Oregon State University

Modeling Female Sex Worker Distribution And Determinants In Sub Saharan Africa To Enhance HIV Prevention Strategies

---

Author

Cyril Nana Boakye Benson

1 INTRODUCTION

HIV continues to be a significant public health challenge in sub-Saharan Africa, which bears a disproportionate burden of the global epidemic. Given the critical role of Female Sex Workers (FSWs) in HIV transmission dynamics, understanding their distribution is essential for the targeted implementation of HIV-prevention services. This study investigates the abundance of FSWs within sub-Saharan Africa by examining various factors that may influence their distribution. The analysis considers variables such as the country (`country`: Mozambique, Malawi, Zimbabwe, Zambia, and Botswana) and the region (`region`) within each country, along with the year (`dataYear`) and month (`month`) the data were collected. Other important factors include the number of Female Sex Workers observed (`FSWCount`), population density within the surveyed transect (`popDensity`), and the built-up index (`built`), which reflects urbanization levels. Environmental variables such as the average length of the growing season (`growingSeason`), average annual rainfall (`rain`), and average annual temperature (`temperature`) are included in the analysis. Additionally, the study evaluates night-time light activity (`nightLight`), distance to the nearest clean water source (`cleanWater`), and proximity to the nearest protected area (`protected`). Health-related factors include the HIV prevalence rate among the population (`hivRate`), average fear of contracting HIV (`hivFear`), and the percentage of the population using insect nets (`insectNet`). Sociological dimensions, such as the average age of first sexual experience (`ageFirstSex`), wealth index relative to the sub-Saharan African region (`wealthIndex`), and the size of the surveyed transect (`surveyArea`), are also considered, along with an indicator of whether the crime rate is high in the transect (`highCrime`).By assessing the influence of these variables, this study aims to provide critical insights into the factors shaping the distribution of FSWs in sub-Saharan Africa. These findings will guide the prioritization of environmental and sociological factors in future data collection efforts and the implementation of HIV-prevention strategies.

2 EXPLORATORY DATA ANALYSIS

The dataset comprises 750 observations, of which 675 are retained for analysis in this study. The response variable, `FSWCount`, represents the number of female sex workers recorded and is a discrete, non-negative integer. Summary statistics, presented in Table 1, were generated using the `summary` function in R (R Core Team, 2020). The recorded counts of female sex workers range from a minimum of 0 to a maximum of 134, with an average count of approximately 5.564. A histogram, displayed in Figure 1 and created using the `ggplot2` package (Wickham, 2019), illustrates the distribution of female sex worker counts. Correlation analysis, as shown in Table 2, reveals that the variables `built`, `growingSeason`, `rain`, `nightLight`, and `cleanWater` exhibit weak negative correlations with `FSWCount`. Conversely, `temperature` shows a positive correlation with `FSWCount`. To further explore these relationships, a scatterplot and correlation matrix are presented in Figures 2 and 3, created using the `GGally` package (Schloerke, 2024). The scatterplot/correlation matrix highlights that most observations cluster around zero counts for variables such as `built`, `growingSeason`, `rain`, `nightLight`, `cleanWater`, and `wealthIndex`. In contrast, a positive correlation was observed between temperature (`temperature`) and FSW counts. These findings provide an overview of the relationships between the response variable and key predictors in the dataset.

3 STATISTICAL PROCEDURE

The model incorporates `country`, `region`, `dataYear`, and `month` as random effects to account for unobserved heterogeneity and variability attributable to spatial and temporal factors. The variable `surveyArea` is specified as an offset term to normalize the response variable, `FSWCount`, by the size of the surveyed regions, ensuring that larger areas do not disproportionately influence the model outcomes. Given the discrete and non-negative integer nature of `FSWCount`, count regression mixed-effects models are deemed appropriate. These models are estimated using the `glmmTMB` function from the `glmmTMB` package (Mollie, 2017), allowing for flexible handling of fixed and random effects while accommodating the characteristics of count data.

4 DISTRIBUTION SELECTION USING ENVIROMENTAL VARIABLES

An initial Poisson mixed-effects model was developed to analyze `FSWCount` as the response variable. The model incorporated `country`, `region`, `dataYear`, and `month` as random effects to account for spatial and temporal variability, along with environmental variables as fixed covariates. Diagnostic evaluations, presented in Figure 4, identified potential issues with overdispersion. Specifically, the Residuals vs. Fitted plot (Figure 4) indicated an increasing spread of residuals with higher fitted values, consistent with overdispersion. Additionally, the QQ plot (Figure 4) revealed

deviations from the normality assumption, with residuals exhibiting heavy-tailed behavior. A formal overdispersion test confirmed the presence of overdispersion, with a dispersion parameter ($\theta = 6.977929$) exceeding what would be expected under a Poisson distribution. To address these challenges, alternative mixed-effects models were considered, including the negative binomial, zero-inflated Poisson, zero-inflated negative binomial, hurdle Poisson, and hurdle negative binomial models. Model selection was guided by the Akaike Information Criterion (AIC), as shown in Table 3. The negative binomial mixed-effects model emerged as the best-performing model, with the lowest AIC value (2873.769), indicating a superior balance between model fit and complexity. Given the observed proportion of zero counts in `FSWCount`, a zero-inflated negative binomial model was also evaluated due to its comparable AIC score. However, upon fitting, the estimated proportion of excess zeros was found to be negligible, suggesting that a zero-inflated structure was unnecessary as the data did not exhibit significant zero inflation. The negative binomial mixed-effects model was therefore deemed the most appropriate for modeling the response variable.

5   DISTRIBUTION SELECTION USING SOCIOLOGICAL VARIABLES

To identify the appropriate distribution using the sociological variables, a Poisson mixed-effects model was fitted with `FSWCount` as the response variable. The model included `country`, `region`, `dataYear`, and `month` as random effects to account for spatial and temporal variability, while all other sociological variables were included as fixed covariates. Diagnostic plots, presented in Figure 5, revealed evidence of overdispersion. Specifically, the Residuals vs. Fitted plot (Figure 5) showed increasing residual variability with higher fitted values, suggesting overdispersion. The QQ plot (Figure 5) provided moderate to strong evidence against the normality assumption, as residuals exhibited heavy-tailed behavior. A formal overdispersion test further confirmed this, with a dispersion parameter ($\theta = 6.324017$) exceeding what would be expected under a Poisson distribution. To address the overdispersion, several alternative mixed-effects models were considered, including negative binomial, zero-inflated Poisson, zero-inflated negative binomial, hurdle Poisson, and hurdle negative binomial models. Model selection was guided by the Akaike Information Criterion (AIC), as shown in Table 4. Among these, the negative binomial mixed-effects model achieved the lowest AIC score (2872.943), indicating the best balance between model fit and information loss. Given the high proportion of zero counts in `FSWCount`, the zero-inflated negative binomial model was also evaluated due to its comparable AIC score. However, after fitting, the estimated proportion of excess zeros was found to be negligible, indicating that the zero-inflated structure did not significantly improve the model fit. Consequently, the negative binomial mixed-effects model was deemed the most appropriate for modeling the response variable.

6   VARIABLE SELECTION USING ENVIRONMENTAL VARIABLES

Using the negative binomial mixed-effects model, the next step is to identify the variables to include in the final model. A preliminary model is fitted with `country`, `region`, `dataYear`, and `month` as random effects, and `popDensity`, `built`, `growingSeason`, `rain`, `temperature`, `nightlight`, and `cleanWater` as covariates, with `surveyArea` specified as an exposure. To determine the random effects structure, Restricted Maximum Likelihood (REML) estimation is used, and an ANOVA test is performed to assess the necessity of including random effects. The test yields a p-value of <0.001 ($\chi_4^2 = 56.25$), providing strong evidence to include random effects in the model. To refine the random effects structure, the random components `month` and `dataYear` are individually removed, and the model is refitted each time. ANOVA tests after each removal yield p-values of 0.6554 ($\chi_1^2 = 0.1991$) for `month` and 0.2255 ($\chi_1^2 = 1.4692$) for `dataYear`, suggesting weak to moderate evidence against the necessity of these components. Consequently, both `month` and `dataYear` are excluded from the model, leading to a slight improvement in the AIC score. With only `country` and `region` retained as random effects, the model is further evaluated to determine whether a random intercept or random slope is necessary. The residuals of the model, grouped by `country` and `region`, are assessed against the seven covariates. The analysis indicates no systematic patterns in the residuals, suggesting that random slopes are not needed for these variables. Next, the fixed effects are selected using Maximum Likelihood (ML) estimation. A preliminary model is fitted with all seven covariates: `popDensity`, `built`, `growingSeason`, `rain`, `temperature`, `nightlight`, and `cleanWater`. The AIC score for this model is 2870.634. The p-values indicate that `growingSeason`, `rain`, `nightlight`, and `cleanWater` are not significant. An ANOVA test confirms this result, with a p-value of 0.1476 ($\chi_4^2 = 6.7866$), leading to the removal of these variables. The revised model, with `popDensity`, `built`, and `temperature` as fixed effects, achieves an improved AIC score of 2869.42. Further evaluation of this model shows that all three remaining covariates are statistically significant. To validate these findings, an ANOVA test against an intercept-only model is conducted, yielding strong evidence (p-value = 7.1×10^{-6}, $\chi_4^2 = 29.209$) that `popDensity`, `built`, and `temperature` are essential for the model. The final negative binomial mixed-effects model is refitted using Maximum Likelihood estimation. The model includes `popDensity`, `built`, and `temperature` as fixed effects, `country` and `region` as random effects, and `surveyArea` as an exposure. This optimized model provides a robust representation of the data, with all

included variables demonstrating significance and contributing to the explanation of the response variable.

Model Specification:

$$\text{FSWCount}_i \sim \text{Negative Binomial}(\mu_i, \eta)$$

$$\log(\mu_i) = \beta_0 + \beta_1 \cdot \text{popDensity}_i + \beta_2 \cdot \text{built}_i + \beta_3 \cdot \text{temperature}_i + \alpha_{\text{country}[i]} + \beta_{\text{region}[i]} + \log(\text{surveyArea}_i)$$

$$\alpha_{\text{country}} \sim \text{Normal}(0, \sigma_\alpha^2), \quad \beta_{\text{region}} \sim \text{Normal}(0, \sigma_\beta^2), \quad \eta \quad \text{(Overdispersion Parameter)}$$

7 VARIABLE SELECTION USING SOCIOLOGICAL VARIABLES

Using the negative binomial mixed-effects model, the next step involves selecting the variables required for the final model. A preliminary model is fitted with `country`, `region`, `dataYear`, and `month` as random effects, and `hivRate`, `protected`, `hivFear`, `highCrime`, `insectNet`, `ageFirstSex`, and `wealthIndex` as covariates, with `surveyArea` specified as an exposure. Restricted Maximum Likelihood (REML) estimation is employed to determine the appropriate random effects structure. An ANOVA test yields a p-value of 2.89×10^{-13} ($\chi_4^2 = 64.76$), providing strong evidence in favor of including random effects in the model. To refine the random effects structure, the random components `month` and `dataYear` are individually removed, and the model is refitted after each removal. ANOVA tests after these removals indicate that `month` has a p-value of 0.7424 ($\chi_1^2 = 0.1081$), suggesting little to no evidence against excluding it, and it is subsequently removed. In contrast, `dataYear` has a p-value of 0.01456 ($\chi_1^2 = 5.9688$), providing moderate to strong evidence for its inclusion, so it is retained in the model. With this refinement, the model's AIC improves slightly. An ANOVA test confirms the necessity of `country`, `region`, and `dataYear` as random effects, with a p-value of 5.96×10^{-14} ($\chi_3^2 = 64.65$). Using the refined random effects structure, the fixed effects are selected through Maximum Likelihood (ML) estimation. A preliminary model is fitted with all seven covariates: `hivRate`, `protected`, `hivFear`, `highCrime`, `insectNet`, `ageFirstSex`, and `wealthIndex`. This model yields an AIC score of 2870.943. However, the p-values indicate that `hivRate`, `protected`, `hivFear`, `ageFirstSex`, and `wealthIndex` are not significant. An ANOVA test is conducted, resulting in a p-value of 0.5092 ($\chi_7^2 = 0.5092$), which provides weak evidence against the null hypothesis that these variables are unnecessary. Consequently, these five variables are removed from the model. The updated model, which retains `highCrime` and `insectNet` as fixed effects, achieves an improved AIC score of 2863.208. Further analysis confirms that both `highCrime` and `insectNet` are significant. To validate this finding, an ANOVA test is conducted against an intercept-only model, yielding strong evidence with a p-value of 2.6×10^{-7} ($\chi_3^2 = 33.421$). This result confirms the necessity of `highCrime` and `insectNet` in the model. Fixed-effect selection is performed using ML estimation, with REML set to `false` in the R code. The final negative binomial mixed-effects model is refitted using Maximum Likelihood estimation. The model includes `highCrime` and `insectNet` as fixed effects, `country`, `region`, and `dataYear` as random effects, and `surveyArea` as an exposure. This optimized model demonstrates no issues with linearity or overdispersion, and all included variables are significant, providing a robust representation of the data.

Model Specification:

$$\text{FSWCount}_i \sim \text{Negative Binomial}(\mu_i, \eta)$$

$$\log(\mu_i) = \beta_0 + \beta_1 \cdot \text{highCrime}_i + \beta_2 \cdot \text{insectNet}_i + \alpha_{\text{country}[i]} + \beta_{\text{region}[i]} + \gamma_{\text{dataYear}[i]} + \log(\text{surveyArea}_i)$$

$$\alpha_{\text{country}} \sim \text{Normal}(0, \sigma_\alpha^2), \quad \beta_{\text{region}} \sim \text{Normal}(0, \sigma_\beta^2), \quad \gamma_{\text{dataYear}} \sim \text{Normal}(0, \sigma_\gamma^2), \quad \eta \quad \text{(Overdispersion Parameter)}$$

8 RESULTS AND CONCLUSION

The analysis revealed significant associations between environmental and sociological variables and the distribution of Female Sex Workers (`FSWs`) across sampled regions. Statistical modeling provided valuable insights into the influence of both fixed and random effects on `FSW` counts. Among the environmental variables, `popDensity` was significantly associated with `FSW` counts, with lower-density areas exhibiting higher prevalence. This suggests that sparsely populated regions may have unique socio-economic or demographic factors influencing `FSW` distribution. Similarly, the `built` index was negatively correlated with `FSW` counts, indicating that less urbanized areas tend to have higher counts, potentially due to disparities in infrastructure or economic conditions. Additionally, `temperature` showed a positive relationship with `FSW` counts, which may reflect region-specific environmental or cultural dynamics.

Sociological variables also played a critical role. `highCrime` was associated with a significant increase in `FSW` counts, suggesting that socio-economic vulnerabilities or unsafe environments contribute to the observed patterns. `insectNet` usage was positively correlated with `FSW` counts. While the mechanism underlying this association is unclear, it may serve as a proxy for regional variations in public health resources or living conditions. The random effects structure of the model accounted for additional unmeasured variability across spatial and temporal dimensions. Random intercepts for `country`, `region`, and `dataYear` were statistically significant, with variance component estimates capturing hierarchical clustering in the data. This highlights the importance of controlling for contextual factors when modeling complex phenomena such as `FSW` distribution.

The final negative binomial mixed-effects model demonstrated improved model fit, with the lowest AIC score of 2863.208. Likelihood ratio tests confirmed the significance of the fixed effects and the necessity of including random effects. These results underscore the importance of integrating both environmental and sociological dimensions to understand the factors influencing `FSW` counts. This study highlights the complex interplay between environmental and sociological variables in shaping `FSW` distribution. However, the findings are generalizable only to the sampled regions and timeframes or to other contexts with similar characteristics. As this was an observational study, causal inferences cannot be drawn. Nonetheless, the statistical evidence provides valuable insights to inform targeted public health interventions and policies. Future research should consider incorporating additional covariates, such as economic indicators or access to healthcare, and explore longitudinal data to capture dynamic changes over time. Such approaches could refine our understanding of the factors affecting `FSW` distribution and guide effective public health strategies.

REFERENCES

[1] Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Müller, K., Ooms, J., Robinson, D., Seidel, D., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. `https://doi.org/10.21105/joss.01686`

[2] Xie, Y. (2021). knitr: A General-Purpose Package for Dynamic Report Generation in R. Available at: `https://CRAN.R-project.org/package=knitr`.

[3] Brooks, M. E., Kristensen, K., van Benthem, K. J., Magnusson, A., Berg, C. W., Nielsen, A., Skaug, H. J., Mächler, M., Bolker, B. M. (2017). glmmTMB Balances Speed and Flexibility Among Packages for Zero-Inflated Generalized Linear Mixed Modeling. *The R Journal*, *9*(2), 378–400. `https://doi.org/10.32614/RJ-2017-066`

[4] Hartig, F. (2020). DHARMa: Residual Diagnostics for Hierarchical (Multi-Level/Mixed) Regression Models. Available at: `https://CRAN.R-project.org/package=DHARMa`.

[5] Greenwood, M. (n.d.). *Stat 505 Course Notes*. [Online].

[6] OpenAI. (n.d.). ChatGPT. [Online].

[7] R Core Team (2023). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. Available at: `https://www.R-project.org/`.

[8] Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., et al. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. `https://doi.org/10.21105/joss.01686`.

[9] Schloerke, B., Cook, D., Larmarange, J., Briatte, F., Marbach, M., Thoen, E., Elberg, A., Crowley, J. (2024). GGally: Extension to 'ggplot2'. R package version 2.2.1. Available at: `https://CRAN.R-project.org/package=GGally`.

9 APPENDIX

9.1 TABLES

Table 1: Summary Statistics of the response variable

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
FSWCount	0	0	0	5.564	5	134

Table 2: Correlation table between FSWCount and the other covariates.

Variable	Correlation
built	-0.2023
growingSeason	-0.1416
rain	-0.05463
temperature	0.2717
nightlight	-0.03247
cleanWater	-0.1469

Table 3: Table of AIC scores for 6 selected mixed-effect models using environmental variables.

Model	AIC
Poisson	5978
Negative Binomial	2874
Hurdle Poisson	4900
Hurdle Negative Binomial	2999
Zero-inflated Poisson	2876
Zero-inflated Negative Binomial	4861

Table 4: Table of AIC scores for 6 selected mixed-effect models using sociological variables.

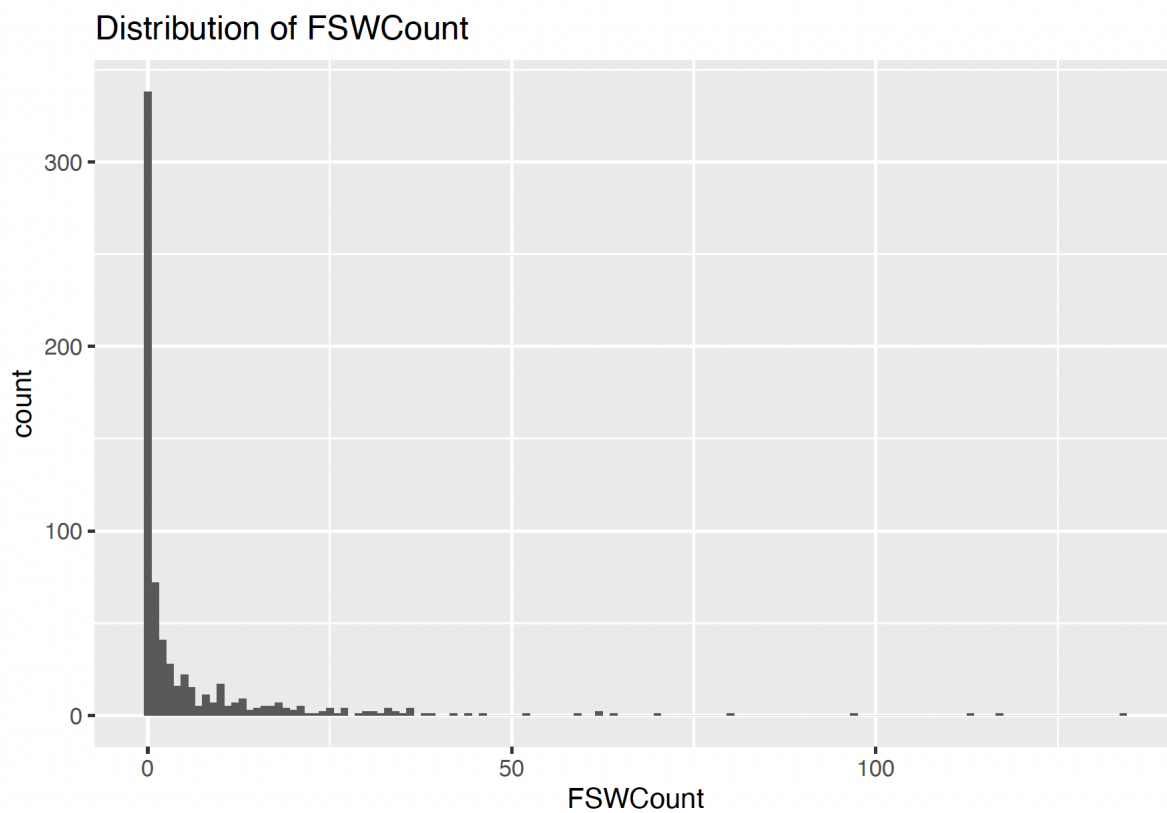| Model	AIC
Poisson	5551
Negative Binomial	2873
Hurdle Poisson	4602
Hurdle Negative Binomial	2989
Zero-inflated Poisson	2875
Zero-inflated Negative Binomial	4562

Distribution of FSWCount

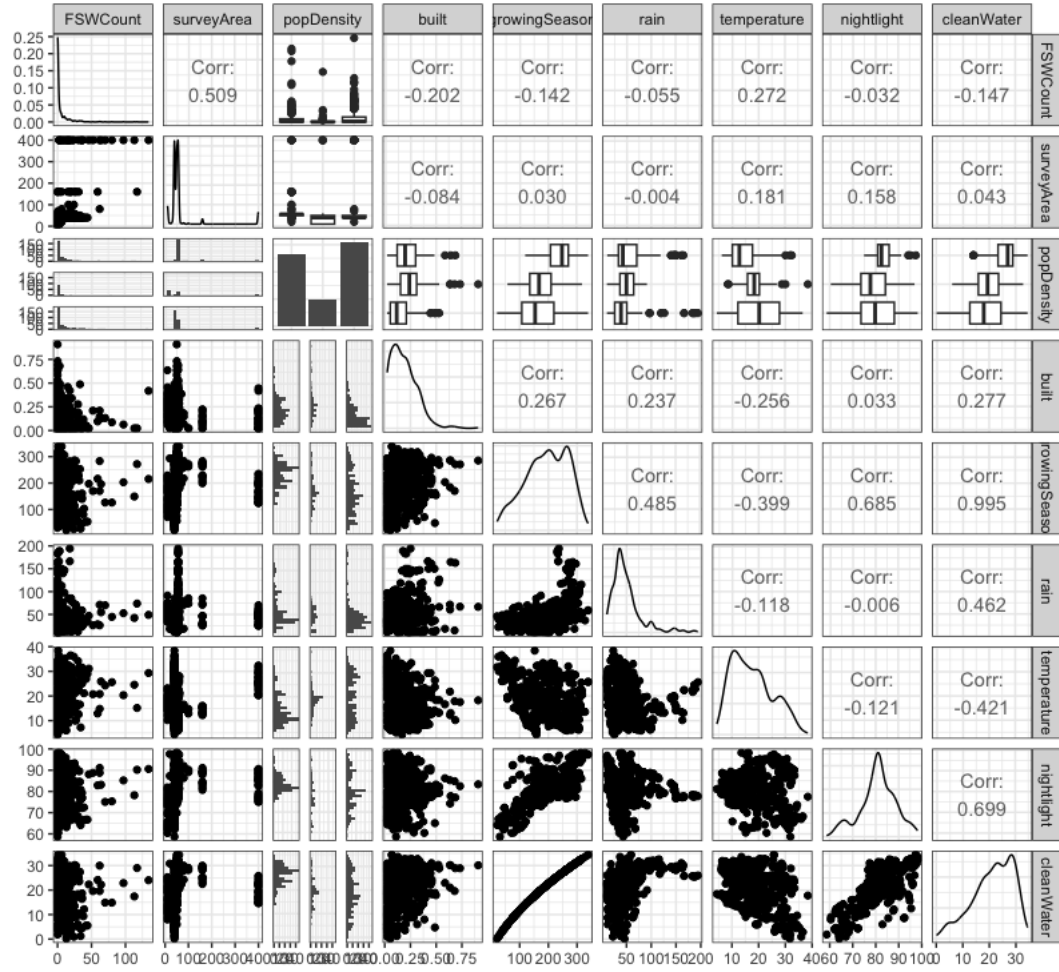Figure 1: The distribution of female sex worker counts

Figure 2: Correlation matrix of the variables showing the correlation between the variables.
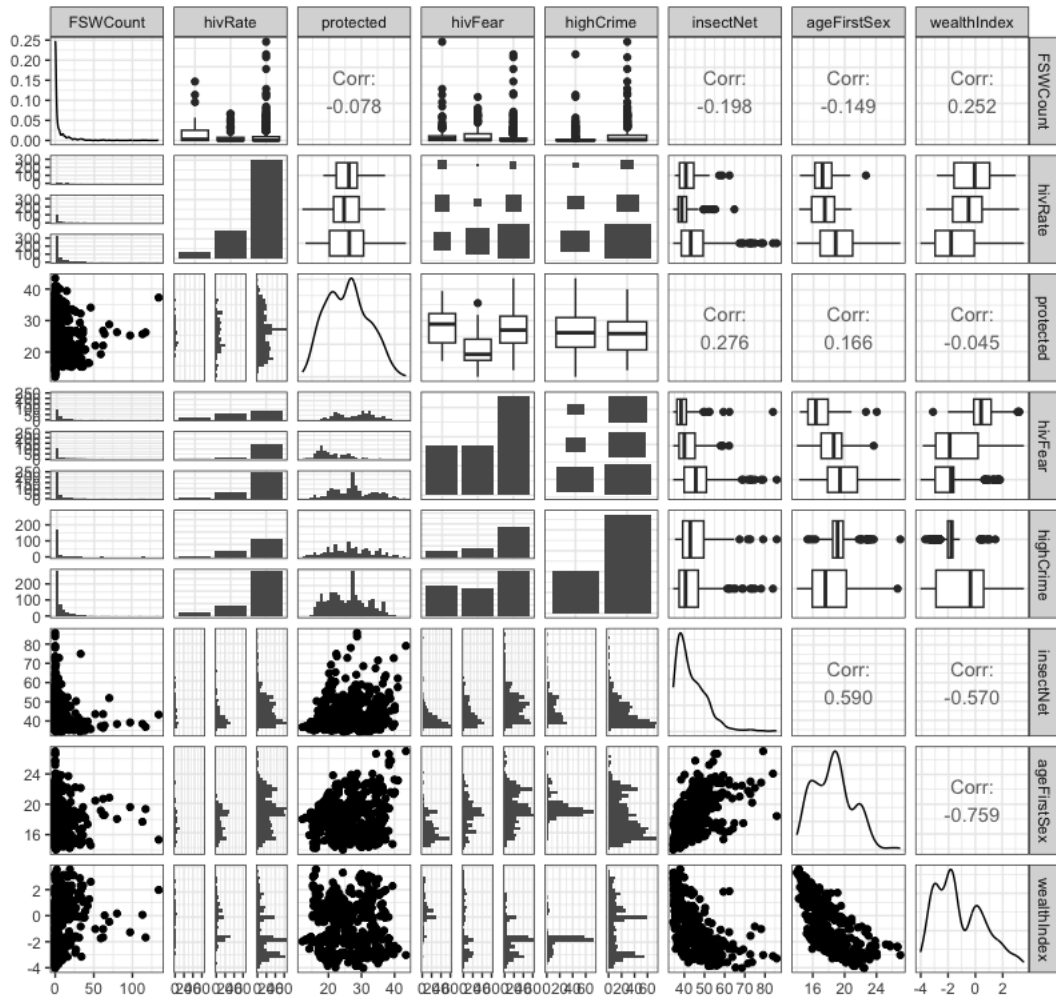
Figure 3: Correlation matrix of the variables showing the correlation between the variables
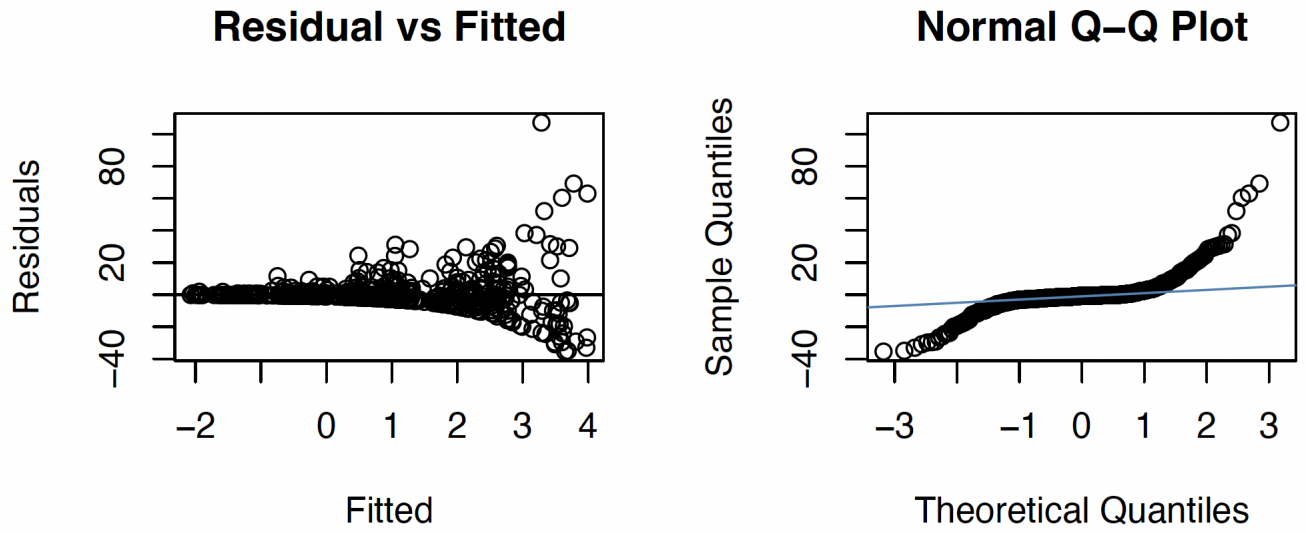

Figure 4: Diagnostic plot of the poisson mixed-effect model using environmental variables.
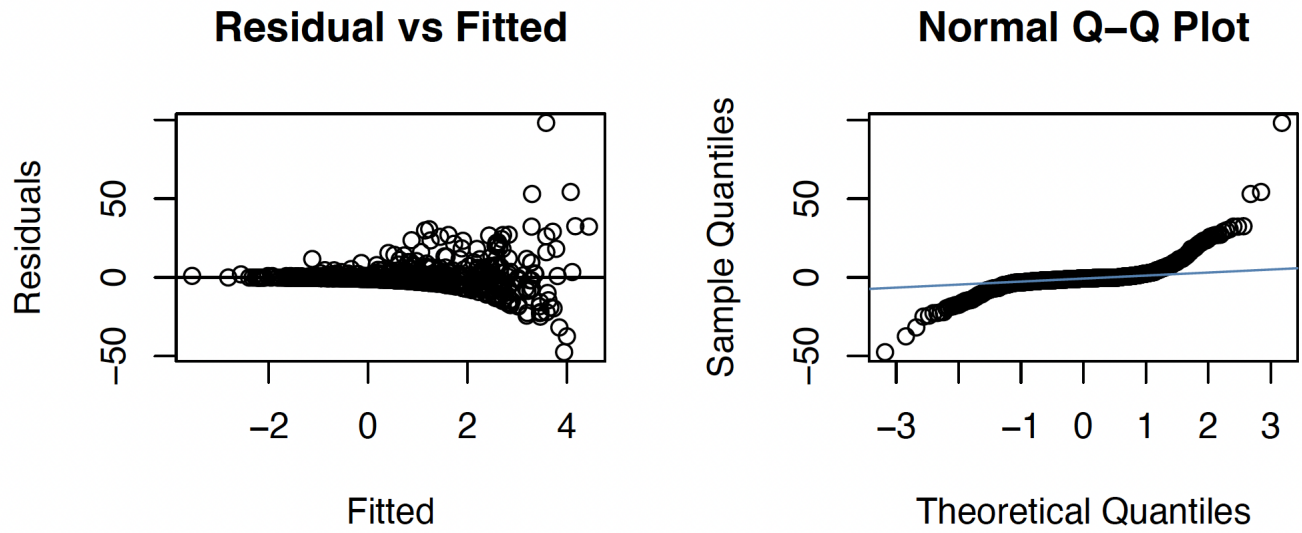
Figure 5: Diagnostic plot of the poisson mixed-effect model using sociological variables.

R CODES

The code below loads various R libraries for data manipulation, visualization, statistical modeling, and diagnostics. It includes tools like `tidyverse` for data wrangling, `lme4` and `glmmTMB` for mixed-effects modeling, and `ggplot2` for visualizations. Diagnostic and post-hoc analysis libraries, such as `DHARMa`, `performance`, and `emmeans`, are also included, while `knitr` options ensure clean output. The `theme_set(theme_bw())` applies a consistent theme to visualizations.

```r
library(tidyverse)
library(easyalluvial)
library(effects)
library(ggmosaic)
library(car)
library(lmerTest)
library(mosaic)
#devtools::install_github("greenwood-stat/ggResidpanel")
library(ggResidpanel)
library(emmeans)
knitr::opts_chunk$set(echo = FALSE)
knitr::opts_chunk$set(warning = FALSE,message = FALSE)
library(tidyverse)
library(lme4)
library(ggplot2)
library(GGally)
library(car)
library(MASS)
library(mosaic)
library(knitr)
library(performance)
library(glmnet)
library(nlme)
library(glmmTMB)
library(DHARMa)
theme_set(theme_bw())
```

The code below preprocesses the dataset by converting specified variables into factors, handling missing data, and preparing separate datasets for training and testing. The training dataset excludes rows with missing values, while the testing dataset retains only rows where FSWCount is missing. It also checks the multicollinearity for both environmental and sociological variables. It also ploted the counts of th FSW.

```r
training <- load.data %>%
  mutate(country = factor(country),
         region = factor(region),
         month = factor(month),
         dataYear = factor(dataYear),
         popDensity = factor(popDensity),
         hivRate = factor(hivRate),
         hivFear = factor(hivFear),
         highCrime = factor(highCrime)) %>%
  drop_na()

testing <- load.data %>%
  mutate(country = factor(country),
         region = factor(region),
         month = factor(month),
         dataYear = factor(dataYear),
         popDensity = factor(popDensity),
         hivRate = factor(hivRate),
         hivFear = factor(hivFear),
         highCrime = factor(highCrime)) %>%
  filter(is.na(FSWCount))



library(GGally)
training %>%
  ggpairs(5:13,
          upper = list(continuous = GGally::wrap(ggally_cor, stars = FALSE))) +
  theme_bw()

library(GGally)
training %>%
  ggpairs(c(5, 14:20),
          upper = list(continuous = GGally::wrap(ggally_cor, stars = FALSE))) +
  theme_bw()


ggplot(training, aes(x = FSWCount)) +
geom_histogram(binwidth = 10,
fill = "lightblue",
color = "black")+
labs(x = "Female Sex Workers", y = "Frequency", title = "Histogram plot of the
    Count of FSW") +
theme_minimal()
training$residmlx <- resid(fit1)
training$predx <- predict(fit1)
par(mfrow=c(1,2))

# Plotting the residuals - random scatter no patterns
plot(training$predx, training$residmlx , ylab="Residuals", xlab="Fitted",
     main="Residual vs Fitted")
abline(0, 0)
```

```r
# Normality of our residuals
qqnorm(training$residmlx)
qqline(training$residmlx, col="steelblue")


training$residsmlx <- resid(fits1)
training$predsx <- predict(fits1)
par(mfrow=c(1,2))

# Plotting the residuals - random scatter no patterns
plot(training$predsx, training$residsmlx , ylab="Residuals", xlab="Fitted",
     main="Residual vs Fitted")
abline(0, 0)


# Normality of our residuals
qqnorm(training$residsmlx)
qqline(training$residsmlx, col="steelblue")
```

The code below performs model selection for environmental variables by fitting six different generalized linear mixed-effects models (GLMMs) using the `glmmTMB` function. Each model evaluates the relationship between `FSWCount` and several covariates (`popDensity`, `built`, `growingSeason`, `rain`, `temperature`, `nightlight`, and `cleanWater`) while accounting for random effects (`country`, `region`, `month`, `dataYear`) and an offset for `surveyArea`. The models differ in their distributional assumptions and structural forms, including Poisson, negative binomial, hurdle, and zero-inflated variations. The `AIC` function is used to compare the models based on their Akaike Information Criterion (AIC) scores.

```r
# MODEL SELECTION FOR ENVIRONMENTAL VARIABLES
library(pscl)
library(glmmTMB)

fit1 <- glmmTMB(FSWCount ~ popDensity + built + growingSeason + rain +
                temperature + nightlight + cleanWater +
                (1 | country) + (1 | region) + (1 | month) + (1 | dataYear),
                offset = log(surveyArea),
                data = training, family = poisson) # Poisson

fit2 <- glmmTMB(FSWCount ~ popDensity + built + growingSeason + rain +
                temperature + nightlight + cleanWater +
                (1 | country) + (1 | region) + (1 | month) + (1 | dataYear),
                offset = log(surveyArea),
                data = training, family = nbinom2) # Negative Binomial

fit3 <- glmmTMB(FSWCount ~ popDensity + built + growingSeason + rain +
                temperature + nightlight + cleanWater +
                (1 | country) + (1 | region) + (1 | month) + (1 | dataYear),
                offset = log(surveyArea), zi = ~1,
                family = truncated_poisson, data = training) # Hurdle Poisson

fit4 <- glmmTMB(FSWCount ~ popDensity + built + growingSeason + rain +
                temperature + nightlight + cleanWater +
                (1 | country) + (1 | region) + (1 | month) + (1 | dataYear),
                offset = log(surveyArea), zi = ~1,
                family = truncated_nbinom2, data = training) # Hurdle Negative
                    Binomial

fit5 <- glmmTMB(FSWCount ~ popDensity + built + growingSeason + rain +
                temperature + nightlight + cleanWater +
                (1 | country) + (1 | region) + (1 | month) + (1 | dataYear),
                offset = log(surveyArea), zi = ~1,
                family = nbinom2, data = training) # Zero-Inflated Negative
                    Binomial

fit6 <- glmmTMB(FSWCount ~ popDensity + built + growingSeason + rain +
                temperature + nightlight + cleanWater +
                (1 | country) + (1 | region) + (1 | month) + (1 | dataYear),
                offset = log(surveyArea), zi = ~1,
                family = poisson, data = training) # Zero-Inflated Poisson

# Compare models using AIC
AIC(fit1, fit2, fit3, fit4, fit5, fit6)
```

The code below performs random effects selection for environmental variables by fitting multiple variations of negative binomial generalized linear mixed-effects models (`glmmTMB`) with different combinations of random effects. Each model includes the same fixed effects (`popDensity`, `built`, `growingSeason`, `rain`, `temperature`, `nightlight`, and `cleanWater`) and an offset for `surveyArea`. The models differ in the inclusion of random intercepts for `country`, `region`, `month`, and `dataYear`. The `AIC` function is used to compare models based on their Akaike Information Criterion (AIC), while the `anova` function evaluates differences in fit between nested models.

```
# RANDOM EFFECT SELECTION FOR ENVIRONMENTAL VARIABLES

fit2A <- glmmTMB(FSWCount ~ popDensity + built + growingSeason + rain +
                 temperature + nightlight + cleanWater +
                 (1 | country) + (1 | region) + (1 | month) + (1 | dataYear),
                 offset = log(surveyArea),
                 data = training, family = nbinom2, REML = TRUE) # Full model

fit2B <- glmmTMB(FSWCount ~ popDensity + built + growingSeason + rain +
                 temperature + nightlight + cleanWater +
                 (1 | country) + (1 | region) + (1 | month),
                 offset = log(surveyArea),
                 data = training, family = nbinom2, REML = TRUE) # Excluding
                     dataYear

fit2C <- glmmTMB(FSWCount ~ popDensity + built + growingSeason + rain +
                 temperature + nightlight + cleanWater +
                 (1 | country) + (1 | region) + (1 | dataYear),
                 offset = log(surveyArea),
                 data = training, family = nbinom2, REML = TRUE) # Excluding
                     month (BEST)

fit2D <- glmmTMB(FSWCount ~ popDensity + built + growingSeason + rain +
                 temperature + nightlight + cleanWater +
                 (1 | country) + (1 | month) + (1 | dataYear),
                 offset = log(surveyArea),
                 data = training, family = nbinom2, REML = TRUE) # Excluding
                     region

fit2E <- glmmTMB(FSWCount ~ popDensity + built + growingSeason + rain +
                 temperature + nightlight + cleanWater +
                 (1 | region) + (1 | month) + (1 | dataYear),
                 offset = log(surveyArea),
                 data = training, family = nbinom2, REML = TRUE) # Excluding
                     country

fit2p <- glmmTMB(FSWCount ~ popDensity + built + growingSeason + rain +
                 temperature + nightlight + cleanWater,
                 offset = log(surveyArea),
                 data = training, family = nbinom2, REML = TRUE) # No random
                     effects

fit2q <- glmmTMB(FSWCount ~ popDensity + built + growingSeason + rain +
                 temperature + nightlight + cleanWater +
                 (1 | country) + (1 | region),
                 offset = log(surveyArea),
                 data = training, family = nbinom2, REML = TRUE) # Only country
                     and region

fit2r <- glmmTMB(FSWCount ~ popDensity + built + growingSeason + rain +
                 temperature + nightlight + cleanWater +
```

```
46                       (1 | country),
47                       offset = log(surveyArea),
48                       data = training, family = nbinom2, REML = TRUE) # Only country
49
50  # Compare models using AIC
51  AIC(fit2A, fit2B, fit2C, fit2D, fit2E)
52
53  # Perform ANOVA tests for nested model comparisons
54  anova(fit2A, fit2B)
55  anova(fit2A, fit2C)
56  anova(fit2A, fit2p)
57  anova(fit2A, fit2q)
58  anova(fit2A, fit2r)
```

The code below performs fixed effects selection for environmental variables using generalized linear mixed-effects models (`glmmTMB`). It compares models with different combinations of fixed effects (`popDensity`, `built`, `temperature`, `growingSeason`, `rain`, `nightlight`, and `cleanWater`) while keeping random effects for `country` and `region`. The `AIC` function is used to compare model fit, and the `anova` function evaluates differences between nested models to identify the most parsimonious model.

```
1   # FIXED VARIABLE SELECTION FOR ENVIRONMENTAL VARIABLES
2
3   # Full model with all fixed effects
4   fit2i <- glmmTMB(FSWCount ~ popDensity + built + growingSeason + rain +
5                    temperature + nightlight + cleanWater +
6                    (1 | country) + (1 | region),
7                    offset = log(surveyArea),
8                    data = training, family = nbinom2, REML = FALSE) # Negative
                         Binomial
9
10  # Reduced model with selected fixed effects
11  fit2t <- glmmTMB(FSWCount ~ popDensity + built + temperature +
12                   (1 | country) + (1 | region),
13                   offset = log(surveyArea),
14                   data = training, family = nbinom2, REML = FALSE) # Negative
                         Binomial
15
16  # Intercept-only model
17  fit2u <- glmmTMB(FSWCount ~ 1 +
18                   (1 | country) + (1 | region),
19                   offset = log(surveyArea),
20                   data = training, family = nbinom2, REML = FALSE) # Negative
                         Binomial
21
22  # Model comparison
23  AIC(fit2i)   # AIC for full model
24  AIC(fit2t)   # AIC for reduced model
25  AIC(fit2u)   # AIC for intercept-only model
26
27  # Perform ANOVA tests for nested model comparisons
28  anova(fit2t, fit2i) # Compare reduced model with full model
29  anova(fit2u, fit2t) # Compare intercept-only model with reduced model
```

The code below performs model selection for sociological variables by fitting six different generalized linear mixed-effects models (`glmmTMB`) with various distributional assumptions. Each model evaluates the relationship between `FSWCount` and sociological covariates (`hivRate`, `protected`, `hivFear`, `highCrime`, `insectNet`, `ageFirstSex`, and `wealthIndex`) while accounting for random effects (`country`, `region`, `month`, and `dataYear`) and an offset for `surveyArea`. The models include Poisson, negative binomial, hurdle, and zero-inflated distributions. The `AIC` function is used to compare models and identify the best-fitting one.

```r
# MODEL SELECTION FOR SOCIOLOGICAL VARIABLES
library(pscl)
library(glmmTMB)

fits1 <- glmmTMB(FSWCount ~ hivRate + protected + hivFear + highCrime + insectNet +
                 ageFirstSex + wealthIndex +
                 (1 | country) + (1 | region) + (1 | month) + (1 | dataYear),
                 offset = log(surveyArea),
                 data = training, family = poisson) # Poisson

fits2 <- glmmTMB(FSWCount ~ hivRate + protected + hivFear + highCrime + insectNet +
                 ageFirstSex + wealthIndex +
                 (1 | country) + (1 | region) + (1 | month) + (1 | dataYear),
                 offset = log(surveyArea),
                 data = training, family = nbinom2) # Negative Binomial

fits3 <- glmmTMB(FSWCount ~ hivRate + protected + hivFear + highCrime + insectNet +
                 ageFirstSex + wealthIndex +
                 (1 | country) + (1 | region) + (1 | month) + (1 | dataYear),
                 offset = log(surveyArea), zi = ~1,
                 family = truncated_poisson, data = training) # Hurdle Poisson

fits4 <- glmmTMB(FSWCount ~ hivRate + protected + hivFear + highCrime + insectNet +
                 ageFirstSex + wealthIndex +
                 (1 | country) + (1 | region) + (1 | month) + (1 | dataYear),
                 offset = log(surveyArea), zi = ~1,
                 family = truncated_nbinom2, data = training) # Hurdle Negative
                     Binomial

fits5 <- glmmTMB(FSWCount ~ hivRate + protected + hivFear + highCrime + insectNet +
                 ageFirstSex + wealthIndex +
                 (1 | country) + (1 | region) + (1 | month) + (1 | dataYear),
                 offset = log(surveyArea), zi = ~1,
                 family = nbinom2, data = training) # Zero-Inflated Negative
                     Binomial

fits6 <- glmmTMB(FSWCount ~ hivRate + protected + hivFear + highCrime + insectNet +
                 ageFirstSex + wealthIndex +
                 (1 | country) + (1 | region) + (1 | month) + (1 | dataYear),
                 offset = log(surveyArea), zi = ~1,
                 family = poisson, data = training) # Zero-Inflated Poisson

# Compare models using AIC
AIC(fits1, fits2, fits3, fits4, fits5, fits6)
```

The code below performs random effects selection for sociological variables using generalized linear mixed-effects models (`glmmTMB`) with a negative binomial distribution. Each model evaluates the relationship between `FSWCount` and sociological covariates (`hivRate`, `protected`, `hivFear`, `highCrime`, `insectNet`, `ageFirstSex`, and `wealthIndex`) while varying the random effects structure (`country`, `region`, `month`, and `dataYear`). The `AIC` function is used to compare model fit, and the `anova` function evaluates differences between nested models to identify the most appropriate random effects structure.

```
# RANDOM EFFECT SELECTION FOR SOCIOLOGICAL VARIABLES

# Full model with all random effects
fits2A <- glmmTMB(FSWCount ~ hivRate + protected + hivFear + highCrime +
    insectNet +
                  ageFirstSex + wealthIndex +
                  (1 | country) + (1 | region) + (1 | month) + (1 | dataYear),
                  offset = log(surveyArea),
                  data = training, family = nbinom2, REML = TRUE) # Negative
                      Binomial

# Excluding dataYear as a random effect
fits2B <- glmmTMB(FSWCount ~ hivRate + protected + hivFear + highCrime +
    insectNet +
                  ageFirstSex + wealthIndex +
                  (1 | country) + (1 | region) + (1 | month),
                  offset = log(surveyArea),
                  data = training, family = nbinom2, REML = TRUE) # Negative
                      Binomial

# Excluding month as a random effect (BEST model)
fits2C <- glmmTMB(FSWCount ~ hivRate + protected + hivFear + highCrime +
    insectNet +
                  ageFirstSex + wealthIndex +
                  (1 | country) + (1 | region) + (1 | dataYear),
                  offset = log(surveyArea),
                  data = training, family = nbinom2, REML = TRUE) # Negative
                      Binomial (BEST)

# Excluding region as a random effect
fits2D <- glmmTMB(FSWCount ~ hivRate + protected + hivFear + highCrime +
    insectNet +
                  ageFirstSex + wealthIndex +
                  (1 | country) + (1 | month) + (1 | dataYear),
                  offset = log(surveyArea),
                  data = training, family = nbinom2, REML = TRUE) # Negative
                      Binomial

# Excluding country as a random effect
fits2E <- glmmTMB(FSWCount ~ hivRate + protected + hivFear + highCrime +
    insectNet +
                  ageFirstSex + wealthIndex +
                  (1 | region) + (1 | month) + (1 | dataYear),
                  offset = log(surveyArea),
                  data = training, family = nbinom2, REML = TRUE) # Negative
                      Binomial

# Model with no random effects
fits2p <- glmmTMB(FSWCount ~ hivRate + protected + hivFear + highCrime +
    insectNet +
                  ageFirstSex + wealthIndex,
```

```r
                    offset = log(surveyArea),
                    data = training, family = nbinom2, REML = TRUE) # Negative
                        Binomial

# Model with only country and region as random effects
fits2q <- glmmTMB(FSWCount ~ hivRate + protected + hivFear + highCrime +
    insectNet +
                    ageFirstSex + wealthIndex +
                    (1 | country) + (1 | region),
                    offset = log(surveyArea),
                    data = training, family = nbinom2, REML = TRUE) # Negative
                        Binomial

# Model with only country as a random effect
fits2r <- glmmTMB(FSWCount ~ hivRate + protected + hivFear + highCrime +
    insectNet +
                    ageFirstSex + wealthIndex +
                    (1 | country),
                    offset = log(surveyArea),
                    data = training, family = nbinom2, REML = TRUE) # Negative
                        Binomial

# Compare models using AIC
AIC(fits2A, fits2B, fits2C, fits2D, fits2E)

# Perform ANOVA tests for nested model comparisons
anova(fits2A, fits2B)
anova(fits2A, fits2C)
anova(fits2A, fits2D)
anova(fits2A, fits2p)
anova(fits2A, fits2q)
anova(fits2C, fits2p)
```

The code below fits several generalized linear models (GLMs) and generalized linear models with a negative binomial distribution (`glm.nb`) to evaluate the relationship between `FSWCount` and various predictors. Models vary in their inclusion of fixed effects such as environmental (`popDensity`, `built`, `temperature`) and sociological (`hivRate`, `highCrime`, `insectNet`) variables, as well as geographic factors (`region`, `dataYear`). The `AIC` function is used to compare model fits, while `anova` performs statistical comparisons between nested models using chi-squared tests. Predictions are generated for new data using the best-fitting model.

```
# Fitting Poisson and Negative Binomial Models

# Poisson model
moda <- glm(FSWCount ~ popDensity + built + growingSeason + rain +
            temperature + nightlight + cleanWater + hivRate + protected +
            hivFear + highCrime + insectNet + ageFirstSex + wealthIndex +
            country + region + dataYear + offset(log(surveyArea)),
            family = "poisson", data = training)

# Full Negative Binomial model
nega <- glm.nb(FSWCount ~ popDensity + built + growingSeason + rain +
               temperature + nightlight + cleanWater + hivRate + protected +
               hivFear + highCrime + insectNet + ageFirstSex + wealthIndex +
               country + region + dataYear + offset(log(surveyArea)),
               data = training)

# Negative Binomial model excluding "country"
negb <- glm.nb(FSWCount ~ popDensity + built + growingSeason + rain +
               temperature + nightlight + cleanWater + hivRate + protected +
               hivFear + highCrime + insectNet + ageFirstSex + wealthIndex +
               region + dataYear + offset(log(surveyArea)),
               data = training)

# Further reduced model excluding "rain"
negc <- glm.nb(FSWCount ~ popDensity + built + growingSeason +
               nightlight + cleanWater + hivRate +
               hivFear + highCrime + insectNet + ageFirstSex + wealthIndex +
               region + dataYear + offset(log(surveyArea)),
               data = training)

# Simplified model
negd <- glm.nb(FSWCount ~ popDensity + nightlight + hivFear +
               highCrime + insectNet + ageFirstSex +
               region + dataYear + offset(log(surveyArea)),
               data = training)

# Final minimal model
nege <- glm.nb(FSWCount ~ popDensity + highCrime + insectNet +
               ageFirstSex + region + dataYear + offset(log(surveyArea)),
               data = training)

negf <- glm.nb(FSWCount ~ popDensity + insectNet +
               highCrime + region + dataYear + offset(log(surveyArea)),
               data = training)

negh <- glm.nb(FSWCount ~ popDensity + ageFirstSex +
               highCrime + region + dataYear + offset(log(surveyArea)),
               data = training)

# Model Comparisons
AIC(negf, negh)                              # Compare AIC values for reduced models
```

```
52  anova(nega, negb, test = "Chisq")    # Chi-squared test between nested models
53  anova(negb, negc, test = "Chisq")    # Compare negb vs negc
54  anova(negd, negc, test = "Chisq")    # Compare negd vs negc
55  anova(negd, nege, test = "Chisq")    # Compare negd vs nege
56  anova(negf, nege, test = "Chisq")    # Compare negf vs nege
```