OJO, C

# INVESTIGATING RISK FACTORS OF GEOHELMINTHS INFECTIONS PREVALENCE

# *Contents*

# *List of Figures*

# *List of Tables*

# ABSTRACT

Soil transmitted helminths (STH), i.e., geohelminths, are parasitic intestinal worms that cause intestinal nematode infections (INI). According to the Lancet's 2019 global burden of disease study INI - comprising *ascariasis*, *trichuriasis*, and *hookworm disease* - are the most prevalent of the neglected tropical diseases. Additionally, INI are the second most burdensome neglected tropical diseases, after dengue.

An experts' hypothesis of geohelminths infections prevalence is that one or more WaSH (water, sanitation, and hygiene) variables impact prevalence rates. This project investigates whether one or more WaSH variables are geohelminths infections prevalence risk factors. The investigation is via Binomial geostatistical logistic models.

The resulting model suggests that a non-linear, statistically significant, association exists between *piped sewage facilities access* and *Hookworm disease infections prevalence*. It suggests that, as the percentage of inhabitants that have access to piped sewage facilities increases from 0 to approximately 24%, *Hookworm disease* prevalence odds decline. However, beyond the 24% point *Hookworm disease* prevalence odds start increasing with increasing access percentage. The increasing prevalence odds after this 24% [inflection] point might be due to a mix of *access density & poor maintenance*. The piped sewage facilities access percentage is the percentage of an area's inhabitants that have access to piped sewage facilities. The measure does not give an insight into access density, i.e., the median number of people per piped [defaecation] facility. Defaecation facilities that have high access densities might not be well-maintained.

# 1
# INTRODUCTION

Soil transmitted helminths (STH), or geohelminths, are parasitic intestinal worms that cause intestinal nematode infections (INI). According to the Lancet's latest global burden of disease study (2, 3), as of 2019 INI - comprising *ascariasis*, *trichuriasis*, and *hookworm disease* - are the most prevalent of the neglected tropical diseases (Table 1.1). Additionally, INI are the second most burdensome neglected tropical diseases, after dengue (Table 1.2). Therefore, intestinal nematode infections are a focus of the World Health Organization (WHO).

| | prev. per $100k$ | confidence interval | |
|---|---|---|---|
| | | lower | upper |
| Intestinal nematode infections | 11,742.37 | 10,919.41 | 12,747.24 |
| Schistosomiasis | 1,808.96 | 1,514.35 | 2,146.62 |
| Lymphatic filariasis | 928.63 | 759.52 | 1,240.81 |
| Other Neglected Tropical Disease | 852.39 | 836.62 | 867.34 |
| Food-borne trematodiases | 433.45 | 381.95 | 491.70 |

prev.: prevalence

Table 1.1: The five most prevalent neglected tropical diseases during the year 2019. The unit of measure is prevalence per $100k$ inhabitants. Data Source: Institute for Health Metrics & The Lancet Global Burden of Disease 2019

Intestinal nematode infections prevention, control, and elimination activities depend on knowledge of prevalence (1), and prevalence metrics depend on field surveys, which are rather expensive to conduct. They are especially expensive when there are no measures to systematically/strategically direct field scientists to survey areas (4). In the absence of systematic/strategic direction, the field scientists have to survey as wide an area as possible, albeit subject to financial, accessibility, etc., constraints.

An option, in place of expensive field surveys, is INI prevalence pre-

dictions. Prediction model development depends on probable risk factors knowledge & understandings (10). In relation to geohelminths infections, WaSH (water, sanitation, and hygiene) variables are amongst the sets of probable risk factors (13). However, although there are a few studies concluding that one or more WaSH variables are geohelminths infections risk factors, the evidence is contentious (19, 20, 21). This project investigates the possibility that one or more WaSH variables are geohelminths infections risk factors.

| | | confidence interval | |
|---|---|---|---|
| | DALY per $100k$ | lower | upper |
| Other Neglected Tropical Disease | 36.25 | 24.91 | 57.74 |
| Dengue | 30.80 | 10.70 | 42.30 |
| Intestinal nematode infections | 25.50 | 16.24 | 38.81 |
| Schistosomiasis | 21.17 | 13.48 | 34.10 |
| Lymphatic filariasis | 21.05 | 12.39 | 35.07 |

DALY: disability adjusted life years

Table 1.2: The five most burdensome neglected tropical diseases during the year 2019. The unit of measure is disability adjusted life years (DALY) per $100k$ inhabitants, i.e., the number of years lost to disability, ill-health, or early death per $100k$ inhabitants. Data Source: Institute for Health Metrics & The Lancet Global Burden of Disease 2019

## 1.1 Geohelminths

An understanding of the biology, ecology, and transmission dynamics of geohelminths is key to proposing possible prevalence risk factors. As previously noted, geohelminths infections comprises *ascariasis*, *trichuriasis*, and *hookworm disease*. The underlying parasites are *A. lumbricoides*, *T. trichiura*, and *A. duodenale/N. americanus*, respectively (Table 1.3).

During their adult stages the geohelminths parasites inhabit different parts of the human intestines, Table 1.3, wherein they 'reproduce sexually and produce eggs' 4. The eggs exit humans via faeces, whilst defaecating, and mature in the environment. Therein, they either (12, 1)

- Mature into infective eggs, which infect humans when they are accidentally ingested. This applies to A. lumbricoides and T. trichuria.
- Hatch into infective larvae in the environment, soil. The larvae subsequently infect humans by penetrating the skin. This applies to N. americanus/A. duodenale.

| | ascariasis | trichuriasis | hookworm disease |
|---|---|---|---|
| *species* | Ascaris lumbricoides | Trichuris trichiura | Ancylostoma duodenale, Necator americanus |
| *abbr.* | A. lumbricoides | T. trichiura | A. duodenale, N. americanus |
| *name* | Roundworm | Whipworm | Hookworm |
| *adult inhabits* | small intestine | caecum, colon | upper small intestine |
| *infective stage* | ova | ova | larvae |

abbrv: abbreviation

Table 1.3: The core geohelminths species that the World Health Organization focuses on. In each case the infective stage organism lives in the environment/soil.

These transmission dynamics illustrate why WaSH variables are plausible risk factors, and a few studies suggests evidence of WaSH interventions impacting **(a)** geohelminths infections prevalence and incidence, and **(b)** incidences of enteric infections, including geohelminth infections (11, 14). However, a few systematic reviews & meta-analyses suggests that the quality of the evidence is patchy (19, 21, 20). The 2022 systematic review & meta-analysis by Garn et al. notes that

> *… The biological plausibility for improved access to WASH to interrupt transmission of STHs is clear, but WASH interventions as currently delivered have shown impacts that were lower than expected. There is a need for more rigorous and targeted implementation research and process evaluations in order that future WASH interventions can better provide benefit to users.*
>
> — Garn et al. 19

In the environment/soil a number of environmental factors aid or affect the survival of the pre-infective eggs, infective eggs (*A. lumbricoides*, *T. trichiura*), and infective larvae (*A. duodenale/N. americanus*). Brooker et al. (4, 22) review probable and known **(a)** environmental factors, and **(b)** ecological associations between geohelminths distributions and a variety of environmental factors. The factors include temperature, soil moisture, relative atmospheric humidity, altitude, and more. A rather critical observation in these reviews is that each species has a different, particular, spectrum of relationships with environmental factors. Hence, this project focuses on a single disease only: Hookworm disease, as of 2019 it is the most burdensome of the three diseases (Table 1.4).

Altogether, the transmission dynamics suggests that WaSH factors are possible risk factors, whereas the ecological associations between

| | DALY per 100k | confidence interval | | prev. per 100$k$ | confidence interval | |
|---|---|---|---|---|---|---|
| | | lower | upper | | lower | upper |
| Ascariasis | 9.74 | 6.19 | 14.77 | 5759.59 | 5095.11 | 6608.87 |
| Trichuriasis | 3.04 | 1.63 | 5.19 | 4656.81 | 4088.06 | 5332.79 |
| Hookworm disease | 12.72 | 8.09 | 18.96 | 2229.84 | 2059.16 | 2413.79 |

prev.: prevalence

Table 1.4: The 2019 disability adjusted life years(DALY) per 100$k$ inhabitants, and disease prevalence per 100$k$ inhabitants. Data Source: Institute for Health Metrics & The Lancet Global Burden of Disease 2019

geohelminths distributions and a variety of environmental factors suggests that a mix of climatic, topographic, and edaphic factors are also possible risk factors. This project predominantly focuses on WaSH factors, this means - most probably - an insufficient collection of variables for a predictive model, but sufficient for **an explanatory model that models the association between a dependent variable and independent variables, whilst concurrently modelling missing knowledge**.

## 1.2 Project Aim

Determine whether one or more WaSH (water, sanitation, and hygiene) variables are geohelminths infections prevalence risk factors.

## 1.3 Project Objectives

Hence, the project's objectives are to

1. Determine which WaSH (water, sanitation, and hygiene) variables have a statistically significant impact on geohelminths infections prevalence.
2. Determine, estimate, the degree to which impacting WaSH variables affect geohelminths infections prevalence.

via geostatistical models.

## 1.4 Manuscript Structure

The background chapter discusses the biology, ecology, and transmission dynamics of geohelminths. Next, the data chapter outlines the underlying data sets of this project. This leads unto the exploratory analysis chapter, which explores the relationship, or otherwise, between the prevalence of a geohelminth infectious disease and indepen-

dent variables. The preliminary investigation & geostatistical binomial logistic models chapters focus on explanatory modelling. The discussion chapter discusses the project's observations, findings, and limitations. Finally, **all the project's code**, and web graphs, reside within github.com/helminthiases; the scheduled annotation completion date is 12 September 2022.

In this text intestinal nematode infections (INI) & geohelminths infections are synonyms, and the comprising diseases are *ascariasis*, *trichuriasis*, and *hookworm disease*.

# 2
# DATA

This chapter outlines the data sets of this project, and their variables. The data sets fall into two groups. Foremost, the geohelminths infections survey experiments data sets. The key details in these data sets are the geographic coördinates of the survey experiments locations, the number of tests/examinations conducted per location, and the year each set of tests occurred.

The second group of data sets are raster maps of probable risk factors. The raster maps comprise measures or estimates that are independent of the survey experiments, i.e., they are extraneous data maps of extraneous independent variables. An extraneous independent variable value at a survey experiments' location depends on the location's geographic coördinates.

## 2.1 Geohelminths Infections Survey Experiments Data

The raw geohelminths infections data is courtesy of the **E**xpanded **S**pecial **P**roject for **E**limination of **N**eglected Tropical Diseases (ESPEN). ESPEN focuses on a set of African countries only. Table 2.1 outlines the set of core variables present in each country's geohelminths infections examinations data set.

The project adds an extra geographic field to each ESPEN data set, named $identifier$. Each set of observations that share the same coördinate values have the same $identifier$ code; this is akin to a site identification code, it addresses the dearth of ESPEN *site_id* codes.

### *2.1.1 Missing Data*

Aside from missing *site_id* codes, many ESPEN data sets are missing core variables values (Table 2.1). Hence, prior to deciding whether to use a data set it is important to understand the missing values patterns (15). Donald Rubin (18, 15, 16) outlines three fundamental

| variable | description |
|---|---|
| *longitude, latitude* | A site's longitude and latitude coördinates. |
| *year* | The observation's year. |
| *hk_examined, hk_positive* | The number of individuals examined, and the number testing positive, respectively.[1] |
| *asc_examined, asc_positive* | *ditto*[2] |
| *tt_examined, tt_positive* | *ditto*[3] |

[1] hk: Hookworm disease
[2] asc: Ascariasis
[3] tt: Trichuriasis

Table 2.1: The core variables of the geohelminths infections survey experiments data sets. Data Source: ESPEN (Expanded Special Project for Elimination of Neglected Tropical Diseases)

missing data mechanisms

- Missing Completely at Random (MCAR): administrative errors, accident.
- Missing at Random (MAR): there's an association between a variable's missing data, and available independent variables & dependent variables.
- Missing Not at Random (MNAR): missing data associated with missing values of the predictor in question or with unobserved predictors.

If the missing values of a data set in question are *missing completely at random* (MCAR) then complete case analysis will suffice because the complete case excerpt is akin to a random sample from a complete population (15). If MCAR does not hold, e.g., data is *missing at random*, then the complete case excerpt is not representative of the underlying population, therefore population inference is not possible via complete case analysis.

The project uses Steyerberg's method (15, Section 7.7.1) to determine whether a data set's missing independent variables are MCAR. In brief, the method's overarching hypotheses are

**$H_0$**: The missing values of an independent variable are not predictable via the dependent variables and other independent variables.

**$H_1$**: It is quite probable that the missing values of the independent variable are predictable via the dependent variables and other independent variables.

The implication of rejecting the null hypothesis $H_0$ is that the missing values **are not missing completely at random**. In the terms of the ESPEN project, particularly the core values of Table 2.1, the independent variables are *year* & *coördinates*, whereby

$$coördinates = (longitude, latitude)$$

whereas dependent variables are the prevalence values vis-à-vis ascariasis, trichuriasis, and hookworm disease.

## 2.2 Extraneous Data

Table 2.2 outlines the project's extraneous variables, i.e., the extraneous independent variables.

Table 2.2: The WaSH, elevation, and population density data variables.

| variable | description |
|---|---|
| *improved_sewer, unimproved_sewer* | Percentage (a) *access to any improved sanitation facility*, and (b) *reliance on unimproved sanitation facilities*, respectively. (WaSH variables) |
| *piped_sewer, unpiped_sewer* | Percentage (a) *access to sewer and septic sanitation facilities*, and (b) *access to a non-piped improved sanitation facility*, respectively. (WaSH variables) |
| *surface_sewer* | Percentage *reliance on open defaecation*. (WaSH variable) |
| *elevation, elevation.km* | The elevation in metres and kilometre, respectively. |
| *p_density, p_density.k* | The number of people per square kilometre, and the number of thousand people per square kilometre, respectively. |

### *2.2.1 WaSH*

The WaSH (water, sanitation, and hygiene) variables values are extracts from **access percentage maps**. The maps are courtesy of the Institute for Health Metrics and Evaluation (IHME), and are due to a Lancet Local Burden of Disease WaSH Collaborators study (8); the access percentages are estimates. There are 18 maps per WaSH variable, one for each year spanning [2000  2017]. An ESPEN data set observation will have WaSH variable values if, and only if,

1. The observation has valid geographic coördinates, and
2. The WaSH variable map covers the area in question.

### *2.2.2 Elevation*

The elevation data is courtesy of WorldClim elevation maps (7), which are derivations of the Shuttle Radar Topography Mission's elevation maps; the **unit of measure is metres**. The elevation value of each examination location, that has verified geographic coördinate values, is an extract of the 30 seconds elevation map.

### *2.2.3 Population Density*

Finally, the project's population density values are derivations of the Gridded Population of the World population estimates; the **unit of measure is the number of people per square kilometre**. The `r` package `geodata` stores the population estimates maps.

The estimates are quinquennial estimates, thus far there are estimates maps for the years 2000, 2005, 2010, 2015, and 2020. This project extracts estimate population density values, per verified examination location coördinates, from these maps. Subsequently, cubic spline interpolation provides intervening years estimates.

# 3
# EXPLORATORY ANALYSIS

The background chapter notes that this project focuses on *Hookworm disease*, henceforth all data explorations, modelling, and analysis are in relation to *Hookworm disease*, and prevalence is the dependent/outcome variable in every case. In terms of data, the project's underlying data set is the Togo data set. It is the best data set vis-à-vis minimal missing points, missing data pattern, and the number of records after excluding incomplete records; the missing data pattern is *missing completely at random*.

## 3.1 In focus: Togo

The country of Togo lies within West Africa. Its bordering countries are the Republic of Ghana, the Republic of Benin, and Burkina Faso. The climate is a tropical wet dry climate (24); the far northern tip is semi-arid. Fig. 3.2 illustrates the elevation pattern across Togo, most of the country has an elevation value less than 400 metres. In general, most of Togo's citizens live in rural areas across the country (24), and the estimated population density across most of the country is within the range (0  $1k$] per square kilometre (Fig. 3.2).

In terms of *Hookworm disease*, the focus of this project, Table 3.1 outlines Togo's latest **(a)** disability adjusted life years (DALY) per $100k$ inhabitants, i.e., the number of years lost to disability, ill-health, or early death per $100k$ inhabitants, and **(b)** prevalence per $100k$ inhabitants (2, 3). The estimate of each metric is much higher than the global [average] estimate; cf. Table 3.1, Table 1.1, and Table 1.2. Over time Togo's prevalence per $100k$ inhabitants has peaked & troughed (3). The reason is not quite clear, but a mass drug administration study by Bronzan et al. (23) *"... found that in areas with high baseline prevalence of hookworm the risk of rebound of infection is high among children who do not receive bi-annual treatment."*; this hints at

a possible programme design problem.

| | | confidence interval | |
|---|---|---|---|
| metric | estimate | lower | upper |
| DALY/100k | 113.79 | 66.77 | 184.18 |
| prev./100k | 17,027.10 | 12,746.98 | 21,643.73 |

prev.: Prevalence

Table 3.1: The 2019 disability adjusted life years (DALY) per 100$k$ inhabitants, and prevalence per 100$k$ inhabitants. Data Source: Institute for Health Metrics & The Lancet Global Burden of Disease 2019

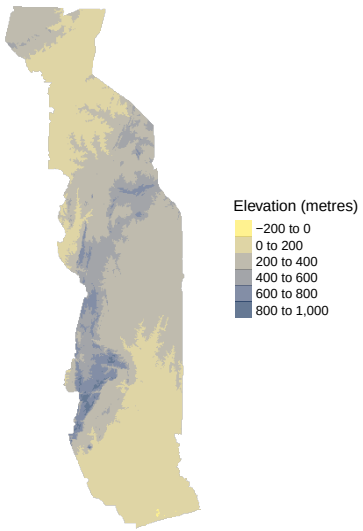


Figure 3.1: The elevation pattern across Togo. Data Source: WorldClim

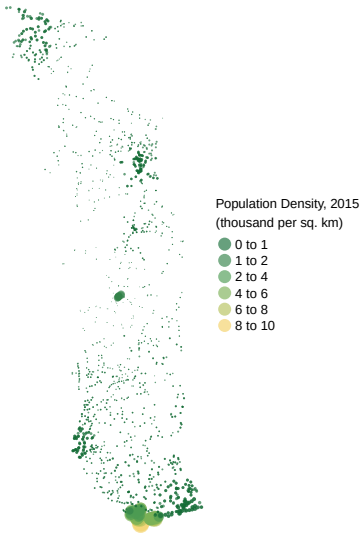


Figure 3.2: The 2015 population density pattern across Togo. Data Source: Gridded Population of the World

## 3.2 Geohelminths Infections Survey Experiments: Distributions of Sites & Prevalence

| year | number of sites |
|---|---|
| *2009* | 1033 |
| *2015* | 1047 |

Table 3.2: A tally of Togo's geohelminths infections survey experiments sites, i.e., records; per year a site is associated with a single record.

Fig. 3.3 illustrates the distribution of survey experiments sites across Togo; Table 3.2 summarises the number of sites per year. The data set's survey experiments occurred during the years 2009 & 2015. However, the years only have 4 survey experiments sites in common, hence the questions

- How similar or dissimilar are proximal survey experiment sites vis-à-vis probable risk factors?
- Is it possible to infer the 2009 prevalence of a site, which was only surveyed in 2015, via the prevalence of proximal 2009 locations - and vice versa?
- Are there sufficient data points for determining possible patterns of association between *prevalence* & *probable risk factors* - considering the disparate distribution of sites across the years?


Figure 3.3: The distribution of geohelminths infections survey experiments sites in Togo during the years 2009 & 2015; the years only have 4 sites in common.

These questions have a bearing on explanatory model development, and partly define the limitations of any developed models.

Fig. 3.4 illustrates the prevalence distributions during the years 2009 & 2015. It is tempting to state that *there have been improvements, over time, in relation to reducing prevalence*; alas, the statement is questionable because the years only have 4 sites in common.

Setting aside the disparate 2009 & 2015 sites, do the prevalence distributions reflect *generic patterns of prevalence distributions over-time due to interventions*? Unclear. Yes, the aim of interventions is the reduction of INI prevalence over time, but it is improbable that all interventions have been successful, and patterns of prevalence distribution over time probably depend on starting prevalence values.

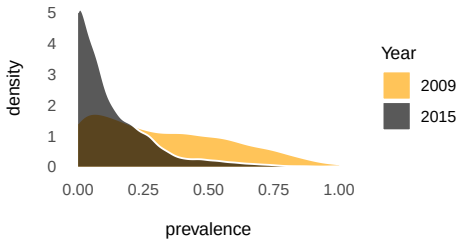


Figure 3.4: The distribution of geohelminths infections prevalence during the years 2009 and 2015

## 3.3 WaSH

Fig. 3.5 illustrates the relationship between *prevalence* & the access values of three WaSH (water, sanitation, and hygiene) variables: *piped sewer*, *unpiped sewer*, and *surface sewer*, i.e., open defaecation. Each graph's prevalence values are actually *empirical logit*(*prevalence*) values:

$$\textit{empirical logit}(\textit{prevalence}) = \frac{p + 0.5}{n + 0.5} \tag{3.1}$$

wherein

$p$ : # of inhabitants that test positive

$n$ : # of inhabitants that test negative

in acknowledgement of the fact that the project's focus is Binomial models. The fitted lines are for pattern investigation purposes, and the fitting functions, throughout this chapter, are **(a)** dashed lines: $y = mx + c$,[1] and **(b)** solid lines: cubic splines.

[1] $y = mx + c$ is the equation of a straight line wherein $m$ is the gradient, and $c$ is the intercept.

It is difficult to definitively state the relationship patterns of Fig. 3.5. The *piped sewer* graph exhibits the strongest relationship pattern - according to the fitting function $y = mx + c$, each year's prevalence decreases with increasing access to piped sewers.

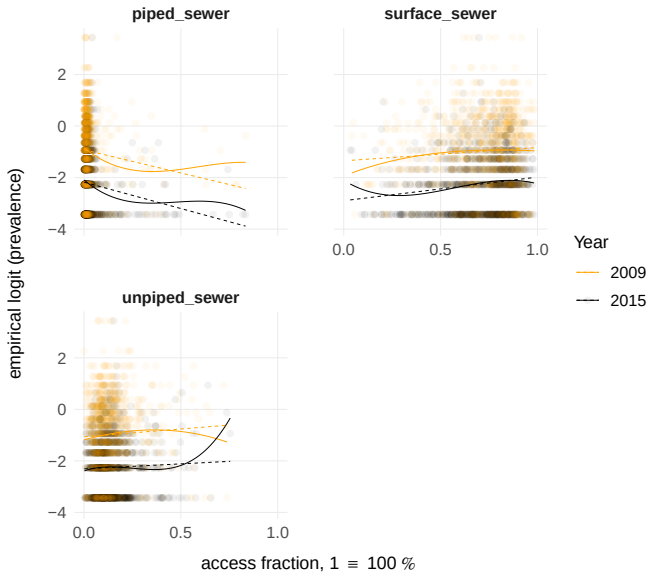


Figure 3.5: The relationships between *Hookworm disease* prevalence & access to sewage facilities; $1 \equiv 100\%$, which means 100% of the population has access. The fitted lines are for pattern observation purposes, and the fitting functions are $y = mx + c$ & *cubic splines*, dashed lines & solid lines, respectively.

## 3.4 Elevation

Fig. 3.6 illustrates the relationship between *prevalence* & *elevation*. A definitive relationship is not discernible. Each year, to varying degrees, the graph suggests an inverse relationship between prevalence and elevation.

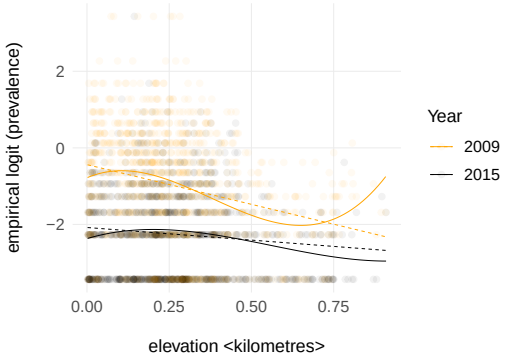


Figure 3.6: The relationships between *Hookworm disease* prevalence and elevation. The fitted lines are for pattern observation purposes, and the fitting functions are $y = mx + c$ & *cubic splines*, dashed lines & solid lines, respectively.

## 3.5 Population Density

Fig. 3.7 illustrates the relationship between *prevalence* & *population density*. Again, a definitive relationship is not discernible, but it is possible that prevalence is inversely proportional to population density.

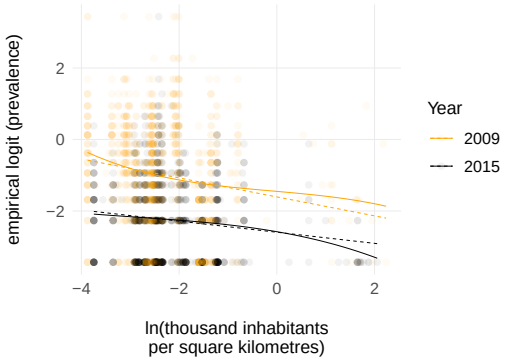


Figure 3.7: The relationships between *Hookworm disease* prevalence and population density. The fitted lines are for pattern observation purposes, and the fitting functions are $y = mx + c$ & *cubic splines*, dashed lines & solid lines, respectively.

# 4
# PRELIMINARY INVESTIGATION

The previous chapter graphically explores empirical relationships between *Hookworm disease* prevalence & a few extraneous independent variables, i.e., probable risk factors. This chapter explores plausible empirical relationships models. As noted before, this project's outcome of interest is [*Hookworm disease*] *prevalence*, which is a proportion value. Hence, all empirical relationship model investigations - between *prevalence* and *zero or more extraneous independent variables* - are **Binomial model** investigations.

In preparation for Binomial geostatistical modelling, the objective of this chapter is to

> *investigate evidence of residual spatial correlation in the residuals of preliminary empirical relationship models between* ***prevalence*** *and* ***zero or more extraneous independent variables****.*

Each *preliminary empirical relationship model*, henceforth *preliminary model*, is a Binomial generalised linear mixed model (GLMM). Evidence of residual spatial correlation, in a preliminary model's residuals, is determined via the empirical variogram; the empirical variogram function determines the degree of spatial correlation in a stochastic process or random field (5, 6). Note that the models of the geostatistical binomial logistic models chapter are derivations of preliminary models that exhibit residual spatial correlation, i.e., an implicit purpose of this chapter is the determination of plausible fixed term effects vectors for geo-statistical modelling.

Table 4.1 summarises this chapter's common notations.

| notation | description |
|---|---|
| $i, n$ | $i$ denotes the $i^{th}$ observation, and $n$ is the number of observations. Hence, $i = 1, \ldots, n$. |
| $x_i$ | The location of the $i^{th}$ observation |
| $\delta(x_i)$ | A vector of terms that depend on zero or more independent variables at location $x_i$ |
| $\beta$ | The vector of regression coefficients vis-à-vis $\delta(x_i)$ |
| $Y_i$ | The outcome at location $x_i$ |
| $p_i$ | The prevalence $\nu_i/e_i$ wherein $\nu$ is the # of positives, and $e_i$ is the # of examinations, at location $x_i$ |

Table 4.1: The notations used at various points of this chapter, and subsequent chapters.

## 4.1 A Generalised Linear Mixed Model with Multiple Random Effects

In general, the key assumptions of generalised linear mixed models are **(a)** the observations are independent, and **(b)** the values of each independent variable are error free. Initially, the preliminary model structure of interest is the GLMM

$$\begin{aligned} Y_i &\sim Binomial(e_i, p_i) \\ \eta_i &= ln\left(\frac{p_i}{1-p_i}\right) \\ &= \delta(x_i)^T\beta + b_i + c_i \end{aligned} \tag{4.1}$$

wherein

$$\begin{aligned} \mathbf{b} &\sim \mathcal{N}(0, \tau_b^2\mathbf{I}_b) \\ \mathbf{c} &\sim \mathcal{N}(0, \tau_c^2\mathbf{I}_c) \end{aligned} \tag{4.2}$$

It is a Binomial GLMM because the prevalence values are proportions. Table 4.1 defines most of the terms of the equations. Expression $\delta(x_i)^T\beta$ is the *fixed effects* expression; $\delta(x_i)$ is a vector of terms that depend on one or more independent variables, whilst $\beta$ is the vector of associated regression coefficients. The independent variables are one or more of the extraneous variables (Table 2.2).

Additionally, $b_i$ & $c_i$ are the random effects terms, in relation to *identifier* & *year*, respectively. Eq. 4.2 outlines the assumptions underlying the random effects vectors, i.e., each is

- Mutually independent, and identically distributed.
- Normally distributed.

The length of the diagonal of identity matrix $\mathbf{I}_b$ is the number of distinct locations, whilst the diagonal length of $\mathbf{I}_c$ is the number of distinct years.

### *4.1.1 Limitations*

There is a key challenge in relation to the GLMM of Eq. 4.1 and Togo's *Hookworm disease* data:

> *The residual spatial correlation tests of each preliminary model is inconclusive. In each case the design, i.e., fixed-effect, matrix is rank deficient due to insufficient data.*

The linear predictors $\eta_i$ of two of the preliminary models in question are

$$\eta_i = \beta_0 + \beta_1 piped\_sewer(x_i) + \beta_2 piped\_sewer(x_i)^2 + \beta_3 elevation.km(x_i) + b_i + c_i \tag{4.3}$$

and

$$\eta_i = \beta_0 + \beta_1 piped\_sewer(x_i) + \beta_2 ln(p\_density.k)(x_i) + \beta_3 elevation.km(x_i) + b_i + c_i \tag{4.4}$$

Each reflects the sometimes inconclusive relationships, between *prevalence* & *extraneous independent variables*, outlined within the exploratory analysis chapter (study Figs. 3.5 to 3.7).

In terms of the models' coefficients, in accordance with the hypotheses

$$H_0 : \beta_j = 0$$

$$H_1 : \beta_j \neq 0$$

there is *preliminary evidence* in favour of the alternative hypothesis, $H_1$, $\forall\ j$, at the significance level 0.05. However, the stability of these observations is uncertain.

Hence, **the project proceeds with the data set of a single year. Specifically, it uses the** 2015 **data of Togo.** Using Togo's 2009 data, rather than the 2015 data, means depending on population density estimates that are derivations of Gridded Population of the World quinquennial population density estimates, as outlined in the data chapter; at present, the quinquennial population estimates are for the years 2000, 2005, 2010, 2015, and 2020.

## 4.2 A Generalised Linear Mixed Model with a Single Random Effect

In the case of single year data sets, the GLMM becomes

$$
\begin{aligned}
Y_i &\sim Binomial(e_i, p_i) \\
\eta_i &= ln\left(\frac{p_i}{1-p_i}\right) \\
&= \delta(x_i)^T\beta + b_i
\end{aligned}
\tag{4.5}
$$

wherein

$$
\mathbf{b} \sim \mathcal{N}(0, \tau_b^2\mathbf{I}_b) \tag{4.6}
$$

The difference between this model and Eq. 4.1 is the absence of a *year* random effects term in this model. The definition of each term remains the same.

Leading on from the afore-stated linear predictors Eqs. 4.3 and 4.4, this section focuses on preliminary models, with linear predictors

$$
\begin{aligned}
\eta_i = \beta_0 &+ \beta_1 piped\_sewer(x_i) + \beta_2 piped\_sewer(x_i)^2 \\
&+ \beta_3 elevation.km(x_i) + b_i
\end{aligned}
\tag{4.7}
$$

and

$$
\begin{aligned}
\eta_i = \beta_0 &+ \beta_1 piped\_sewer(x_i) + \beta_2 ln(p\_density.k)(x_i) \\
&+ \beta_3 elevation.km(x_i) + b_i
\end{aligned}
\tag{4.8}
$$

i.e., linear predictors Eqs. 4.3 and 4.4 without *year* random effects terms.

### *4.2.1 The Estimated Coefficients*

Table 4.2 & Table 4.3 outline the coefficient estimates vis-à-vis the fixed terms of linear predictors Eq. 4.7 & Eq. 4.8, respectively. In each case, and in accordance with the hypotheses

$$H_0 : \beta_j = 0$$

$$H_1 : \beta_j \neq 0$$

there is evidence in favour of the alternative hypothesis, $H_1$, $\forall\ j$, at significance level 0.05; each coefficient estimate is significant at the 0.05 significance level.

| | | | confidence interval | | | |
|---|---|---|---|---|---|---|
| | coef. | coef. est. | lower | upper | s.e. | $p_{value}$ |
| *1* | $\beta_0$ | -1.947 | -2.247 | -1.646 | 0.153 | 0.000 |
| *piped_sewer* | $\beta_1$ | -11.313 | -17.179 | -5.446 | 2.993 | 0.000 |
| $piped\_sewer^2$ | $\beta_2$ | 15.757 | 5.343 | 26.172 | 5.314 | 0.003 |
| *elevation.km* | $\beta_3$ | -1.583 | -2.428 | -0.738 | 0.431 | 0.000 |

coef.: coefficient
est.: estimate
s.e.: standard error

Table 4.2: The estimated coefficients of the GLMM with linear predictor $\eta_i = \beta_0 + \beta_1 piped\_sewer(x_i) + \beta_2 piped\_sewer(x_i)^2 + \beta_3 elevation.km(x_i) + b_i$

| | | | confidence interval | | | |
|---|---|---|---|---|---|---|
| | coef. | coef. est. | lower | upper | s.e. | $p_{value}$ |
| *1* | $\beta_0$ | -2.518 | -2.993 | -2.043 | 0.242 | 0.000 |
| *piped_sewer* | $\beta_1$ | -3.393 | -6.278 | -0.509 | 1.472 | 0.021 |
| *log(p_density.k* | $\beta_2$ | -0.212 | -0.405 | -0.020 | 0.098 | 0.031 |
| *elevation.km* | $\beta_3$ | -1.898 | -2.794 | -1.002 | 0.457 | 0.000 |

coef.: coefficient
est.: estimate
s.e.: standard error

Table 4.3: The estimated coefficients of the GLMM with linear predictor $\eta_i = \beta_0 + \beta_1 piped\_sewer(x_i) + \beta_2 ln(p\_density.k)(x_i) + \beta_3 elevation.km(x_i) + b_i$

### *4.2.2 Residual Spatial Correlation*

Next, spatial correlation assessments of the residuals of the generalised linear mixed models with linear predictors Eq. 4.7 & Eq. 4.8. Their empirical variograms - Fig. 4.1 & Fig. 4.2, respectively - illustrate evidence of residual spatial correlation. Hence, proceeding to geostatistical binomial logistic modelling with the fixed terms of the linear predictors Eq. 4.7 & Eq. 4.8.

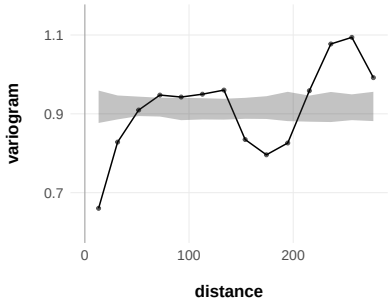


Figure 4.1: The empirical variogram of the GLMM with linear predictor $\eta_i = \beta_0 + \beta_1 piped\_sewer(x_i) + \beta_2 piped\_sewer(x_i)^2 + \beta_3 elevation.km(x_i) + b_i$

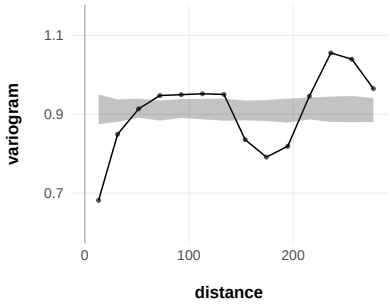


Figure 4.2: The empirical variogram of the GLMM with linear predictor $\eta_i = \beta_0 + \beta_1 piped\_sewer(x_i) + \beta_2 ln(p\_density.k)(x_i) + \beta_3 elevation.km(x_i) + b_i$

### *4.2.3 The Random Effects*

The two models, with linear predictors Eq. 4.7 & Eq. 4.8, have similar *estimated random intercept values* distributions, and the variances of these distributions are more or less equal.

**In the case of**

$$\eta_i = \beta_0 + \beta_1 piped\_sewer(x_i) + \beta_2 piped\_sewer(x_i)^2 + \beta_3 elevation.km(x_i) + b_i \quad (4.7)$$

the variance $\tau_b^2$ of the random effects vector $\mathbf{b}$ is 1.584. Fig. 4.3 is the quantile-quantile plot of the random intercept values $b_i$ — the distribution is approximately normal.

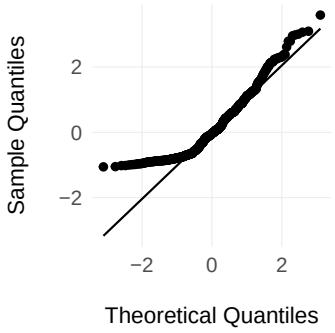


Figure 4.3: The quantile-quantile plot of the random intercept estimates w.r.t. the GLMM with linear predictor $\eta_i = \beta_0 + \beta_1 piped\_sewer(x_i) + \beta_2 piped\_sewer(x_i)^2 + \beta_3 elevation.km(x_i) + b_i$

**In the case of**

$$\eta_i = \beta_0 + \beta_1 piped\_sewer(x_i) + \beta_2 ln(p\_density.k)(x_i) + \beta_3 elevation.km(x_i) + b_i \quad (4.8)$$

the variance $\tau_b^2$ of the random effects vector $\mathbf{b}$ is 1.585. Fig. 4.3 is the quantile-quantile plot of the random intercept values $b_i$ — the distribution is approximately normal.

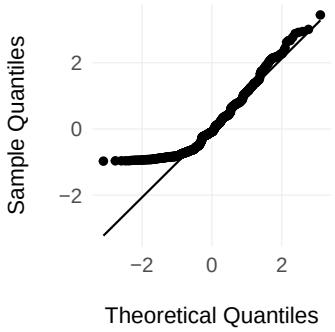


Figure 4.4: The quantile-quantile plot of the random intercept estimates w.r.t. the GLMM with linear predictor $\eta_i = \beta_0 + \beta_1 piped\_sewer(x_i) + \beta_2 ln(p\_density.k)(x_i) + \beta_3 elevation.km(x_i) + b_i$

# 5
# GEOSTATISTICAL BINOMIAL LOGISTIC MODELS

The predominant focus of this chapter is explanatory models. An explanatory model determines which independent variables have a statistically significant impact on a dependent variable in question, it also outlines the degree of impact of the independent variables. Hence, and in relation to the project's aim & objectives, the objective herein is to

> *Determine which extraneous independent variables - WaSH variables, population density, and elevation - have a statistically significant impact on the prevalence of Hookworm disease infections, and outline their degree of impact.*

Leading on from the data, exploratory analysis, preliminary investigation chapters, a few important points: **(a)** The outcomes - prevalence - are proportions, hence the project's focus is Binomial models. **(b)** The preliminary investigation identified Binomial generalised linear mixed models with strong evidence of residual spatial correlation, hence *this chapter focuses on geostatistical binomial logistic [explanatory] models.* **(c)** Modelling is via the data set of a single year; Togo, 2015.

## 5.1 Geo-statistical binomial logistic modelling

Here we explore *geostatistical binomial logistics models* wherein the parameter estimates are Monte Carlo Maximum Likelihood estimates. The model structure is

$$Y_i \sim Binomial(e(x_i), p(x_i)) \tag{5.1}$$

$$\eta(x_i) = ln\left(\frac{p(x_i)}{1-p(x_i)}\right) \tag{5.2}$$

$$= \delta(x_i)^T\beta + S(x_i) + U_i \tag{5.3}$$

Again, Table 4.1 defines most terms. $x_i$ is the spatial/geographic location of the $i^{th}$ observation. $\delta(x_i)$ is the $i^{th}$ vector of fixed terms - i.e., the $i^{th}$ row vector of the model's design matrix - it depends on zero or more independent variables. Whilst $\beta$ is the vector of the concurrent fixed effects coefficients. $S(x_i)$ is the stochastic process term, it models the *unexplained spatial variation* (9). Herein, the assumption is that $S(\mathbf{x})$ is

- A zero-mean gaussian process, i.e., the joint probability of $S(x_1), \ldots, S(x_n)$ is multi-variate normal.
- Stationary, i.e., it has a constant variance - denoted $\sigma^2$.
- Isotropic, i.e., the correlation between $S(x_i)$ & $S(x_j)$ depends on the distance between $x_i$ & $x_j$ only.

A consequence of these assumptions is that the covariance $\gamma(x_i, x_j)$ is defined as

$$\gamma(x_i, x_j) = \sigma^2 \rho(\|x_i - x_j\|) \tag{5.4}$$

wherein $\rho(u) = \rho(\|x_i - x_j\|)$ is the correlation function; $u = \|\cdot\|$ denotes distance. Throughout this paper correlation function in use is the *exponential correlation function* (9)

$$\rho(u; \phi) = e^{-u/\phi}, \qquad u \geq 0 \tag{5.5}$$

The term $\phi$ is the scale parameter.

In contrast to $S(x_i)$, $U_i$ are independent Normally distributed random variables

$$U_i \sim \mathcal{N}(0, \tau^2) \tag{5.6}$$

that capture the effects of unknown/unmeasured independent variables, i.e., $U_i$ represent the *unstructured variation in* $Y_i$ (9). The unknown variables either **(a)** lack a spatial structure, or **(b)** *"… are spatially varying but on a scale smaller than the minimum observed distance.''* (9)

### *5.1.1 Linear Predictors*

The preliminary investigation chapter identifies two generalised linear mixed models whose **(a)** residuals exhibit spatial correlation, and **(b)** coefficient estimates $\beta_j$ are significant $\forall\ j$, at significance level 0.05. The linear predictors of those models are

$$\begin{aligned}\eta_i = {} & \beta_0 + \beta_1 piped\_sewer(x_i) + \beta_2 piped\_sewer(x_i)^2 \\ & + \beta_3 elevation.km(x_i) + b_i\end{aligned}$$

and

$$\eta_i = \beta_0 + \beta_1 piped\_sewer(x_i) + \beta_2 ln(p\_density.k)(x_i) + \beta_3 elevation.km(x_i) + b_i$$

Each linear predictor's fixed-term expression is the olive coloured part of the linear predictor. Hence, and building on these preliminary investigation fixed-term expressions, the geostatistical binomial logistic models in focus are the models with linear predictors

$$\eta_i = \beta_0 + \beta_1 piped\_sewer(x_i) + \beta_2 piped\_sewer(x_i)^2 + \beta_3 elevation.km(x_i) + S(x_i) + U_i \tag{5.7}$$

and

$$\eta_i = \beta_0 + \beta_1 piped\_sewer(x_i) + \beta_2 ln(p\_density.k)(x_i) + \beta_3 elevation.km(x_i) + S(x_i) + U_i \tag{5.8}$$

The next few sections discuss the models' estimates, and whether the models successfully capture **(a)** unexplained spatial variation, and **(b)** unstructured variation in $Y_i$.

## 5.2 The Models

In this section all parameter estimates are Monte Carlo Maximum Likelihood estimates.

### *5.2.1 The Estimated Coefficients*

Table 5.1 & Table 5.2 outline the coefficient estimates vis-à-vis the fixed-terms of linear predictors Eq. 5.7 & Eq. 5.8, respectively. By the hypotheses

$$H_0 : \beta_j = 0$$

$$H_1 : \beta_j \neq 0$$

each fixed-term coefficient estimate of Table 5.1 is statistically significant at the 0.05 significance level, i.e., there is evidence in favour of the alternative hypothesis, $H_1$, $\forall\ j$, at significance level 0.05. These Table 5.1 observations are in-line with the preliminary investigation's estimated coefficients observations.

Table 5.1: The estimated coefficients of the Binomial geostatistical logistic model with linear predictor $\beta_0 + \beta_1 piped\_sewer(x_i) + \beta_2 piped\_sewer(x_i)^2 + \beta_3 elevation.km(x_i) + S(x_i) + U_i$

| | coef. | coef. est. | confidence interval lower | confidence interval upper | s.e. | $p_{value}$ |
|---|---|---|---|---|---|---|
| *1* | $\beta_0$ | -1.957 | -2.402 | -1.512 | 0.227 | 0 |
| *piped_sewer* | $\beta_1$ | -7.605 | -11.380 | -3.829 | 1.926 | 0 |
| *I(piped_sewer$^2$)* | $\beta_2$ | 15.766 | 9.394 | 22.138 | 3.251 | 0 |
| *elevation.km* | $\beta_3$ | -2.152 | -3.191 | -1.114 | 0.530 | 0 |

coef.: coefficient
est.: estimate
s.e.: standard error

Table 5.2: The estimated coefficients of the Binomial geostatistical logistic model with linear predictor $\beta_0 + \beta_1 piped\_sewer(x_i) + \beta_2 ln(p\_density.k)(x_i) + \beta_3 elevation.km(x_i) + S(x_i) + U_i$

| | coef. | coef. est. | confidence interval lower | confidence interval upper | s.e. | $p_{value}$ |
|---|---|---|---|---|---|---|
| *1* | $\beta_0$ | -2.238 | -2.852 | -1.625 | 0.313 | 0.000 |
| *piped_sewer* | $\beta_1$ | 0.784 | -1.209 | 2.776 | 1.016 | 0.441 |
| *log(p_density.k)* | $\beta_2$ | -0.018 | -0.215 | 0.178 | 0.100 | 0.856 |
| *elevation.km* | $\beta_3$ | -1.843 | -2.856 | -0.829 | 0.517 | 0.000 |

coef.: coefficient
est.: estimate
s.e.: standard error

In contrast, not all the fixed-term coefficient estimates of Table 5.2 are significant; the coefficients of the *piped_sewer* & $ln(p\_density.k)$ terms

are not statistically significant at the 0.05 significance level. **Hence, the rest of this chapter** discusses the estimates of the Binomial geostatistical logistic model with linear predictor Eq. 5.7.

### *5.2.2 The Impact of the Estimated Coefficients*

A variable's coefficient estimate is a direct, or indirect, measure of the variable's impact on an outcome. Focusing on the fixed-term effects of Eq. 5.7

$$\begin{aligned}\eta_i &= \beta_0 + \beta_1 piped\_sewer(x_i) + \beta_2 piped\_sewer(x_i)^2 \\ &\quad + \beta_3 elevation.km(x_i) + S(x_i) + U_i\end{aligned}$$

$$\Rightarrow$$

$$\frac{p(x_i)}{1-p(x_i)} \propto e^{\beta_0} e^{\beta_1 piped\_sewer(x_i)} e^{\beta_2 piped\_sewer(x_i)^2} e^{\beta_3 elevation.km(x_i)} \tag{5.9}$$

by Eqs. 5.1 to 5.3 and 5.7. The effects *piped_sewer* and *elevation.km* on prevalence are deducible via Eq. 5.9.

#### Piped Sewage Facilities

Equating the derivative of Eq. 5.9 w.r.t. (with respect to) *piped_sewer* to zero leads to

$$-7.605 + \big(2 \times 15.766 \times piped\_sewer\big) = 0 \tag{5.10}$$

This suggests that the effect of *piped_sewer* on *Hookworm disease* infections prevalence varies about the inflection point $7.605/31.532 = 0.241$. As access to piped sewage facilities increases from 0 to 0.241 the prevalence odds decrease. Beyond the inflection point, however, the prevalence odds increase as access to piped sewage facilities increases; the range is (0.241  1]. The increasing prevalence odds after the inflection point might due to a mix of

- access density
- poor maintenance

The measure *piped_sewer* denotes the percentage of an area's inhabitants that have access to piped sewage facilities. The measure does not give an insight into access density, i.e., the median number of people per piped [defaecation] facility. Defaecation facilities that have high access densities require robust *hygiene maintaining cleaning schedules*, which might not be budgeted for. Consequently, potential facility users might sometime opt for options detrimental to public health.

Elevation

For every unit increase in elevation, i.e., every kilometre increase in elevation, the prevalence odds declines

$$100 \times \left(e^{-2.152} - 1\right) = -88.375 \tag{5.11}$$

i.e., decreases by 88.375%.

### *5.2.3 The Variance & Scale Estimates*

Table 5.3 outlines the variance, and scale parameter, estimates of linear predictor Eq. 5.7. The confidence interval of the scale parameter $\phi$ is rather wide. On the other hand, $\sigma^2$, the variance of the stochastic process $S(\mathbf{x})$, has a rather narrow confidence interval.

| | | confidence interval | | | confidence interval | |
|---|---|---|---|---|---|---|
| | est. | lower | upper | $e^{est.}$ | lower | upper |
| $ln(\sigma^2)$ | 0.310 | 0.028 | 0.593 | 1.364 | 1.028 | 1.810 |
| $ln(\phi)$ | 2.838 | 2.394 | 3.282 | 17.083 | 10.955 | 26.640 |
| $ln(\tau^2)$ | -1.377 | -2.277 | -0.477 | 0.252 | 0.103 | 0.621 |

est.: estimate

Table 5.3: The Binomial geostatistical logistic model variance estimates vis-à-vis $\beta_0 + \beta_1 piped\_sewer(x_i) + \beta_2 piped\_sewer(x_i)^2 + \beta_3 elevation.km(x_i) + S(x_i) + U_i$. $\kappa = 0.5$

### *5.2.4 The Distribution of the Random Effects*

Fig. 5.1 is the Normal quantile-quantile plot of the medians of the random effects samples w.r.t. Eq. 5.7. According to the plot, the medians of the random effects samples have a Normal distribution, line with model definition expectations.

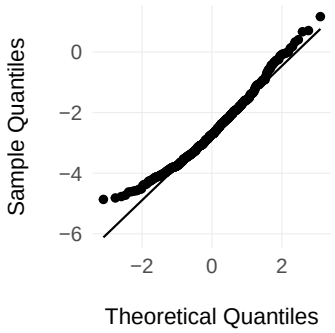


Figure 5.1: The Normal quantile-quantile plot of the medians of the random effects samples w.r.t. $\beta_0 + \beta_1 piped\_sewer(x_i) + \beta_2 piped\_sewer(x_i)^2 + \beta_3 elevation.km(x_i) + S(x_i) + U_i$. $\kappa = 0.5$

### *5.2.5 Residual Spatial Correlation*

The purpose of the Binomial geostatistical logistic model's [linear predictor] stochastic term $S(x_i)$ is the modelling/capturing of *unexplained spatial variation* (9). If the $S(x_i)$ term of Eq. 5.7 successfully models the unexplained spatial variation, then the associated standardised residual errors will be free of residual spatial correlation. As before, we test for residual spatial correlation via the empirical variogram. Foremost, the calculation of the standardised residual errors.

The standardised residual errors are

$$r_i = \frac{1}{\psi}\epsilon_i \left(1 - h_i\right)^{-1/2}, \quad \textit{wherein} \tag{5.12}$$

$$\psi^2 = \frac{1}{n-\rho}\sum_{i=1}^{n}\epsilon_i^2 \tag{5.13}$$

The number of observations is $n$, $\rho$ is the number of fixed effects parameters, i.e., the number of model coefficients. The raw residuals are $\epsilon_1, \ldots, \epsilon_n$; each $\epsilon_i$ is the difference between observed outcome $Y_i$ and its fitted value. $\psi$ is the standard error of the raw residuals. Finally, $h_i$ is the leverage of the $i^{th}$ observation. The definition of leverage value is

$$\mathbf{h} = diag\Big(X\left(X^T X\right)^{-1} X^T\Big) \tag{5.14}$$

wherein $X$ denotes a model's design matrix.

Fig. 5.2 outlines the empirical variogram of the standardised residual errors of the Binomial geostatistical logistic model with linear predictor Eq. 5.7. The empirical variogram does not exhibit residual spatial correlation, i.e., the Binomial geostatistical logistic model in question successfully models the unexplained spatial variation.

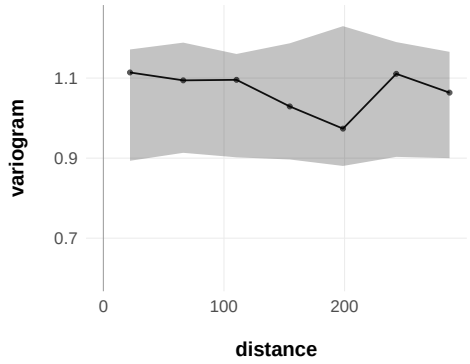


Figure 5.2: The variogram of the Binomial geostatistical logistic model with linear predictor $\beta_0 + \beta_1 piped\_sewer(x_i) + \beta_2 piped\_sewer(x_i)^2 + \beta_3 elevation.km(x_i) + S(x_i) + U_i$.

### *5.2.6 Prevalence Estimates*

The project set aside a portion of Togo's 2015 data set for preliminary prediction analysis purposes. Thus far, the portion of the Togo 2015 data set in use is, more or less, the *training* data set.

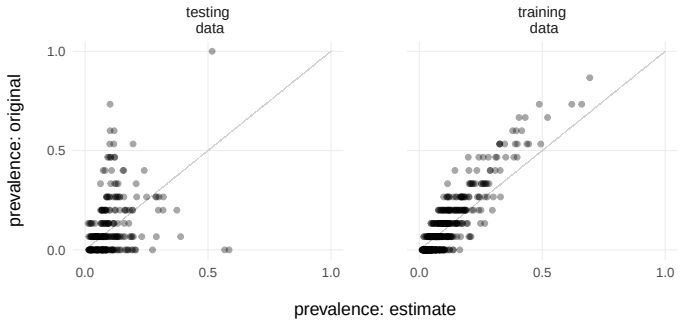


Figure 5.3: The prevalence estimates and originals vis-à-vis the Binomial geostatistical logistic model with linear predictor $\beta_0 + \beta_1 piped\_sewer(x_i) + \beta_2 piped\_sewer(x_i)^2 + \beta_3 elevation.km(x_i) + S(x_i) + U_i$.

The training data's original & estimated prevalence values are not quite equivalent. In general, the estimates are under-estimates, hence the negative bias value in Table 5.4. Hence, this model requires further adjustment. An adjustment option is additional risk factors; potential factors include land surface temperature, enhanced vegetation index values, and soil moisture (4).

Unsurprisingly, the prediction performance illustrated by the testing data set is poor. The predominant focus is explanatory modelling, with a special focus on WaSH variables; not all possible variables that might impact geohelminths infections prevalence. Sometimes the range of factors considered during explanatory modelling are sufficient for prediction purposes, sometimes they are not.

Interestingly, Brooker et al. (4) report moderate accuracy w.r.t. the prediction of *Hookworm disease* infections prevalence via binomial logistic regression analysis. Therein the independent variables are land surface temperature, normalized difference vegetation index, and elevation.

| | Bias | RMSE |
|---|---|---|
| *training* | -0.006 | 0.071 |
| *testing* | -0.014 | 0.146 |

Table 5.4: The bias and root mean square error values vis-à-vis the Binomial geostatistical logistic model with linear predictor $\beta_0 + \beta_1 piped\_sewer(x_i) + \beta_2 piped\_sewer(x_i)^2 + \beta_3 elevation.km(x_i) + S(x_i) + U_i$.

# 6
# DISCUSSION

This project studies the association between WaSH variables and geohelminths infections prevalence. It especially focuses on *Hookwookworm* disease prevalence, and investigates whether one or more WaSH variables are geohelminths infections prevalence risk factors. The investigation is via explanatory Binomial geostatistical logistic models.

The resulting model suggests that *piped sewage facilities access* is a statistically significant risk factor. The model illustrates a non-linear relationship between *piped sewage facilities access* and *Hookworm disease infections prevalence*. As the percentage of inhabitants that have access to piped sewage facilities increases from 0 to approximately 24.12% *Hookworm disease* prevalence odds decline. However, beyond the 24.12% access point *Hookworm disease* prevalence odds start increasing with increasing access percentage.

The increasing prevalence odds after the 24.12% inflection point might be due to a mix of *access density* & *poor maintenance*. The piped sewage facilities access percentage is the percentage of an area's inhabitants that have access to piped sewage facilities. The measure does not give an insight into access density, i.e., the median number of people per piped [defaecation] facility. Defaecation facilities that have high access densities might not be well-maintained.

## 6.1 Assumptions

The key/implicit assumptions:

- The ESPEN (Expanded Special Project for Elimination Neglected Tropical Diseases) survey measures, when available, are error free.
- If an ESPEN record's *number of survey examinations* value is zero or null, then the examinations' data is missing.
- If an ESPEN record's *number of positive tests* value is null, then the positives test data is missing.

- The WaSH access percentage estimates, and the population density estimates, are approximately accurate.

## 6.2 Limitations

The assumptions hint at the project's limitations. The data sets of a number of countries have missing data cells; the critical missing data being geographic coördinates, year, the number of examinations, and the number of positive tests. In the end, instead of a place agnostic model, the project's models use the data of a single country: Togo. Extending to a place agnostic explanatory model for, e.g., every country affected by *Hookworm disease*, means using a data set with geographic coördinates spanning these countries.

It is quite possible that the accuracy of the WaSH access percentage estimates, and of the population density estimates, varies between and within countries. Hence, the findings herein should be re-explored via one or more alternative estimates.

## 6.3 Validity

Finally, the project's validity (17) can be understood via

- **Internal Validity**: The project's focus is not the determination of a cause & effect relationship, hence an internal validity valuation is inapplicable.
- **Construct Validity**: Each explanatory model determines **(a)** whether there is a statistically significant association between an extraneous independent variable and *Hookworm disease* infections prevalence, and subsequently **(b)** the variable's impact. This is in line with the project's aim & objectives, i.e., the operational setting does reflect the construct.
- **External Validity**: At present, the study's conclusions can be generalised to Togo only.
- **Conclusion Validity**: Each of the resulting model's MCML (Monte Carlo Maximum Likelihood) coefficient estimates is statistically significant. Hence, it is reasonable to conclude that *piped_sewer* is probably a risk factor. The case would be stronger in the case of a *temporal Binomial geostatistical logistic model*.

# REFERENCES


1. A. Montresor, Helminth control in school-age children. (2011). https://www.who.int/publications/i/item/9789241548267.
2. T. Lancet, The global burden of disease study 2019. **396** (2020) 1129–1306. https://www.thelancet.com/journals/lancet/issue/vol396no10258/PIIS0140-6736(20)X0042-0.
3. I. for Health Metrics, GBD compare data visualization. (2020). https://vizhub.healthdata.org/gbd-compare/.
4. S. Brooker, A. C. A. Clements, & D. A. P. Bundy, Global epidemiology, ecology and control of soil-transmitted helminth infections. In S.I. Hay, A. Graham, & D.J. Rogers,eds., *Global mapping of infectious diseases: Methods, examples and emerging applications* (Academic Press, 2006), pp. 221–261. https://doi.org/https://doi.org/10.1016/S0065-308X(05)62007-6.
5. N. Cressie, *Statistics for spatial data* (John Wiley & Sons, Inc.).
6. G. Matheron, Principles of geostatistics. *Economic Geology*, **58** (1963) 1246–1266. https://doi.org/10.2113/gsecongeo.58.8.1246.
7. S. Fick & R. Hijmans, WorldClim 2: New 1-km spatial resolution climate surfaces for global land areas. *International Journal of Climatology*, **37** (2017) 4302–4315. https://doi.org/10.1002/joc.5086.
8. A. Deshpande, M. K. Miller-Petrie, P. A. Lindstedt, & et al., Mapping geographical inequalities in access to drinking water and sanitation facilities in low-income and middle-income countries, 2000–17. *The Lancet Global Health*, **8** (2020) e1162–e1185. https://doi.org/https://doi.org/10.1016/S2214-109X(20)30278-3.
9. P. J. Diggle & E. Giorgi, *Model-based geostatistics for global public health: Methods and applications* (Chapman; Hall/CRC Press, 2019).
10. C. J. L. Murray, A. Y. Aravkin, P. Zheng, C. Abbafati, & et al., Global burden of 87 risk factors in 204 countries and territories, 1990–2019: A systematic analysis for the global burden of disease study 2019. *The Lancet*, **396** (2020) 1223–1249. https://doi.org/https://doi.org/10.1016/S0140-6736(20)30752-2.

11. C. A. A. P. Mascarini-Serra Luciene Maura AND Telles, Reductions in the prevalence and incidence of geohelminth infections following a city-wide sanitation program in a brazilian urban centre. *PLOS Neglected Tropical Diseases*, **4** (2010) 1–7. https://doi.org/10.1371/journal.pntd.0000588.
12. J. Bethony, S. Brooker, M. Albonico, S. M. Geiger, A. Loukas, D. Diemert, & P. J. Hotez, Soil-transmitted helminth infections: Ascariasis, trichuriasis, and hookworm. *The Lancet*, **367** (2006) 1521–1532. https://doi.org/https://doi.org/10.1016/S0140-6736(06)68653-4.
13. M. Albonico, A. Montresor, D. W. T. Crompton, & L. Savioli, Intervention for the control of soil-transmitted helminthiasis in the community. In D.H. Molyneux,ed., *Control of human parasitic diseases* (Academic Press, 2006), pp. 311–348. https://doi.org/https://doi.org/10.1016/S0065-308X(05)61008-1.
14. G. Norman, S. Pedley, & B. Takkouche, Effects of sewerage on diarrhoea and enteric infections: A systematic review and meta-analysis. *The Lancet Infectious Diseases*, **10** (2010) 536–544. https://doi.org/https://doi.org/10.1016/S1473-3099(10)70123-7.
15. E. W. Steyerberg, *Clinical prediction models: A practical approach to development, validation, and updating* (Springer, 2010). https://doi.org/10.1007/978-1-4419-2648-7.
16. R. Little & D. B. Rubin, *Statistical analysis with missing data, third edition* (Wiley, 2019). https://doi.org/10.1002/9781119482260.
17. W. M. Trochim & J. P. Donnelly, *Research methods knowledge base* (Atomic Dog, 2006).
18. D. B. Rubin, Inference and missing data. *Biometrika*, **63** (1976) 581–592. https://doi.org/10.1093/biomet/63.3.581.
19. W. Garn JV & M. Freeman, Interventions to improve water, sanitation, and hygiene for preventing soil-transmitted helminth infection. *Cochrane Database of Systematic Reviews*, (2022). https://doi.org/10.1002/14651858.CD012199.pub2.
20. M. C. Freeman, J. V. Garn, G. D. Sclar, S. Boisson, K. Medlicott, K. T. Alexander, G. Penakalapati, D. Anderson, A. G. Mahtani, J. E. T. Grimes, E. A. Rehfuess, & T. F. Clasen, The impact of sanitation on infectious disease and nutritional status: A systematic review and meta-analysis. *International Journal of Hygiene and Environmental Health*, **220** (2017) 928–949. https://doi.org/https://doi.org/10.1016/j.ijheh.2017.05.007.
21. D. G. A. S. Strunz Eric C. AND Addiss, Water, sanitation, hygiene, and soil-transmitted helminth infection: A systematic review and meta-analysis. *PLOS Medicine*, **11** (2014) 1–38. https://doi.org/10.1371/journal.pmed.1001620.

22. S. Brooker & E. Michael, The potential of geographical information systems and remote sensing in the epidemiology and control of human helminth infections. *Remote sensing and geographical information systems in epidemiology* (Academic Press, 2000), pp. 245–288. https://doi.org/https://doi.org/10.1016/S0065-308X(00)47011-9.
23. A. M. A. A. Bronzan Rachel N. AND Dorkenoo, Impact of community-based integrated mass drug administration on schistosomiasis and soil-transmitted helminth prevalence in togo. *PLOS Neglected Tropical Diseases*, **12** (2018) 1–23. https://doi.org/10.1371/journal.pntd.0006551.
24. CIA, The world factbook. (2022). https://www.cia.gov/the-world-factbook/countries/togo/.

# *A*
# PROJECT SCOPE

**Project host**: World Health Organization (WHO)
**World Health Organization supervisor**: Professor Antonio Montresor
**Lancaster Medical School/**CHICAS **supervisor**: Dr. Emanuele Giorgi

Soil transmitted helminths (STH), i.e., geohelminths, are parasitic intestinal worms that cause intestinal nematode infections such as ascariasis, trichuriasis, and hookworm disease. According to the Lancet's latest global burden of disease study (2, 3), soil transmitted helminths infections are the **(a)** second most burdensome neglected tropical diseases, after dengue, and **(b)** most prevalent. Therefore, geohelminths infections are a focus of the World Health Organization (WHO). Geohelminths infections prevention, control, and elimination activities depend on knowledge of prevalence (1). Prevalence metrics depend on field surveys, which are rather expensive to conduct. They are especially expensive when there are no measures to systematically/strategically direct field scientists to survey areas. In the absence of systematic/strategic direction, the field scientists have to survey as wide an area as possible, albeit subject to financial, accessibility, etc., constraints.

A key STH prevalence hypothesis, of domain experts, is that WASH (water, sanitation, and hygiene) variables might be critical prognostic factors of STH prevalence. Therefore, the project will focus on understanding the association between STH prevalence and WASH variables, and investigate whether they are prognostic factors of STH prevalence.

## A.1 Project Aim

Determine whether one or more WaSH (water, sanitation, and hygiene) variables are geohelminths infections prevalence risk factors.

## A.2 Project Objectives

Hence, the project's objectives are to **(a)** determine which WaSH (water, sanitation, and hygiene) variables have a statistically significant impact on geohelminths infections prevalence, and **(b)** determine, estimate, the degree to which impacting WaSH variables affect geohelminths infections prevalence — via geostatistical models.

## A.3 Project Data

The project will rely on the

- The soil transmitted helminths' data sets of ESPEN (Expanded Special Project for Elimination of Neglected Tropical Diseases) – **for** ascariasis, hookworm infection, and trichuriasis examinations and cases data per site.
- The estimated WASH variables of IHME (Institute for Health Metrics and Evaluation) – **for** determining site level WASH variables values. The ESPEN estimates, which depend on the IHME methodology, are implementation level estimates.
- The National Oceanic & Atmospheric Administration (NOAA), WorldClim, and DIVA GIS – **for** historical geospatial and/or environmental variables values.

## A.4 Deliverables

The deliverables are

- Geo-statistical STH prevalence modelling & analysis, via GitHub.
- Thesis manuscript.
- Datasheet.

## A.5 Timeline

| | Item |
|---|---|
| 2022/06/20 - 2022/07/08 | Addressing site level records identification issues (refer to the quality constraints section further below). The derivation of site level variable values; WASH and environmental variables values. |
| 2022/07/12 - 2022/07/30 | Exploratory data analysis.<br>Exploration of geospatial binomial logistic models. |
| 2022/08/01 - 2022/08/21 | Final modelling stage; two models at most.<br>Model validation.<br>Model testing. |
| 2022/08/22 - 2022/09/06 | Manuscript writing. |
| 2022/09/08 - 2022/09/11 | Finalise modelling and analysis, repository and data sheet. |
| 2022/09/11 - 2022/09/15 | Poster preparation. |

## A.6 Out of Scope

The prediction of STH Prevalence for countries outwith the continent of Africa because the ESPEN data project focuses on African countries only, i.e., it only has the data of a set of African countries. Consequently, model development will be via the data of one or more African countries. Additionally, model validation, internal & external, will be via ESPEN countries only.

## A.7 Project Assumptions

Key assumptions:

- The soil transmitted helminth parasites of concern are roundworms (Ascaris lumbricoides), hookworms (Ancylostoma duodenale, Necator americanus), and whipworms (Trichuris trichiura).
- In terms of data granularity, the focus of this project is **site level prevalence**, and hence site level measures.
- The historical site level positive cases of ascariasis, hookworm infection, and trichuriasis, detailed by ESPEN, are dependable.

## A.8 Project Constraints

| | Date |
|---|---|
| Project start date | 6 June 2022 |
| Project end date | 9 September 2022 |
| Deadlines | Masters Thesis: 9 September 2022<br>Poster Session: 16 September 2022<br>Model Repository: 12 September 2022<br>Data Sheet: 12 September 2022 |

### *A.8.1* Time Constraints

The final week of the masters modules overlaps with the first week of the masters project, therefore only a third of the week ending 10 June 2022 shall be spent on the masters project. Future planning may occasionally occur during week days, the time will be recovered during the evenings and/or weekends.

### *A.8.2* Budget Constraints

The project has no budget, i.e., no budget/payment for personnel or computational resources used; the student bears the cost of conducting the project.

### *A.8.3* Quality Constraints

The site level data of ESPEN has 3 core issues. Foremost, and per country, many records do not have a site level identification code. Second, records that have the same coordinates, do not usually have the same site identification code. Finally, records associated with the same site sometimes have slightly different coordinate values. Prior to exploratory analysis & modelling, these discrepancies will have to be addressed, i.e., **(a)** each distinct site must have a unique site identifier and distinct coordinate values, and **(b)** records associated with the same site must have the same unique site identifier & coordinate values.

### *A.8.4* Equipment Constraints

No equipment has been provided for this project. The project's tasks will be, are, conducted via a personal computer. However, it is quite probable that the computer will stop working any time soon; the computer is quite old, and in recent times bits & pieces have stopped

working. Hence, access to a fall-back laptop is critical.