Journal of Environmental Research and Public Health*. 2021;18(3):1243.

15. Siroka A, Law I, Macinko J, et al. Assessing the impact of social determinants on tuberculosis: A global review. *European Respiratory Journal*. 2016;48(4):1299–1313.

16. Wilkinson RG. Income inequality and social cohesion. *American Journal of Public Health*. 1999;89(9):1321–1322.

17. Bartram J, Cairncross S. Hygiene, sanitation, and water: Forgotten foundations of health. *PLoS Medicine*. 2010;7(11):e1000367.

18. Prüss A, Kay D, Fewtrell L, et al. Estimating the burden of disease from water, sanitation, and hygiene at a global level. *Environmental Health Perspectives*. 2002;110(5):537–542.

19. Lienhardt C, Fielding K, Sillah JS, et al. Risk factors for tuberculosis infection in sub-saharan africa: A contact study in the gambia. *American Journal of Respiratory and Critical Care Medicine*. 2001;163(7):1324–1330.

20. Matos ED, Moreira ED, Alencar AP, et al. Tuberculosis and socioeconomic indicators: Systematic review of the literature. *Revista Panamericana de Salud Pública*. 2013;33(5):356–361.

21. Marmot M, Friel S, Bell R, et al. Closing the gap in a generation: Health equity through action on the social determinants of health. *The Lancet*. 2008;372(9650):1661–1669.

22. Aagaard-Hansen J, Chaignat C-L. Integrated approaches to health and development: The role of water, sanitation and hygiene. *Tropical Medicine & International Health*. 2014;19(9):1116–1120.

23. Gupta RK, Lucas SB, Fielding KL, et al. Addressing the social determinants of tuberculosis: A review of the evidence and policy options. *The Lancet*. 2015;386(10010):964–972.

24. Reid A, Gupta R, Ditiu L, et al. Building a tuberculosis-free world: The lancet commission on tuberculosis. *The Lancet*. 2019;393(10178):1331–1384.

25. Onozaki I, Law I, Sismanidis C, et al. Impact of investment in tuberculosis control on economic growth in countries with high tuberculosis burden. *Tropical Medicine & International Health*. 2015;20(1):29–37.