

The School of Mathematics



THE UNIVERSITY  
*of* EDINBURGH

# Analysis on Risk Factors for Leptospirosis Based on Generalized Linear Mixed Models

by

Yile Shi

Dissertation Presented for the Degree of  
MSc in Statistics with Data Science

August 2021

Supervised by  
Dr. Gail Robertson and Dr. Amy Wilson

## Executive Summary

**Background:** Leptospirosis, a widely distributed zoonotic disease, has become a major public health concern due to its significant effect on human and animal morbidity and mortality [1]. Although various researches have been conducted to study this zoonotic infection in different domains, work on identifying individuals at risk, especially in low-income rural areas in developing countries, is still deficient.

**Research question:** This research aims to shed light on the potential risk factors that affect individual's leptospirosis infection in a rural area of Kenya, and further generate knowledge to help inform health intervention programmes for leptospirosis infections reduction. We consider 8 explanatory variables: gender, age, land use, occupation, distance to local hospital, altitude, family size, and occupation of household head, as well as 3 environmental variables with nested structures.

**Data:** We use a subset with 595 subjects of the sampled data collected by the International Livestock Research Institute from members of households in villages in Tana River County, Kenya, including the leptospirosis test results using ELIZA (enzyme-linked immunosorbent assay).

**Methods:** We construct binomial generalized linear mixed models including the random effect of environmental variables, where the response of interest is individual's ELIZA teste result. The random intercept model with village effect is selected due to its best performance by AIC (Akaike Information Criterion). Moreover, we conduct model diagnosis and evaluation, including residual plots and ROC (Receiver Operating Characteristic) curves (Figure 1).

**Results:** We obtain the random intercept among villages is normally distributed with mean 0 and standard deviation 0.5053. Considering the village effect, gender, land use (at 0.1 significance level) and occupation of household head tend to have significant effects on individual's leptospirosis infection. Specifically, male individuals (OR 0.6427; 95% CI 0.4194 to 0.9764) have less risk for leptospirosis seropositivity than female individuals; people working on pastoral land (OR 0.3582; 95% CI 0.1151 to 1.1372) or riverine land (OR 0.5738; 95% CI 0.1657 to 1.8141) are less likely to be infected, comparing with irrigation land group; individuals with household heads working as pastoralists (OR 6.4987; 95% CI 2.2190 to 19.3265) take more risk for leptospirosis seropositivity than farmers group.

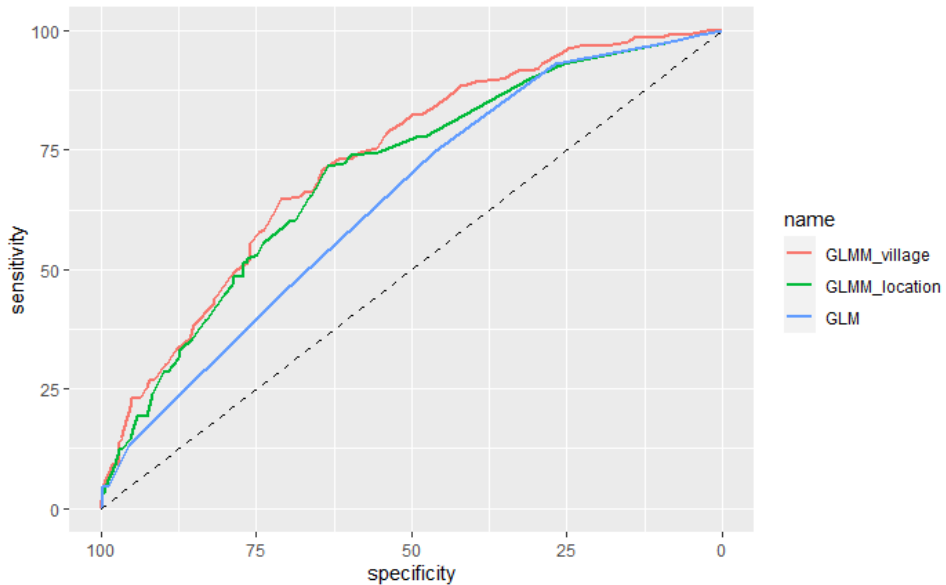


Figure 1: ROC curves of different models

## Acknowledgments

I am sincerely grateful to the supervisors of this project, Dr. Gail Robertson and Dr. Amy Wilson, as well as PhD supporter Miss. Rebecca Akeresola, for their advice and guidance. I would also like to thank the International Livestock Research Institute for providing the data and background information used in the research.

## University of Edinburgh – Own Work Declaration

This sheet must be filled in, signed and dated - your work will not be marked unless this is done.

Name: Yile Shi

Matriculation Number: s2168022

Title of work: Analysis on Leptospirosis Risk Factors Based on Generalized Linear Mixed Model

We confirm that all this work is my own except where indicated, and that We have:

- Clearly referenced/listed all sources as appropriate
- Referenced and put in inverted commas all quoted text (from books, web, etc)
- Given the sources of all pictures, data etc. that are not my own
- Not made any use of the report(s) or essay(s) of any other student(s) either past or present
- Not sought or used the help of any external professional academic agencies for the work
- Acknowledged in appropriate places any help that We have received from others (e.g. fellow students, technicians, statisticians, external sources)
- Complied with any other plagiarism criteria specified in the Course handbook

We understand that any false claim for this work will be penalised in accordance with the University regulations (<https://teaching.maths.ed.ac.uk/main/msc-students/msc-programmes/statistics/data-science/assessment/academic-misconduct>).

Signature:

A handwritten signature in black ink that reads "Yile Shi". The signature is written in a cursive, slightly slanted style.

Date: 2022/8/10

# Contents

<b>1</b>	<b>Introduction</b>	<b>2</b>
<b>2</b>	<b>Exploratory Data Analysis</b>	<b>2</b>
2.1	Missing Data . . . . .	2
2.2	Feature Selection . . . . .	3
2.2.1	Nominal Variables . . . . .	3
2.2.2	Numerical Variables . . . . .	5
2.2.3	Environmental Variables . . . . .	5
<b>3</b>	<b>Implementation</b>	<b>8</b>
3.1	Generalized Linear Mixed Model . . . . .	8
3.2	Model Construction and Selection . . . . .	8
3.2.1	Step 1: Initial Model . . . . .	8
3.2.2	Step 2: Find the Optimal Fixed Structure . . . . .	8
3.2.3	Step 3: Find the Optimal Random Structure . . . . .	9
<b>4</b>	<b>Results</b>	<b>11</b>
4.1	Model Diagnosis . . . . .	11
4.2	Model Evaluation . . . . .	11
4.3	Model Interpretation . . . . .	13
<b>5</b>	<b>Conclusions</b>	<b>15</b>
	<b>Appendix</b>	<b>17</b>
	<b>Word Count</b>	<b>18</b>

# 1 Introduction

Leptospirosis, a zoonosis found worldwide, remains a major public health issue in many developing countries, which is expected to become more important due to a rapid urbanization, global warming, and extreme climatic events such as floods [2]. Due to the difficulty of accurate data collection, identification of risk factors for individual’s leptospirosis infection in rural areas of developing countries becomes an active research topic. This report aims to contribute to this topic by exploring potential risk factors behind leptospirosis diagnosis. We expect that a better understanding of the influence of factors on leptospirosis infection can help to inform the development of lifestyle interventions for leptospirosis risk reduction.

We apply exploratory data analysis (EDA) on sampled data collected by the International Livestock Research Institute from members of households in villages in Tana River County, Kenya, and extract a subset of 595 individuals with 11 columns as well as the leptospirosis test results using ELIZA (enzyme-linked immunosorbent assay). We build generalized linear mixed models including the random effect of environmental variables, and select the best performed one. Our goal is to find, among the following potential risk factors, the most likely to be relevant to leptospirosis seropositivity, with an environment effect into consideration: gender; age; land use; occupation; distance to local hospital; altitude; family size; and occupation of household head.

## 2 Exploratory Data Analysis

### 2.1 Missing Data

We detect a serious problem of missing values in the dataset after dropping duplicated data. Table 1 displays the number and proportion of missing values, arranged in descending order.

variable	count	proportion
occupation	343	0.3649
disthosp	259	0.2755
livestk_home	236	0.2511
location	236	0.2511
landuse	4	0.0043
gender	1	0.0011
age	1	0.0011

Table 1: Number and proportion of missing values in variables

`landuse`, `gender` and `age` contain missing values less than 1% and we drop the corresponding observations without much loss of information of the original dataset.

According to data description, `livestk_home` is a binary variable indicating whether or not livestock is kept in the household of sampled person. Although contact with animals increases risk of leptospirosis [3], from figure 2, we observe a significant imbalance, which probably leads to insignificant results. Thus, we exclude this column from further modelling.

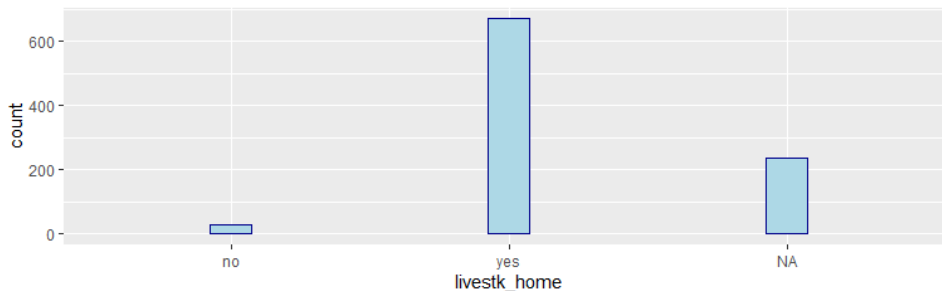


Figure 2: Distribution of `livestk_home`

**disthosp** is the Euclidean distance from the sampled person’s household to local hospital. Here, we assume that people from the same village go to the same hospital. Then, we impute the missing values with the mean distance to the local hospital of each village. However, we observe that data in **disthosp** in village 12, 13 and 23 are completely missing and fail to impute. Instead, we drop corresponding rows in these villages.

**location** is the anonymised location of area where sampling was done. Again, we consider using **village** to determine the corresponding location in the same row. However, we still meet the problem that data in **locations** in some villages are completely missing. Besides, we observe that subjects from the same village could belong to different locations. possibly because the village is on the boundary of two locations or it is just a mistake in data input. As a result, we drop all missing values in this column instead of imputation.

Individual’s occupation could have great influence on leptospirosis infection [3], associated with various aspects. Figure 3 shows the conditional distributions of **occupation** in juvenile (< 18 years old) and adult ( $\geq 18$  years old) groups.

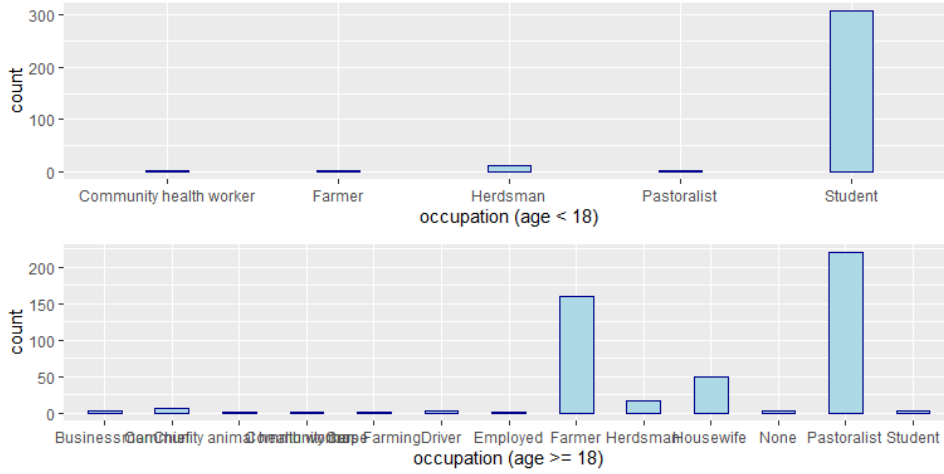


Figure 3: Distribution of **occupation**, conditioning on **age**

It is shown that juveniles are most likely to be students, hence we impute the missing occupation for juvenile individuals with "Student". However, the case for adults is complicated. A p-value  $< 2.2e - 16$  from the  $\chi^2$  test shows strong evidence that individual’s occupation is correlated with the type of land use. We also observed this relationship in Figure 4. Specifically, most sampled adults working on irrigation or riverine land are farmers, while individuals pastoral land are most likely to be pastoralists. Besides, **gender** and **constituency** are also observed to be correlated with **occupation**.

It is hard to determine an appropriate imputation method for adult group. Instead, we drop adult subjects missing in **occupation**.

## 2.2 Feature Selection

We now determine the columns used for modelling. **result**, which represents whether the sample individual is leptospirosis seropositive or not based on ELISA (enzyme-linked immunosorbent assay), is considered as the response variable.

As for potential risk factors, identifier columns including **sampleid** and **parent** are excluded from our model, as well as **relationshiphh** and **genhhid**. Besides, as mentioned before, we do not take **livestk\_home** into consideration. We select features from the rest columns based on plots and the work of Goarant et al (2016) [4].

### 2.2.1 Nominal Variables

We start with nominal variables **gender**, **occupation**, **landuse**, **hhoccup** and **hhgender**. First, we count the number of observations in each category of **occupation** and **hhoccup**, the occupation of household head, grouped by two columns respectively.



Figure 4: Distributions of `occupation` in adult group, conditioning on `landuse`

occupation		hhoccup	
category	count	category	count
Student	239	Pastoralist	345
Pastoralist	159	Farmer	247
Farmer	156	Chief	15
Housewife	36	Businessman	5
Herdsman	20	Civil Servant	4
Chief	6	Herdsman	3
Businessman	2	Crop Farming	2
Driver	2	None	2
Community animal health worker	1	Casual Labourer	1
Community nurse	1	Driver	1
Crop Farming	1		
Employed	1		
None	1		

Table 2: Number of observations in each category in `occupation` and `hhoccup`

As shown in Table 2, most categories contain observations less than 20. Small and insufficient group size would lead to unstable and biased results. On the other hand, numerous occupation groups could result in larger degrees of freedom, which increases model complexity. As a remedy, we merge occupation groups with similar meanings and risks of exposure, such as "Crop Farming" and "Farmer", "Herdsman" and "Pastoralist". However, most smaller occupation groups have very different risk profiles, e.g. drivers are unlikely to have the same contact with animal risk as farmers. In this case, we disregard such categories and discard corresponding observations.

We plot stacked bar charts for nominal variables and report the prevalence in each category. From Figure 5 and Table 3, `gender`, `occupation`, `landuse` and `hhoccup` have significantly different prevalence among categories, hence we select them as feature candidates for modelling. Although `hhgender` also tends to have significant effect, we exclude this column due to its imbalance where over 80% observations are males.



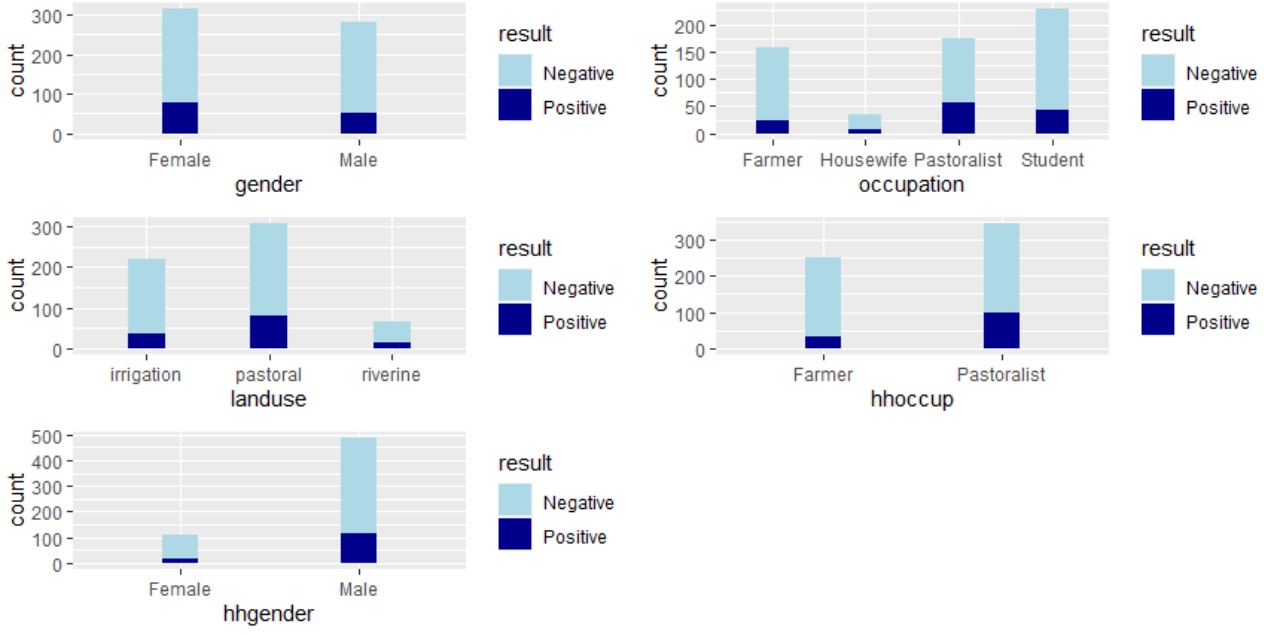


Figure 5: Stacked bar plot of **result** in factor variables

factor	category	prevalence
gender	Female	0.2508
	Male	0.1821
occupation	Farmer	0.1465
	Student	0.1834
	Pastoralist	0.3257
	Housewife	0.2353
landuse	irrigation	0.1636
	pastoral	0.2597
	riverine	0.2090
hhoccup	Farmer	0.1325
	Pastoralist	0.2803
hhgender	Female	0.1481
	Male	0.2341

Table 3: Prevalence in each category of factor variables

### 2.2.2 Numerical Variables

We make box plots for numerical variables including **age**, **altitude**, **nmales**, **nfemales**, **famsize** and **disthosp**, grouped by **result**.

Comparing with nominal variables, numerical variables do not show such significant difference between "Negative" and "Positive" groups in Figure 6. Based on [4], we select **age**, **altitude**, **famsize**, the number of people in the sampled household and **disthosp** as feature candidates for modelling.

### 2.2.3 Environmental Variables

Environmental variables including **village**, **location** and **constituency** indicate the geographical information of sampled people, with an order **village** < **location** < **constituency**. Again, we plot the stacked bar charts of the response variable and compute the prevalence, grouped by the levels in **village**, **location** and **constituency** respectively.

Nested structures are observed in environmental variables and we would like to take the influence of them into consideration due to great difference of prevalence among groups. However, we might consider the random effect of these variables rather than fixed effect, due to the hierarchical structures

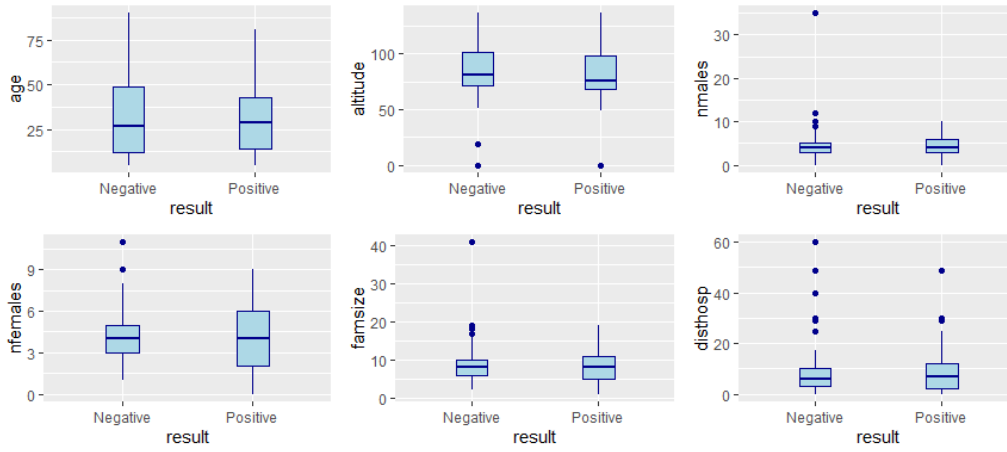


Figure 6: Box plots of numerical variables, grouped by **result**

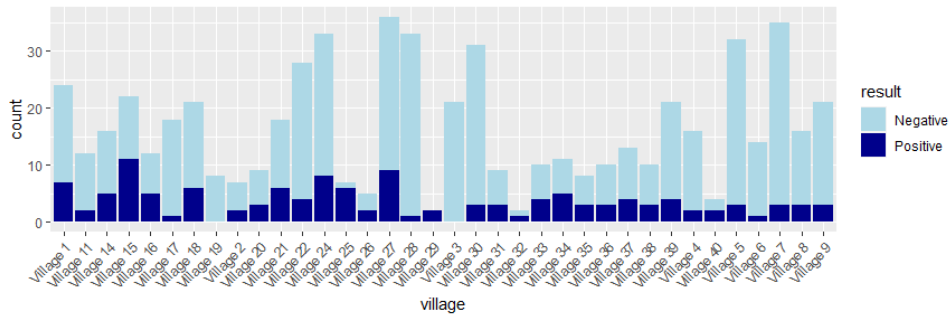


Figure 7: Stacked bar plots of **result** in **village**

category	prevalence	category	prevalence	category	prevalence
Village 1	0.2917	Village 16	0.4167	Village 29	1.0000
Village 2	0.2857	Village 17	0.0556	Village 30	0.0968
Village 3	0	Village 18	0.2857	Village 31	0.3333
Village 4	0.1250	Village 19	0	Village 32	0.5000
Village 5	0.0938	Village 20	0.3333	Village 33	0.4000
Village 6	0.0714	Village 21	0.3333	Village 34	0.4545
Village 7	0.0857	Village 22	0.1429	Village 35	0.3750
Village 8	0.1875	Village 24	0.2424	Village 36	0.3000
Village 9	0.1429	Village 25	0.8571	Village 37	0.3077
Village 11	0.1667	Village 26	0.4000	Village 38	0.3000
Village 14	0.3125	Village 27	0.2500	Village 39	0.1905
Village 15	0.5000	Village 2	0.0303	Village 40	0.5000

Table 4: Prevalence in each category of **village**

category	prevalence	category	prevalence	category	prevalence
Location 1	0.5556	Location 7	0.2857	Location 13	0.3333
Location 2	1.0000	Location 8	0.3182	Location 14	0.2857
Location 3	0.1571	Location 9	0.0962	Location 15	0.0667
Location 4	1.0000	Location 10	0.2424	Location 16	0.3000
Location 5	0.3125	Location 11	0.5000	Location 17	0.1667
Location 6	0.1429	Location 12	0.4250	Location 18	0.0000
				Location 19	0.4000

Table 5: Prevalence in each category of **location**

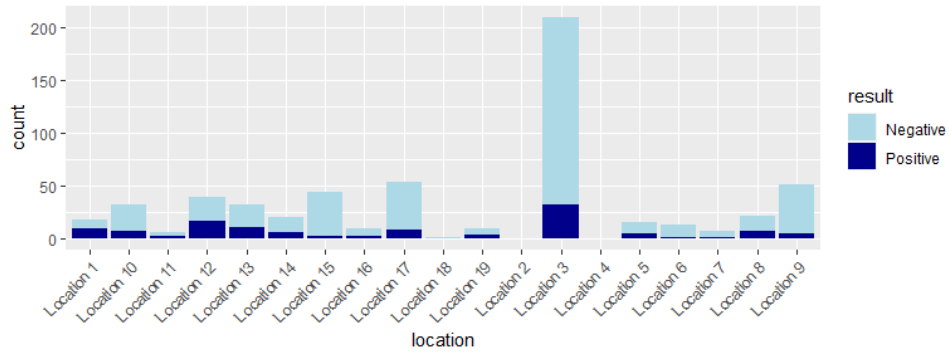


Figure 8: Stacked bar plots of **result** in **location**

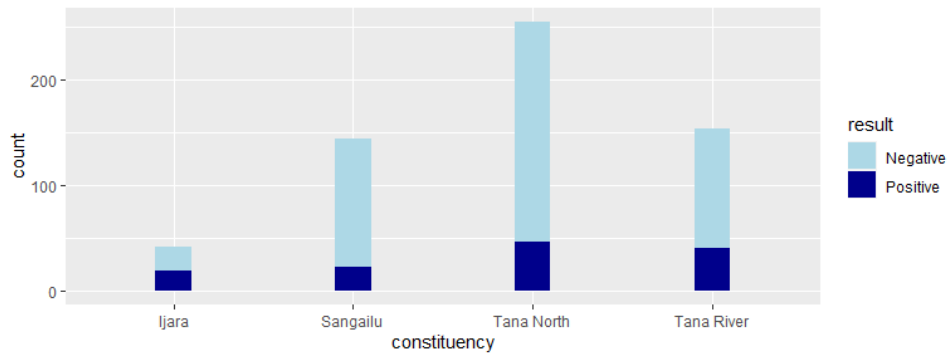


Figure 9: Stacked bar plots of **result** in **constituency**

category	prevalence	category	prevalence
Tana North	0.1843	Tana River	0.2662
Ijara	0.4524	Sangailu	0.1597

Table 6: Prevalence in each category of **constituency**

with numbers of levels in **village** and **location**. Models with random effect will be explained in the next section.

So far, we have extracted a subset for modelling with 595 observations, including 8 feature candidates and 3 environmental variables.

## 3 Implementation

### 3.1 Generalized Linear Mixed Model

The exposition in this subsection follows that in Zuur (2009) [5].

Since the response variable in our model is binomial, we start with the generalized linear model (GLM) for binomial data, i.e. logistic regression.

Suppose the response  $\mathbf{Y} = (Y_1, \dots, Y_n)$   $Y_i \sim \text{Bin}(1, p_i)$ , we have:

$$\text{logit}(p_i) = \ln \frac{p_i}{1 - p_i} = \mathbf{x}_i^T \boldsymbol{\beta}$$

GLMs do not always work well. Sometimes data may contain multiple levels, which is known as nested or hierarchical structure, and data in each level may be correlated. In this case, regression results in different levels could vary and model assumptions such as homogeneity could be violated. To deal with it, we consider an extended model where part of parameters could change across different levels in the hierarchical variables (random effect), and the rest parameters keep consistent across levels (fixed effect). The GLM including both fixed and random effect is called the generalized linear mixed model (GLMM).

Now, for the binomial GLMM for response  $\mathbf{Y}$ , we have:

$$\text{logit}(\mathbf{p}) = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \boldsymbol{\epsilon}$$

Here,  $\mathbf{p} = (p_{ij})$  is matrix where  $p_{ij}$  denotes the probability for individual  $i$  in level  $j$ ;  $\mathbf{X}$  and  $\boldsymbol{\beta}$  are the fixed effects design matrix and fixed effect respectively;  $\mathbf{Z}$  and  $\mathbf{b}$  are the random effects design matrix and random effect respectively;  $\boldsymbol{\epsilon}$  is the matrix for unexplained information. Note that  $\mathbf{b}$  and  $\boldsymbol{\epsilon}$  follow normal distributions with mean 0.

Besides the allowance for correlations among observations and nested data structure, GLMMs also reduce the model complexity. If we include the hierarchical variables with numerous levels in the fixed structure, we will obtain complicated models without stable results and generality. Considering the random effect instead only requires one extra parameter and can obtain more general statements. That's why we consider the random effect of environmental variables rather than their fixed effect.

### 3.2 Model Construction and Selection

#### 3.2.1 Step 1: Initial Model

We start with a binomial GLM, containing feature candidates except geographical variables. Here, we do not contain the interaction terms of variables in the model as we do not detect much interaction among features.

#### 3.2.2 Step 2: Find the Optimal Fixed Structure

The `summary` and `step` function are both used to determine the optimal fixed structure among candidates. The latter function chooses the optimal model and significant features by AIC and the  $\chi^2$  test. Table 7 and 8 display the output of two functions.

We select `gender` and `hhoccup`, which are statistically significant at 0.05 significance level in the output of both functions. Besides, `landuse` is also selected, as it is significant at 0.1 significance level in stepwise selection and one of its levels is statistically significant at 0.05 level in model summary.

Although `occupation` shows its significance in stepwise algorithm, none of its levels is significant in model summary. On the other hand, we observe high correlation between `occupation` and `landuse`, which probably causes collinearity and poor model performance. Therefore, we do not include `occupation` in the fixed structure.

In conclusion, the optimal fixed structure consists of `gender`, `landuse` and `hhoccup`.

	<i>t</i> value	Pr(>   <i>t</i>  )
(Intercept)	-0.341	0.73317
genderMale	-2.428	0.01520 *
age	-1.798	0.07219 .
occupationHousewife	-0.757	0.44890
occupationPastoralist	0.754	0.45068
occupationStudent	-1.423	0.15463
landusepastoral	-2.347	0.01892 *
landuseriverine	-1.607	0.10810
altitude	-1.681	0.09283 .
famsize	0.346	0.72900
hhoccupPastoralist	3.181	0.00147 **
disthosp	0.301	0.76322
Significance codes: 0, '***' 0.001, '**' 0.01, '*' 0.05, '.' 0.1, ' ' 1		

Table 7: Feature significance in initial GLM using `summary` function

	Df	AIC	Pr(>Chi)
(none)		602.72	
altitude	1	603.45	0.0990659 .
age	1	603.86	0.0765869 .
landuse	2	604.59	0.0531339 .
occupation	3	605.55	0.0316715 *
gender	1	606.75	0.0141019 *
hhoccup	1	611.90	0.0008302 ***
Significance codes: 0, '***' 0.001, '**' 0.01, '*' 0.05, '.' 0.1, ' ' 1			

Table 8: Significant features in stepwise algorithm using `step` function

### 3.2.3 Step 3: Find the Optimal Random Structure

With the optimal fixed structure, we now introduce the random effect of environmental variables and determine GLMM with the best performance, using R package `lme4` [6]. As mentioned, we investigate whether there is a village or location effect.

We start with fitting the random intercept model of `village`. To assess whether the model with mixed effect model is better than the ordinary binomial GLM, we refit the latter one over the selected features without random intercept and compare the models using `anova` function. The output is given in Table 9:

	npar	AIC	BIC	logLik	deviance	Chisq	Df	Pr(>Chisq)
GLM	5	603.63	625.57	-296.81	593.63			
GLMM( <code>village</code> )	6	601.87	628.20	-294.93	589.87	3.7611	1	0.05246 .
Significance codes: 0, '***' 0.001, '**' 0.01, '*' 0.05, '.' 0.1, ' ' 1								

Table 9: Comparison between GLM and GLMM of `village`

Lower AIC value indicates that the mixed model including village effect is preferred. Moreover, we extend the random part by adding random slopes. Specifically, the `landuse` effect may be different per village and the same may hold for the `hhoccup` effect. Table 10 show the output of model comparison using `anova` function:

Extending the model with random slopes gives no improvement as the AIC values of models with random slopes get higher. Now we consider the alternative model with the random intercept of `location`. Comparison between random intercept models of `village` and `location` by AIC is displayed in Table 11:

The AIC value of the random intercept model with the village effect is slightly lower. Besides, since

	npar	AIC	BIC	logLik	deviance	Chisq	Df	Pr(>Chisq)
random intercept	6	601.87	628.20	-294.93	589.87			
random intercept + <b>landuse</b>	8	605.75	640.85	-294.87	589.75	0.122	2	0.9408
random intercept + <b>hhoccup</b>	11	611.48	659.76	-294.74	589.48	0.264	3	0.9667
Significance codes: 0, '***' 0.001, '**' 0.01, '*' 0.05, '.' 0.1, ' ' 1								

Table 10: Comparison among GLMMs of **village** with different random structures

	df	AIC
GLMM( <b>village</b> )	6	601.8683
GLMM( <b>location</b> )	6	601.8942

Table 11: Comparison between random intercept models of **village** and **location**

the column **village** contains more levels than **location**, the models considering the random effect on village aspect obtain more detailed results. The random intercept and slope models of **location** performs even worse. Consequently, we choose the random intercept model with village effect as the optimal model.

Note that we do not consider the constituency effect as **constituency** only contains 4 levels. It might be better to include this variable in model's fixed structure. However, after introducing the fixed effect of **constituency**, other variables in the fixed structure become insignificant. Moreover, the standard deviation of the random village or location effect turns to 0, indicating no difference among villages or locations. Thus, the influence of **constituency** is ruled out.

## 4 Results

### 4.1 Model Diagnosis

In linear regressions, we usually look at the residual plots and normal Q-Q plots to check if the model assumptions are satisfied. However, these are probably not suitable for GL(M)Ms. Standard residual plots, when interpreted in the same way as for linear models, seem to show all kind of problems, such as non-normality, heteroscedasticity, even if the model is correctly specified. For example, the standard residual plot of our optimal model, shown in Figure 10, has an obvious pattern. In this case, we get confused and are not able to tell whether such pattern in GLMM residuals is a problem or not.

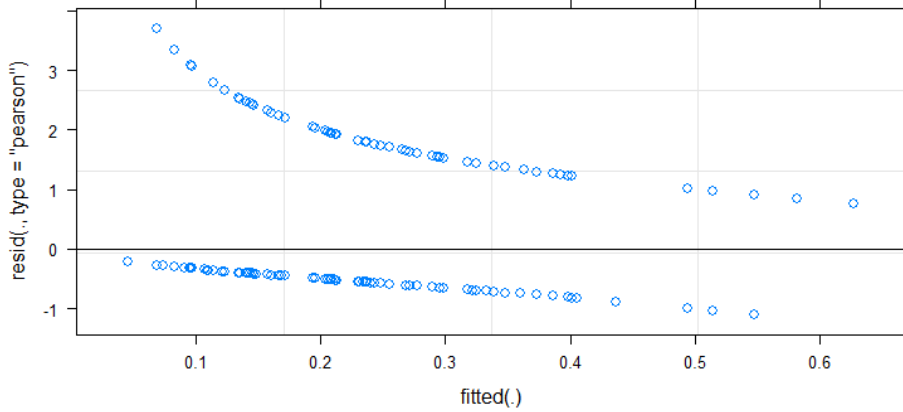


Figure 10: Standard residual plot of the optimal model

To address this problem, we adopt the DHARMa package [7]. It aims to create readily interpretable residuals for GL(M)Ms that are standardized to values between 0 and 1. DHARMa also provides Q-Q plots to detect overall deviations from the expected distribution. The basic steps are:

1. Simulate new response data from the fitted model.
2. Calculate the empirical cumulative density function for the simulated observations, which describes the possible values (and their probability) at the predictor combination of the observed value, assuming the fitted model is correct.
3. The residual is then defined as the value of the empirical density function at the value of the observed data.

We simulate 1000 observations based on our optimal model, and plot the scaled residuals. Intuitively, nothing in the Figure 11 is highlighted in red, indicating no violation of model assumptions is detected. Specifically, well-performed Q-Q plot indicates that the residuals are normally distributed. In the residual plot on the right, we do not observe obvious pattern of scaled residuals. As a result, we do not detect significant problems in the optimal model.

Besides, as recommended in [7], we plot the residuals against specific predictors including `landuse` and `occuparion` in Figure 12. Again, no significant problem is detected.

### 4.2 Model Evaluation

Due to binomial response, for model evaluation, we plot the ROC (Receiver Operating Characteristic) curve and compute the corresponding AUC (Area Under Curve) value with 95% confidence intervals using R package [8]. We repeat the same steps for the ordinary GLM and the random intercept model with location effect. Furthermore, we compute the mean AUC of models using 10-fold cross validation using R package `caret` [9].

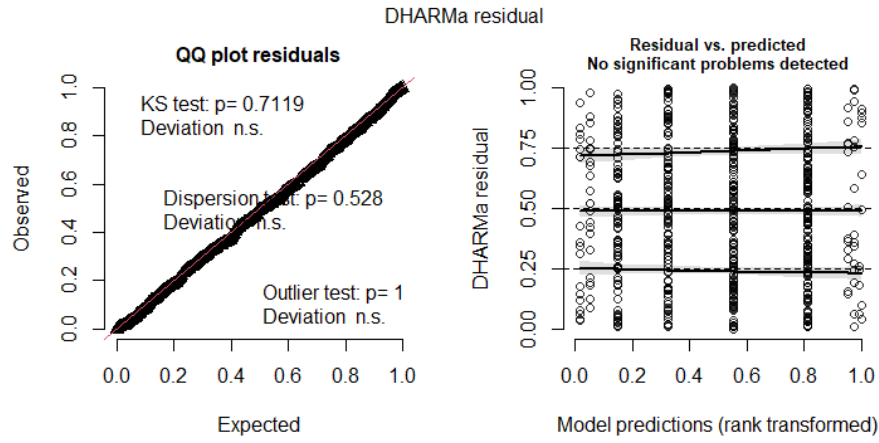


Figure 11: DHARMA residual plot and Q-Q plot of the optimal model

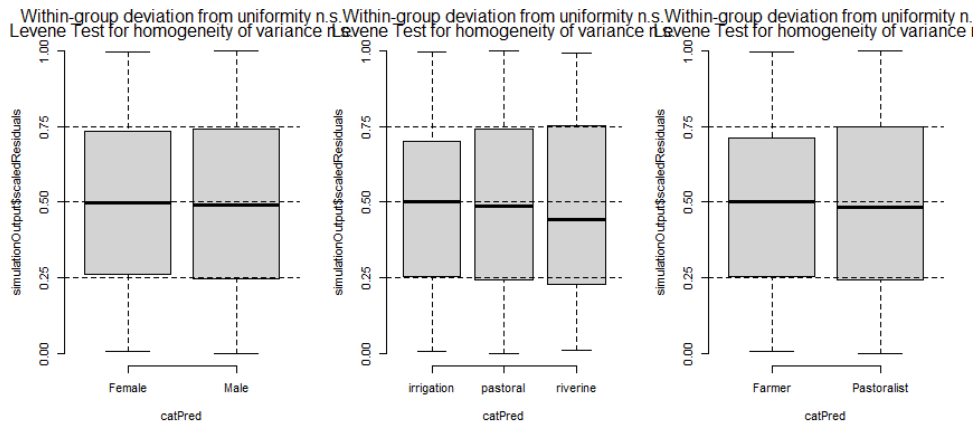


Figure 12: DHARMA residual plot and Q-Q plot of the optimal model

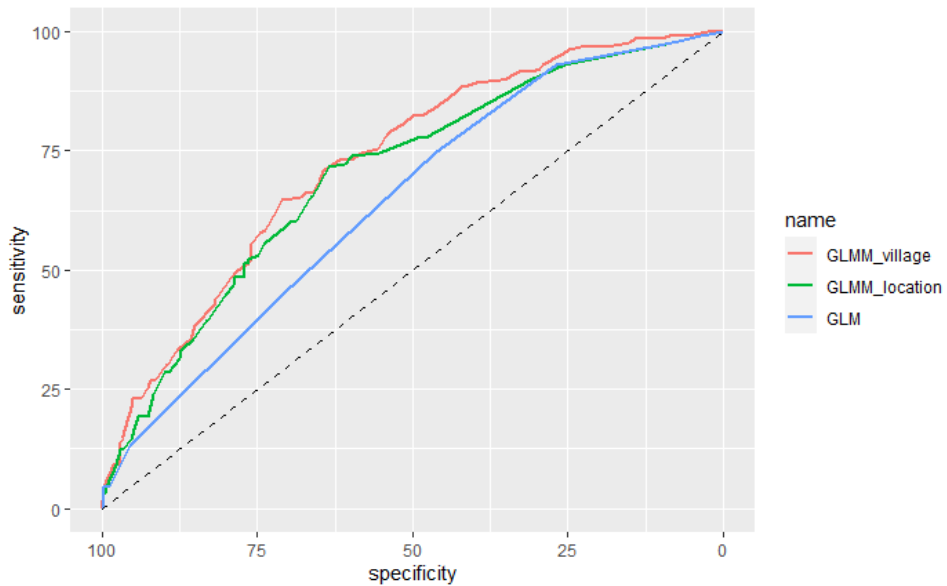


Figure 13: ROC curves of models

According to Figure 13 and Table 12, ROC curve beyond the random baseline and  $AUC = 0.7275 > 0.5$  indicate good performance of the optimal model. Moreover, the random intercept model with village effect over-performs the rest two models, implying its better ability to explain and predict.



	AUC	95% CI	meanAUC in 10-fold CV
GLMM(village)	0.7275	(0.6806, 0.7745)	0.6687
GLMM(location)	0.7018	(0.6520, 0.7516)	0.6681
GLM	0.6437	(0.5942, 0.6932)	0.6454

Table 12: AUC values of models

### 4.3 Model Interpretation

Table 13 and Table 14 display the detailed model summary.

	Estimate	Std.Error	<i>t</i> value	Pr(>   <i>t</i>  )
(Intercept)	-1.6179	0.2589	-6.249	4.13e-10 ***
genderMale	-0.4421	0.2141	-2.065	0.03895 *
landusepastoral	-1.0268	0.5697	-1.802	0.07148 .
landuseriverine	-0.5554	0.5850	-0.949	0.34245
hhoccupPastoralist	1.8716	0.5401	3.465	0.00053 ***
Significance codes: 0, '***' 0.001, '**' 0.01, '*' 0.05, '.' 0.1, ' ' 1				

Table 13: Summary of the fixed effect

Groups	Name	Variance	Std.Dev.
village	(Intercept)	0.2553	0.5053

Table 14: Summary of random effect

**gender**, **hhoccup** and **landuse** (0.1 significance level) show significant influence on leptospirosis seropositivity, hence they are considered as risk factors for leptospirosis. The optimal model is given by:

$$\begin{aligned}
Y_{ij} &\sim \text{Bin}(1, p_{ij}) \\
\text{logit}(p_{ij}) &= -1.6179 - 0.4421 \times \mathbb{I}(\text{gender}_{ij} = \text{Male}) \\
&\quad - 1.0268 \times \mathbb{I}(\text{landuse}_{ij} = \text{pastoral}) - 0.5554 \times \mathbb{I}(\text{landuse}_{ij} = \text{riverine}) \\
&\quad + 1.8176 \times \mathbb{I}(\text{hhoccup}_{ij} = \text{Pastoralist}) + v_i \\
v_i &\sim N(0, 0.5053^2)
\end{aligned}$$

Here,  $Y_{ij}$  denotes the test result of individual  $j$  in village  $i$ , with the probability  $p_{ij}$ . The notation *logit* stands for the logistic link function. The notation  $\mathbb{I}$  represents the indicator function, with the general expression:

$$\mathbb{I}_X(x) = \begin{cases} 1, & \text{if } x \in X \\ 0, & \text{if } x \notin X \end{cases}$$

$v_i$  denotes the random intercept across villages, following a Normal distribution with mean  $\mu = 0$  and standard deviation  $\sigma = 0.5053$ , indicating that the majority of the values (95% to be more exact) of  $v_i$  fall between  $(-1.96 \times 0.5053, 1.96 \times 0.5053)$ .

For example, a male individual from Village 1, with riverine land and household head occupation "Farmer", has a probability  $p$  of being leptospirosis seropositive, where  $p$  satisfies:

$$\text{logit}(p) = \ln \frac{p}{1-p} = -1.6179 - 0.4421 - 0.5554 + v_1 = -2.6154 + v_1, \quad v_1 \sim N(0, 0.5053^2)$$

Based on [10] and [5], the intercept terms including  $v_i$  denote the log odds of leptospirosis seropositivity in the baseline model. The reference groups for features in the fixed structure are **gender** =

*Female*, **landuse** = *irrigation*, and **hhoccup** = *Farmer*. In village  $i$ , we have

$$\text{odds}_{\text{result}} = \frac{P(\text{Positive} \mid \text{female}, \text{irrigation}, \text{farmer})}{P(\text{Negative} \mid \text{female}, \text{irrigation}, \text{farmer})} = e^{-1.6179+v_i} \approx 0.1983 \times e^{v_i}$$

Particularly, considering an average village effect, i.e.  $v_i = 0$ , we obtain an average odds 0.1983 for leptospirosis seropositivity in the baseline model above, with the 95% confidence interval (0.1159, 0.3316).

Holding the average village effect, we have the odds ratios of factors with 95% confidence intervals:

- With **landuse** and **hhoccup** fixed, the odds of leptospirosis seropositivity for those in male group is 0.6427 (0.4194, 0.9764) times the odds of those in female group. Male individuals take less risk for leptospirosis seropositivity, probably because they have better resistance and immunity than females.
- With **gender** and **hhoccup** fixed, the odds of leptospirosis seropositivity for those working on pastoral land is 0.3582 (0.1151, 1.1372) times the odds of those working on irrigation land; the odds of leptospirosis seropositivity for those working on riverine land is 0.5738 (0.1657, 1.8141) times the odds of those working on irrigation land. People working on the irrigation land contact more with both polluted water and animal urine, which may contain the bacteria, and have higher risk for leptospirosis infection [11].
- With **gender** and **landuse** fixed, the odds of leptospirosis seropositivity for people whose household heads work as pastoralists is 6.4987 (2.2190, 19.3265) times the odds of those in farmers group. Comparing with farmers, pastoralists have more contact with animals and are more likely to be exposed to the bacteria. Further, leptospirosis has higher probability to be spread in the households of pastoralists.

## 5 Conclusions

We extract a subset of the data sampled from households in villages of Tana River County, Kenya, and construct a generalized linear mixed model including the random effect of villages, to identify risk factors on leptospirosis. As a result, individual's gender, land use type (0.1 significance level) and the occupation of household head appear to have significant influence on the seropositivity of leptospirosis. Specifically, male individuals seem to have lower risk of transmission of this zoonotic disease. People from household with the irrigation land are most likely to be infected, comparing with those from household with pastoral or riverine land. Besides, members from a household with the head working as a pastoralist tends to take more risk for leptospirosis seropositivity than those from a household with the head working as a farmer.

There are several limitations of this research. First, we delete over 300 observations in EDA section, due to missing or other problems, which leads to a great loss of information and further possibly biased model results. We expect improvement of model accuracy and performance, based on more complete dataset with less mistakes. Second, the response column for ELIZA test result has an imbalanced distribution where most observations have negative results. This may result in model bias as our classifier tends to over-adapt to the negative group. In this case, a possible future research could be the exploration of zero-inflated models on the count data, by concerting the observations from individual to household aspects. Moreover, we could explore the models using Bayesian methods with proper and informative priors in future.

## References

- [1] David A Haake and Paul N Levett. Leptospirosis in humans. *Leptospira and leptospirosis*, pages 65–97, 2015.
- [2] Mathieu Picardeau. Diagnosis and epidemiology of leptospirosis. *Médecine et maladies infectieuses*, 43(1):1–9, 2013.
- [3] Elizabeth Anne Jessie Cook, William Anson de Glanville, Lian Francesca Thomas, Samuel Kar-iuki, Barend Mark de Clare Bronsvort, and Eric Maurice Fèvre. Risk factors for leptospirosis seropositivity in slaughterhouse workers in western kenya. *Occupational and Environmental Medicine*, 74(5):357–365, 2017.
- [4] Cyrille Goarant. Leptospirosis: risk factors and management challenges in developing countries. *Research and reports in tropical medicine*, 7:49, 2016.
- [5] Alain F Zuur, Elena N Ieno, Neil J Walker, Anatoly A Saveliev, Graham M Smith, et al. *Mixed effects models and extensions in ecology with R*, volume 574. Springer, 2009.
- [6] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48, 2015.
- [7] Florian Hartig. *DHARMa: Residual Diagnostics for Hierarchical (Multi-Level / Mixed) Regression Models*, 2021. R package version 0.4.3.
- [8] Xavier Robin, Natacha Turck, Alexandre Hainard, Natalia Tiberti, Frédérique Lisacek, Jean-Charles Sanchez, and Markus Müller. proc: an open-source package for r and s+ to analyze and compare roc curves. *BMC Bioinformatics*, 12:77, 2011.
- [9] Max Kuhn. *caret: Classification and Regression Training*, 2022. R package version 6.0-92.
- [10] Nicholas P Jewell. *Statistics for epidemiology*. chapman and hall/CRC, 2003.
- [11] R. Wesley Farr. Leptospirosis. *Clinical Infectious Diseases*, 21(1):1–8, 07 1995.

## Appendix

Programming part in this research is completed with R.

### R packages:

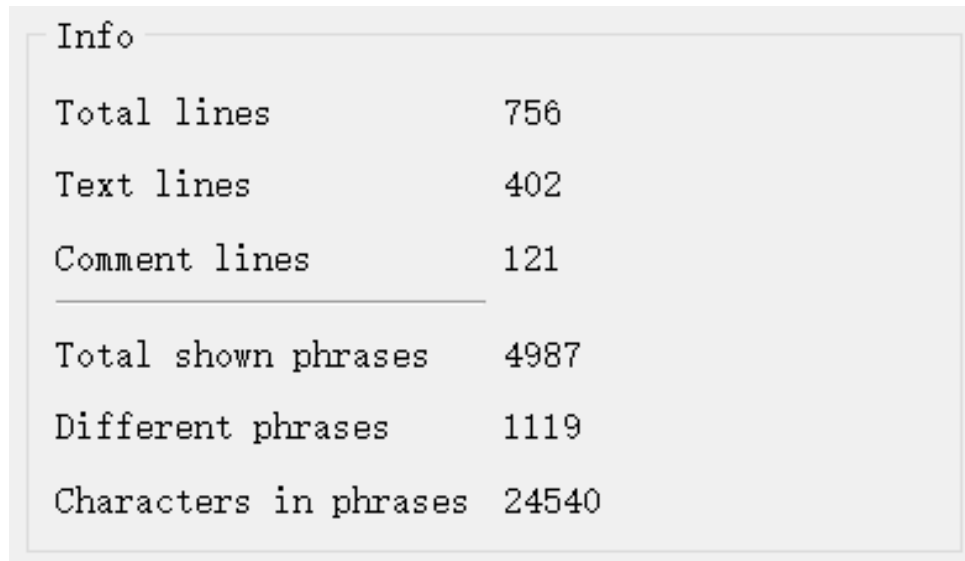
For EDA section, we use the `tidyverse` and `gridExtra` packages. We construct models with `lme4` package. `DHARMA` package is used for model diagnosis. `pROC` and `caret` packages are used for model evaluation. All histograms and line graphs are generated with `ggplot2`.

### R code:

The complete code is available via this [Github repository](#).

## Word Count

This report contains 4987 words, including executive summary, main text, and references. The screenshot using **Analyse Text** function in TeXstudio is provided.

The image shows a screenshot of the 'Info' window in TeXstudio. It displays a table of statistics for the current document. The table has two columns: the first column lists the metric (Total lines, Text lines, Comment lines, Total shown phrases, Different phrases, Characters in phrases) and the second column shows the corresponding value (756, 402, 121, 4987, 1119, 24540). A horizontal line separates the line counts from the phrase and character counts.

Info	
Total lines	756
Text lines	402
Comment lines	121
<hr/>	
Total shown phrases	4987
Different phrases	1119
Characters in phrases	24540

Figure 14: Word count in TeXstudio