

REMOTE SENSING OF CLIMATE VARIABILITY AND CHOLERA IN SUB-SAHARAN AFRICA; SPATIAL ASSESSMENT

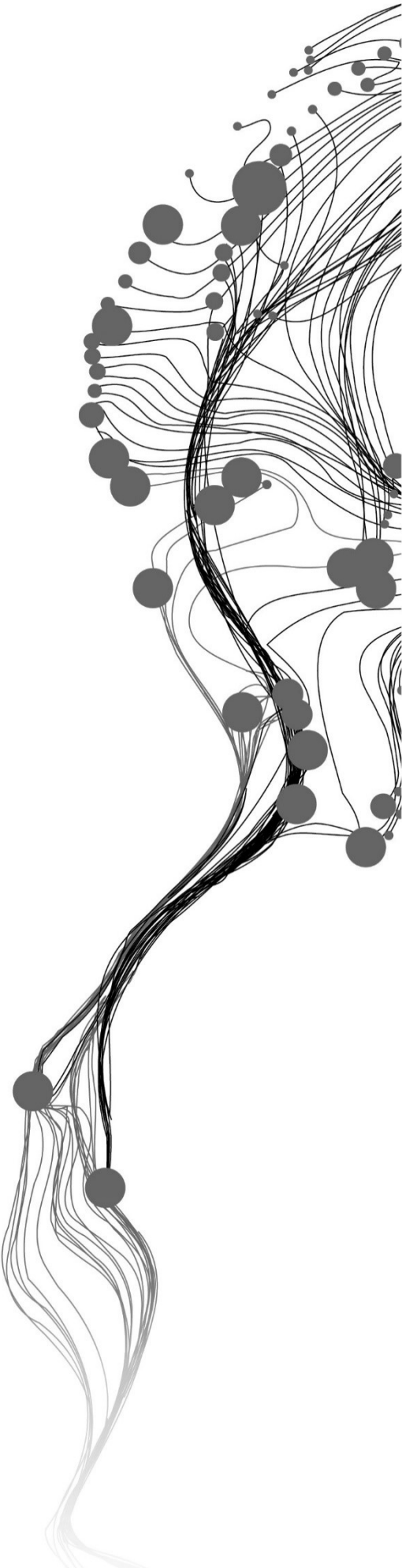
GETINET TAFESE TUCHO

March, 2019

SUPERVISORS:

Dr. Frank Osei

Dr. Peng Jia



REMOTE SENSING OF CLIMATE VARIABILITY AND CHOLERA IN SUB- SAHARAN AFRICA; SPATIAL ASSESSMENT

GETINET TAFESE TUCHO

Enschede, the Netherlands, March 2019

Thesis submitted to the Faculty of Geo-Information Science and Earth Observation of the University of Twente in partial fulfilment of the requirements for the degree of Master of Science in Geo-information Science and Earth Observation.

Specialization: Geoinformatics

SUPERVISORS:

Dr. Frank Osei

Dr. Peng Jia

THESIS ASSESSMENT BOARD:

Prof.dr.ir. A. Stein (Chairman)

Advisor: drs. J.P.G. Bakx (Advisor)

Dr. E.J.J. Rood (External Examiner)

DISCLAIMER

This document describes work undertaken as part of a programme of study at the Faculty of Geo-Information Science and Earth Observation of the University of Twente. All views and opinions expressed therein remain the sole responsibility of the author, and do not necessarily represent those of the Faculty.

ABSTRACT

Cholera is a severe endemic disease in Sub-Saharan Africa which is associated with climate variability related potential covariates. Cholera incidence can be influenced by socio-economic factors for instance; inadequate public infrastructure, poor cultural practice and lack of communal awareness about the disease incidence and its risk. Thus, investigation of the possible risk factors of the disease would be of great importance, especially for public health authorities to design intervention strategies to mitigate the disease risk. Geostatistical analysis methods are essential tools to model disease spatial variability. Generalized linear model (GLM) is a statistical model that allows response variables which have a distribution other than the normal distribution with the assumption of Poisson distribution in which mean and variance are equal. The objective of this particular study was to investigate impacts of climate variability and potential environmental covariates that have a relatively strong association with the annual count of cholera cases in Sub-Saharan Africa using a Poisson generalized linear model (GLM).

KEYWORDS: Cholera, Sub-Saharan, Climate-variability, Spatial epidemiology, Spatial regression, Regression kriging

ACKNOWLEDGMENTS

First of all, I thank the Almighty Lord God, Jesus Christ, for all the wonderful protection, mercy, grace and the amazing works of his hands in my life, without whom I wouldn't have been here. I acknowledge the University of Twente-ITC for giving me the opportunity of academic admission through which my dream came to being. I am grateful to the Netherlands Organization for international cooperation in higher education (nuffic) for allocating the funds for my study. I would like to sincerely thank Jimma University-College of Social Science and Humanities for allowing me to use this scholarship opportunity. I would like to give a special thank my first supervisor Dr Frank B. Osei for all his invaluable advice, comments, critics and guidance in all course of this thesis work. I extend my heartfelt gratitude to my second supervisor Dr Peng Jia for his kind comments and encouragement. I am grateful to Professor Alfred Stein and Mr Wan Bakx for all their constructive ideas and assistance in all my educational life and this thesis work. I would also like to extend my sincere gratitude to the library officers for their kind cooperation in facilitating and availing the required reading materials. I would also like to give my sincere gratitude to Drs. Petra Budde (P.E.) for her absolute kindness and cooperation in providing me the possible data sources related to my thesis topic. I would like to thank my family; my Moms-Ametu Kajela & Bidity Keno, my wife –Bontu Hailu and my 2 little daughters (Nanati and Amerti) for all their encouragements, prayer and support. I would like to give my special thanks to my father; Mr. Tafese Tucho Kumsa for all his provision and inspiration through which he put the building block of my today's life; I wish he could have seen this too! I would like to thank my friends, classmates and also my Ethiopian fellow friends who had in one way or the other touched my life. I will always have a great memory of my spectacular student life in Enschede, and about friends I made in the walk of my life in ITC.

TABLE OF CONTENTS

1. Chapter 1	7
Introduction	7
1.1. Motivation and Problem Statement	7
1.2. Research Identification	8
1.3. General objective	8
2. Chapter 2	10
Literature Review	10
2.1. The Disease Mapping Basics	10
2.2. Related Works	17
3. Chapter 3	21
Description of study area and data	21
3.1. The study area	21
3.2. Data preparation	22
4. Chapter 4	24
Methods and Data	24
4.1. Correlation test	25
4.2. Fitting the generalized linear model (GLM)	26
4.3. Semi-variogram modelling	27
4.4. Regression Kriging Interpolation and Mapping	27
4.5. Cross-Validation	28
5. Chapter 5	29
Result and Discussion	29
5.1. Descriptive statistics of the study data	29
5.2. Standardized mortality ratio (SMR)	30
5.3. Spatial regression modelling; the generalized linear model (GLM)	31
5.4. Prediction of Cholera risk in Sub-Saharan Africa, 2006	32
5.5. Prediction of Cholera incidence in Sub-Saharan Africa in 2007	33
5.6. Prediction of Cholera incidence in Sub-Saharan Africa in 2008	34
5.7. Prediction of Cholera incidence in Sub-Saharan Africa in 2009	35
5.8. Prediction of Cholera incidence in Sub-Saharan Africa in 2010	36
5.9. Prediction of Cholera incidence in Sub-Saharan Africa in 2011	37
5.10. Prediction of Cholera incidence in Sub-Saharan Africa in 2012	38
5.11. Prediction of Cholera incidence in Sub-Saharan Africa in 2013	39
5.12. Prediction of Cholera incidence in Sub-Saharan Africa in 2014	40
5.13. Prediction of Cholera incidence in Sub-Saharan Africa in 2015	41
5.14. Cross-validation of the regression kriging prediction	42
6. Chapter 6	44
Conclusion and Recommendation	44
6.1. Conclusion	44
6.2. Recommendation	46
Reference	47

LIST OF FIGURES

Figure 1. Study area map (overlaid with precipitation raster of 2006)	21
Figure 2. Processed dataset for the study.....	23
Figure 3. Flowchart of the applied methodology	24
Figure 4. Relative risk map by SMR.....	31
Figure 5. Exponential model variogram with cutoff=4000 width=500 (left), Predicted risk (right)	33
Figure 6. Exponential model variogram (left), Predicted risk (right).....	34
Figure 7. Exponential model variogram (left), Predicted risk (right).....	35
Figure 8. Linear model with cutoff=6000 width=400 (left), Predicted risk (right).....	36
Figure 9. Exponential variogram cutoff=6000 width=500 (left), Predicted risk (right)	37
Figure 10. Linear model variogram (left), Predicted risk (right).....	38
Figure 11. Linear model variogram (left), Predicted risk (right).....	39
Figure 12. Exponential variogram (left), Predicted risk (right).....	40
Figure 13. Exponential variogram (left), Predicted risk (right).....	41
Figure 14. Exponential variogram (left), Predicted risk map (right).....	42

LIST OF TABLES

Table 1. Pearson's product moment correlation matrix	26
Table 2. Descriptive statistics of cholera incidence in Sub-Saharan Africa, 2006-2015	30
Table 3. Summary of the model (GLM).....	32
Table 4. Variogram model parameters, 2006.....	33
Table 5. Variogram model parameters, 2007.....	34
Table 6. Variogram model parameters, 2008.....	35
Table 7. Variogram model parameters 2009.....	36
Table 8. Variogram model parameters, 2010.....	37
Table 9. Variogram model parameters, 2011.....	38
Table 10. Variogram model parameters, 2012	39
Table 11. Variogram model parameters, 2013	40
Table 12. Variogram model parameters, 2014	41
Table 13. Variogram model parameters, 2015	42
Table 14. Cross-validation of the regression kriging prediction.....	43

1. CHAPTER 1

Introduction

1.1. Motivation and Problem Statement

In Sub-Saharan Africa, cholera is the persistent life-threatening epidemic disease, influenced by severe climate variability and related environmental and socio-economic factors. According to Lessler et al. (2018) cholera remains a considerable threat to human health in sub-Saharan Africa, severely affecting the poor and most vulnerable social groups. Lobitz et al. (2000) stated that drinking contaminated water and bathing in unpurified water body increases the probability of cholera infection. Cholera can affect all social groups indiscriminately and can cause sudden death through dehydration unless treated as early as possible (Ali et al., 2015).

Drought and harsh climatic condition combined with inadequate public health infrastructure and poor hygiene can influence the incidence of cholera in developing countries such as those within the sub-Saharan African region. The study by Magny, Guégan, Petit, & Cazelles (2007) in five coastal adjoining West African countries suggested that regional climate variability and environmental degradation influenced both the temporal dynamics and the spatial synchrony of cholera epidemic. According to Osei, Duker, & Stein (2012), the spread of cholera incidence is enhanced by socio-economic and environmental factors once there is an outbreak. Reyburn et al. (2011) state that the outbreaks exhibit strong seasonality, tending to occur after increased rainfall and warm temperatures. Likewise, Constantin et al. (2009) emphasize that climate variability and environmental changes influence the emergence of cholera. According to Koelle (2009), climate variability can affect the extent of the epidemic disease outbreak and the level of vulnerability.

In line with the advancement of earth observation science, the short revisit time of high spatiotemporal resolution satellites, nowadays, made it possible to acquire a large amount of high-quality image time series. This, in turn, enabled the extraction of spatial, temporal and spatiotemporal information over a long period in a particular area under study. The progressive improvements in modelling capabilities of geographic information system (GIS) and spatial

statistics combined with remote sensing (RS) data have made possible the monitoring and mapping of the spatial and temporal patterns of infectious diseases. Hence, this study integrates geostatistical modelling techniques, GIS and RS to investigate the spatial variability of cholera incidence across the region using the annual case count and related environmental covariates.

1.2. Research Identification

Cholera remains a persistent threat to human health in Sub-Saharan Africa despite the remarkable research progress. Different factors influence cholera incidence; mainly having both social-economical and natural aspects. The major socio-economic factors are lack of adequate public awareness about the epidemiological nature of the disease and its mode of transmission, poor personal and communal hygiene, unsafe drinking water, inadequate health care infrastructure. The natural aspects would include both climatic and environmental factors that influence the emergence and spread of the cholera epidemic. Hence, results from this research investigation of the possible risk factors that influence the spatial trends of cholera incidence would be of great importance, especially to the regional health authorities to understand the spatial variability of the cholera risk across the region as to plan mitigation and prevention strategies.

1.3. General objective

The primary objective of this research is to investigate the impacts of environmental factors and climate variability on the spatial and temporal trends of cholera incidence in sub-Saharan Africa from 2006 to 2015. The primary objective will be achieved through the following specific objectives.

1.3.1. Specific objectives

1. To investigate the influences of temperature and precipitation on the spatial variability of cholera incidence across the region
2. To use the generalized linear model (GLM) and regression kriging (RK) methods to map the risk of cholera incidence
3. To assess the spatial variability of cholera incidence across the region

1.3.2. Research question

1. What is the influence of temperature and precipitation on cholera incidence across the region?
2. How can we map the risk of cholera incidence across the region?
3. What are the spatial trends in cholera incidence across the region from 2006 - 2015?

1.3.3. Thesis Outline

This thesis is sub-divided into six chapters. Chapter 1 is introduction, motivation and problem statement, research identification, research objectives, and questions. Chapter 2 outlines the review of the literature. Chapter 3 is about the study area and the datasets used for the study. Chapter 4 is about the methodology of the study. Chapter 5 outlines the results and discussions. Chapter 6 is about the conclusion and recommendations.

2. CHAPTER 2

Literature Review

2.1. The Disease Mapping Basics

2.1.1. Spatial epidemiology

Spatial epidemiology is a branch of medical science concerned with studying, measuring, analyzing and interpreting spatial variations in disease distribution (Elliot, Wakefield, Best, & Briggs, 2000). Since the first London cholera outbreak mapping by Dr John Snow in September 1854, several studies have been carried out on cholera epidemiology and its burden on a human population across the globe. Dr John, in his research, was able to theorize that cholera reproduced in a human body and was spread through contaminated water which was contradictory with the prevailing theory that the disease was spread by "miasma" or fog in the air. London sewage system was even more ad hoc in which the pervasive stench of animal and human feces combined with rotting garbage which made the miasma theory more plausible. Snow then mapped the public wells and all known risks of cholera incidence around them and realized that there was a spatial clustering of the outbreaks around the water pump location points. Eventually, as he had the pump handles removed, and then the disease outbreak reduced (Johnson & Collection., 2007).

Most epidemiological studies of cholera have then focused on the pathogenies and biological characteristics of *V. cholerae*. Nevertheless, such kind of research does not specify the disease risk and its related risk factors (Frank Badu Osei, 2010). Spatial epidemiological mapping methods can depict areas with high-risk estimates and helps to formulate a hypothesis about the potential factors influencing such variations based on spatial information about the outbreak and for optimal site selection for allocation of health facilities (Wang, Zhao, & Wang, 2018; Frank Badu Osei, 2010).

2.1.2. Disease mapping

In disease mapping, the target region is partitioned into n neighbouring areas such that the areas that exhibit high disease risks will be detected. The observed disease case in each respective

region is denoted by Y_i where $i = 1, \dots, n$. The expected numbers of disease cases are denoted by E_i and are based on the size and demographic structure of the population living within each study region. Therefore, to assess which areas that exhibit relatively higher disease risk, the expected number of cases to occur in each area will be computed by dividing the population living in each region into a number of strata. Then the number of people in each stratum is multiplied by the incidence rate for that particular stratum. And the result is summed up based on the strata to produce the expected number of the cases (Moraga, 2017). The expected disease cases can be computed as;

$$E_i = \sum_{j=1}^m r_j * n_j \quad 2.1$$

r_j is an incidence rate (total number of observed disease case divided by the total number of population in the study region), whereas n_j is the population in a strata j . Disease risk is estimated by Standardized Mortality Ratio (SMR) as;

$$SMR = \frac{Y_i}{E_i} \quad 2.2$$

Y_i and E_i are observed and expected disease cases in area $i = 1, \dots, n$ respectively, $SMR > 1$ shows more cases are observed than expected and $1 < SMR$ shows fewer cases are observed than expected. To capture a large spread with the increasing trend across the study area, the variance of the estimated SMR would be estimated as;

$$var(SMR_i) = \frac{SMR_i}{E_i} \quad 2.3$$

In this case, as E_i becomes smaller, $var(SMR)$ will become larger irrespective of the study area size as E_i is a function population size. This high variability sometimes suggests that the extreme SMRs will be based on a lower expected number of the disease case within the study area. In this regard, SMR often is misleading for areas with a small number of population as it shows larger disease risk than expected. Hence, it is quite useful to give an understanding of the spatial variability of the disease risk across the study area whereby the related potential covariates and correlation information from the neighbourhood regions can be incorporated into the model to produce the smooth prediction the target response variable (Waller & Gotway, 2004).

2.1.3. Cholera

Cholera is an infectious disease caused by the bacteria *Vibrio cholerae*. This deadly disease is often manifested by the sudden onset of rice water looking diarrhoea, vomiting, rapid dehydration, shock, and death unless treated as early as possible. Cholera is one of the most widespread infectious diseases that appear to be influenced by climate, geography and other environmental factors (Xu et al., 2016). Constantin et al. (2009) stated that climate variability and environmental changes influence the emergence and incidence of cholera in a human population.

Drought and harsh climatic condition combined with inadequate public health infrastructure and poor hygiene can influence the incidence of cholera in developing countries such as sub-Saharan African countries. According to Osei, Duker, & Stein (2012), the spread of cholera incidence is enhanced by socio-economic and environmental factors once there is an outbreak. Reyburn et al. (2011) state that the outbreaks exhibit strong seasonality, tending to occur after increased rainfall and warm temperatures. Likewise, Constantin et al. (2009) emphasize that climate variability and environmental factors influence the emergence of cholera in a human population. According to Koelle (2009), climate change influences the extent of the epidemic disease and the level of exposure of the community in the area of the outbreak.

Cholera can transmit in two possible modes of transmissions. One occurs through exposure to an environmental reservoir of *V. cholerae*; the other is through the faecal-oral route through utilization of faecal-contaminated water resources (Frank Badu Osei, 2010). Likewise, Lobitz et al. (2000) emphasize that drinking contaminated water and bathing in an unpurified brackish river can increase the probability of cholera infection based on the water temperature and aquatic ecosystem in the marine environment.

2.1.4. Climate variability

Climate variability is one among the significant environmental challenges confronting most developing countries like Africa nowadays. Climate variability refers to a long-term change in the weather pattern of the extreme weather condition (Koelle, 2009b). Climate variability includes changes in one or more climate variables such as temperature, precipitation, wind and sunshine which in turn impact the survival and reproduction of living things including disease pathogens (Wu, Lu, Zhou, Chen, & Xu, 2016). This change, to the higher dimension, can

disrupt the healthy life of the human population, by multiplying existing health problems both at the local and global level.

2.1.5. Spatial regression model

A spatial regression model is a statistical model that characterizes the spatial relationship between a variable of interest, and one or more explanatory variables. For spatially referenced data, the spatial random effect (model residual) associated with each area of the outbreak helps to model the underlying spatial dependence structure (Lawson, 2010). Spatial dependence is measured by spatial autocorrelation, which is a property of data that arises whenever there is a spatial pattern in the values, as opposed to a random pattern where there is no spatial autocorrelation (Wakefield, 2007). The underlying process may vary systematically over space due to correlations with other explanatory variables. This can be modelled by ordinary least squares (OLS) estimation as a simple regression model;

$$y_i = x\beta_i + \varepsilon_i \quad 2.4$$

Including the y-intercept, equation 2.9 can be written as $y_i = \beta_0 + \beta_1 x_i$; y_i is an observed response variable, X_i is an explanatory variable (covariate), β_0 is a y-intercept, β_1 is regression coefficient and ε is uncorrelated random effect (residual) normally distributed with Gaussian distribution as $E[\varepsilon_i] \sim N(0, \sigma^2)$. Therefore, mean of the response variable Y_i is estimated as;

$$E[Y_i] = \beta_0 + \beta_1 X_{ij} \quad 2.5$$

$X_{i,j}$ is a j^{th} predictor variable measured for the i^{th} observation. The main assumptions for the errors ε_i is that $E[\varepsilon_i] = 0$, for $i = 1, \dots, n$.

The ordinary least squares (OLS) is a simple linear model with one predictor variable that explains the relationship between the dependent variable and independent variable with unspecified trend in the dataset. Multiple regression model contains two or more explanatory variables, and is typically helpful to perform a precise prediction of the response variable with more than one explanatory variables as (Kutner, 2005) and (Waller & Gotway, 2004);

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \dots + \beta_p X_{i,p} + \varepsilon_i \quad 2.6$$

It can be written as;

$$Y_i = \beta_0 + \sum_{k=0}^p \beta_k x_{ik} + \varepsilon_i \quad 2.7$$

Assuming that $E[\varepsilon_i] = 0$, the mean of response variable $E[Y_i]$ is:

$$E[Y_i] = \beta_0 + \beta_1 X_{i,1} + \dots + \beta_p X_{i,p} \quad 2.8$$

2.1.6. Poisson Regression model

Poisson regression model is one of a generalized linear model (GLM) used to model the count of data in which the mean of the response variable Y_i is assumed to have a Poisson distribution over a fixed time and space. The probability mass function of Poisson sampling distribution is given as;

$$P(Y = y) = \frac{e^{-\lambda} \lambda^y}{y!} \quad 2.9$$

Where y is a realization of the Poisson distributed discrete random variable Y , λ is the average rate of occurrence of the Poisson distributed event y . The Poisson regression model is;

$$Y_i | \theta_i \sim \text{Poisson}(\mu_i) \\ Y_i = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p + \varepsilon_i \quad 2.10$$

Where $i=1, \dots, N$, the expected count of Y_i is $E[Y_i] = \mu_i$ and $E[\varepsilon_i] = 0$. Hence, the above equation is simplified as;

$$E[Y_i] = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p \quad 2.11$$

The equivalent log-linear model is;

$$\log(E[Y_i]) = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p$$

This equation can farther be simplified as;

$$\mu_i = \beta_0 + \exp(\beta_1 X_1 + \dots + \beta_p X_p) \quad 2.12$$

Y_i is observed count of cholera outbreak to be estimated (mean of the response variable), θ_i is a relative risk, $E[]$ is expectation, β_0 is y-intercept, $(\beta_1 + \dots + \beta_p)$ are regression coefficients which quantify association of cholera outbreak with explanatory variable $(X_1 + \dots + X_p)$ measured in the study region, the residual follows Gaussian distribution with the mean of $E[\varepsilon_i] \sim \text{Gauss}\{0, \sigma^2\}$.

2.1.7. Variogram model

Variogram $2\gamma(h)$, is a mathematical function that describes spatial autocorrelation and spatial structure that it provides a measure of how much two samples of the same variable $Z(s)$ taken from different observation locations will vary depending on the distance (h) between them (Robinson & Dietrich, 2016). The equation is:

$$2\gamma(h) = Var(Z(s+h) - Z(s)) \quad 2.13$$

Semivariance $\gamma(h)$ is the geostatistical measure of autocorrelation among the target pairs of points based on the sum of the average squared difference that are separated by a spatial lag distance h from one another in a single direction which is half of the variogram model (Robinson & Dietrich, 2016) and is given as:

$$\gamma(h) = \left(\frac{1}{2}\right) Var(Z(s+h) - Z(s)) \quad 2.14$$

In intrinsic stationarity, the mean and variance are constant. However, as the lag distance increases the variance tends to change as a function of the lag (h). The mean is constant at all observation locations implies that expected value at each observation location is the same. That is; $E[Z(s+h)] = E[Z(s)] = \mu$. Thus, the equation is derived as:

$$E[Z(s+h) - Z(s)] = E[Z(s+h)] - E[Z(s)] = 0$$

Variance is estimated as;

$$Var[Z(s+h) - Z(s)] = E[Z(s+h) - Z(s)]^2 \quad 2.15$$

Therefore, semivariance is estimated as an average of the squared differences among the pair of observation points (Robinson & Dietrich, 2016) and the equation becomes;

$$\gamma(h) = \left(\frac{1}{2}\right) E[Z(s+h) - Z(s)]^2 \quad 2.16$$

$\gamma(h)$ is semivariogram, $Z(s)$ is the observation at location s , whereas h is spatial lag (distance) between pairs of points in a given bin.

2.1.8. Regression Kriging Interpolation

Kriging interpolation is the process of predicting values for the variable of interest at unsampled locations from the surrounding sampled locations within the study area. Interpolations fall into two broad categories; deterministic (inverse-distance weighting) and probabilistic (kriging) interpolation (Lance A. Waller, 2004). Deterministic interpolation is a mathematical model which does not have a measure of uncertainty associated with it, as it is a function of the weighted inverse distance between the observation and prediction locations; whereas kriging is a probabilistic interpolation method having its foundation in statistical theory to assume a statistical model for the data to perform an optimal spatial prediction (Lawson, Banerjee, Haining, & Ugarte, 2016).

Literature shows that there is a significant interest in the fields of spatial epidemiology to interpolate disease incidence to map the risk of disease occurrence from a regional database onto a continuous surface within the study region. Hence, regression kriging is one among the numerous geostatistical interpolation techniques that perform a smoothed prediction of a spatial mean of the target response variable that exhibits a spatial variability across space (Berke, 2004) concerned in this study. In regression kriging, estimation of the regression coefficient is made separately using generalized least squares (GLS) whereas the GLM model residuals are interpolated using ordinary kriging prediction and added back to the mapping equation (Noel A. C. Cressie, 1993).

$$\hat{z}_{Rk}(s_0) = \hat{m}(s_0) + \hat{e}(s_0) \quad 2.17$$

Where $\hat{m}(s_0)$ non-spatial trend component predicted from the GLM model (trend surface) and $\hat{e}(s_0)$ is the interpolated and back-transformed the model residual known as a spatial random effect.

$$\hat{z}_{Rk}(s_0) = \sum_{k=0}^p \hat{\beta}_k \cdot q_k(s_0) + \sum_{i=1}^n w_i(s_0) \cdot e(s_i) \quad 2.18$$

Where $\hat{\beta}_k$ is regression coefficient estimated from the data, w_i is the weight matrix determined by semivariance function in the semi-variogram model, e is the model residual.

2.2. Related Works

Since the first London cholera outbreak mapping by Dr John Snow in September 1854, several studies have been conducted on cholera epidemiology and its burden on the human population across the globe. Dr John was able to theorize that cholera reproduced in the human body and was spread through contaminated water which was contradictory with the prevailing "miasma" theory that the disease was spread fog in the air. Snow used a point pattern analysis techniques to investigate the disease clusters around the broad street pumps and could be able to map the public wells and all corresponding cholera-prevalent locations around them (Johnson & Collection., 2007). The work had put a remarkable basement for the emergence of the present day spatial epidemiology as it could be able to disprove the prevailing miasma theory which had been believed that cholera spread was influenced by the poisonous fog in the air which did not have scientific evidence. However, there was no means in the study to include the spatial dependence structure and environmental covariates into the model that could influence the spread of cholera incidence. The point pattern approach rather helped him to investigate the variation in the intensity of the point patterns as a function of the weighted Euclidean distance to the source of the disease spread (the broad street pumps) order to locate the closest and the more cholera-prevalent areas across the city of London.

Bandyopadhyay, Kanji, & Wang (2012) quantified the impacts of temperature and rainfall on the regional incidence of diarrheal infection in Sub-Saharan Africa. In their study, they utilized the demographic and health survey dataset from 14 Sub-Saharan countries combining with the temperature and rainfall raster between 1992 and 2001. In their study, they used the ordinary least square estimation (OLS) to test the impacts of climate variability on diarrheal infection in the 14 countries of Sub-Saharan Africa. Their result showed that decreased average rainfall and increased mean temperature were together found to increase the incidence of diarrhoea infection. However, disease data by its nature is a discrete count data in which observation can only take non-negative integer which is preferably modelled by Poisson log-linear regression model that accounts for non-normal data distribution and over-dispersed data in the disease cases which is where the OLS fails to model.

A joint study was conducted on the burden of cholera on children in Africa; Beira (Mozambique) and Asia; Jakarta (Indonesia) and Kolkata (India) by Deen et al. (2008). They mainly focused on the laboratory experiment of isolation of cholera bacterium, the *V.cholerae* from the rectal swabs collected from all age groups. In the study, they computed the incidence rate (per 1000 population) of the cholera incidence across the study sites in which they indicated that there were higher burden of cholera on children over 5 years of ages in each respective study area among which Africa (Mozambique) had been shown to have taken the highest cholera incidence level.

Luquero, Francisco J. (2009) estimated the risk of cholera incidence in Guinea-Bissau in time and space. They used the historical cholera data to compute the incidence and case fatality rate of the disease. They fitted a Poisson regression model to the log-transformed cholera data to assess the spatiotemporal trends of cholera incidence. They applied the non-parametric Cubic splines for smoothing the prediction. The study revealed that there were spatial trends in the disease incidence, whereas there was no that much significant secular trend of cholera incidence in the region. They also performed a sensitivity analysis by varying different degrees of freedom while smoothing the prediction based on the results of adjusted R^2 from the regression model.

Osei, Duker, & Stein (2011) used the hierarchical Bayesian modelling of the space-time diffusion patterns of cholera in Kumasi, Ghana in which they analyzed the joint effects of the two modes of cholera transmissions (environment-to-human and human-to-human) on space-time diffusion dynamics. According to the authors, the primary and the most responsible sparking route for space-time diffusion patterns of the cholera epidemic is the human interaction with an aquatic reservoir of the cholera bacterium (*V. cholera*) whereas, the second is through fecal-oral based transmission from the pre-infected person. The authors discussed the gaps in the traditional diffusion modelling technique in which the classical linear regression is used with the assumption of the response variable to be normally distributed and linearly interacts with the related explanatory variables that it ignores the possible nonlinearity and spatial effects of the explanatory variables. Hence, to investigate these transmissions, they developed the integrated statistical models; one of which is hierarchical Bayesian modelling for joint analysis of nonlinear effects on continuous covariates (proximity to primary case location and population density) and spatially structure (a reference to the community) and unstructured random effects. The authors also performed the joint variogram modelling to define the extent of the spatial autocorrelations among the observation points within the contiguous case locations in which the heterogeneous

cholera case counts were assumed to be the realizations of a discrete random variable which follows a Poisson distribution. The authors also estimated the unknown model parameters by a fully Bayesian approach in which the prior assumptions were first specified to smooth the posterior estimation of the unknown model parameters using MCMC simulation techniques. The study confirmed that the integrated hierarchical Bayesian modelling approach supported with different variogram modelling of spatial autocorrelation between each Poisson distributed cholera cases observation point well explored the space-time diffusion patterns of cholera incidence in Kumasi, Ghana. However, the method is computationally intensive, and also the interpretation of the final output from such interconnected complex methodologies require a careful understanding of the intermediate result at each hierarchical level as the resulting posterior estimate is from the prior ones.

A study was conducted on cholera incidence in the same study region by Lessler et al. (2018). In this study, the authors used a Bayesian modelling framework to maps the risk of cholera incidence from 2010 - 2016. The authors integrated the reported cholera cases from several spatiotemporal scales mainly focusing on socio-economic factors. They formulated an approach in which they divided the study areas into 20km x 20km grid cells. They used conditional autoregressive (CAR) to model the spatially correlated random effects. Each observation was mapped to the corresponding grid cell would be the sum of the expected number of cases from all grid cells in the area. That way, they could find substantial heterogeneities of the incidence based on geographical locations such as within and between countries of the study region. However, the Bayesian approach by its nature is highly influenced by prior knowledge about the distribution of the disease cases which sometimes is up to the personal knowledge about the approximate prior information that will be incorporated into the model; and that will make it difficult to validate (Ainsworth & Dean, 2006). Besides, the study seems to neglect the current punitive climate variability and environmental factors which are thought to influence the spatial variability of cholera incidence while giving more emphasis to the socio-economic aspects only, which are mainly about the presence or absence of infrastructures within the community.

This study differs from the preceding ones in its focuses on the investigation of impacts of the punitive climatic variability and environmental factors that are thought to have potentially enhanced the risk of cholera incidence across Sub-Saharan Africa. The Poisson generalized log-linear model (GLM) is used to assess the spatiotemporal variability of cholera incidence and produce a coarse resolution map of relative risk across the region. The data required for this

study are; cholera cases dataset from 2006 to 2015 and the corresponding explanatory variables; the mean annual temperature and mean annual precipitation are taken from NASA earth resource database which is formerly derived from MODIS (Aqua) image time series. Regression kriging (alternatively; Universal Kriging) helps to map the relative risk of cholera incidence for 41 countries of Sub-Saharan Africa countries for ten years (2006 – 2015). The accuracy assessment is carried out to test the performance of the regression models for all of the ten models developed on yearly basis (2006-2015) using a leave-one-out cross-validation technique. The findings from this research would give a spatial highlight to the regional public health authorities to design their intervention strategies to mitigate the disease risk.

3. CHAPTER 3

Description of study area and data

3.1. The study area

Sub-Saharan Africa is one among the many regions of the world in which cholera has been persistently occurring in both endemic and epidemic situations. After about 40 years of its reappearance in 1970, cholera remained a public health issue being an immense disease burden frequently affecting the poor and most vulnerable social groups. From 1970 until 2011, 3,221,050 cholera suspects were reported to WHO which accounts for 46% of the total global report (Mengel, Delrieu, Heyerdahl, & Gessner, 2014). Among the 63658 cholera deaths reported by WHO between 2000 and 2015, about 83% (52 812) occurred in Sub-Saharan African (Lessler et al., 2018a).

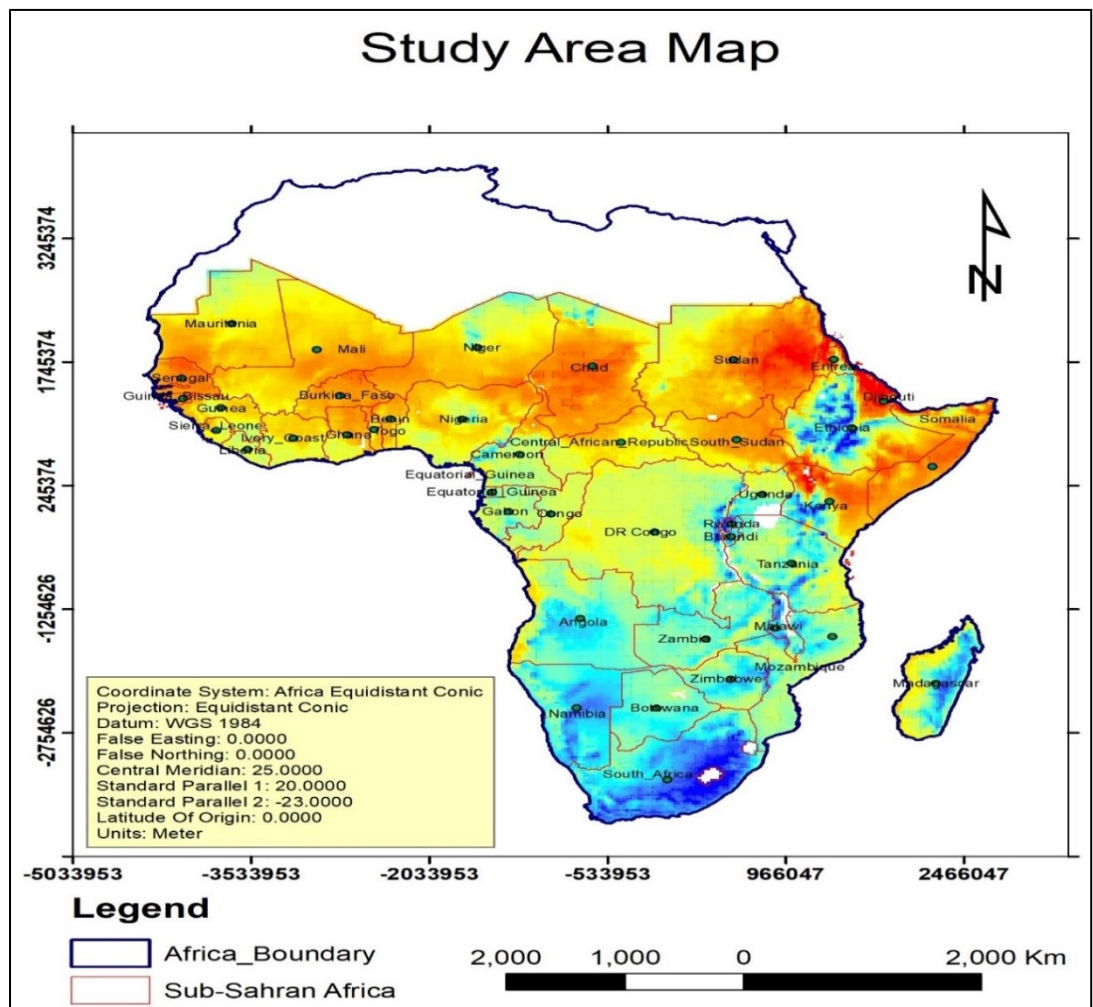


Figure 1. Study area map (overlaid with precipitation raster of 2006)

3.2. Data preparation

3.2.1. Cholera and population data

Cholera incident dataset was collected from WHO weekly epidemiological record (WER) from all respective countries of the study region from January 01, 2006 to December 31, 2015. The dataset by its nature is a zero-inflated dataset as the case was not uniformly reported to WHO from all countries within the study periods. This may be due to either there was no cholera incidence recorded in that country within that particular year or maybe lack of regularity in reporting on a yearly basis. The Poisson generalized log-linear model (GLM) is preferred as it allows to model the relationship between the target count variable and one or more explanatory variables.

3.2.2. Temperature and precipitation raster

For this analysis, the gridded (raster) temperature (°F) and precipitation (mm) time series having a spatial resolution of 0.250 x 0.250 degrees were taken from NASA earth resource database; ([https://giovanni.gsfc.nasa.gov/giovanni/...*](https://giovanni.gsfc.nasa.gov/giovanni/...)). The raster datasets were processed in ArcGIS to extract the mean values that were later incorporated into the spatial regression model as explanatory variables to investigate their influence on cholera incidence across the region.

3.2.3. Image processing

The images of the selected environmental covariates; temperature and precipitation time series were imported into the GIS environment (ArcGIS) and then clipped to the boundary of the study area, sub-Saharan Africa. As some of them were in different spatial resolutions, resampling had to be carried out and done accordingly. Then the minimum, maximum and the mean annual temperature and precipitation values for were computed using spatial analyst toolset in ArcToolbox (raster calculator) and exported as CSV file format for further analysis in R programming language (the open-source statistical software) (Figure 2).

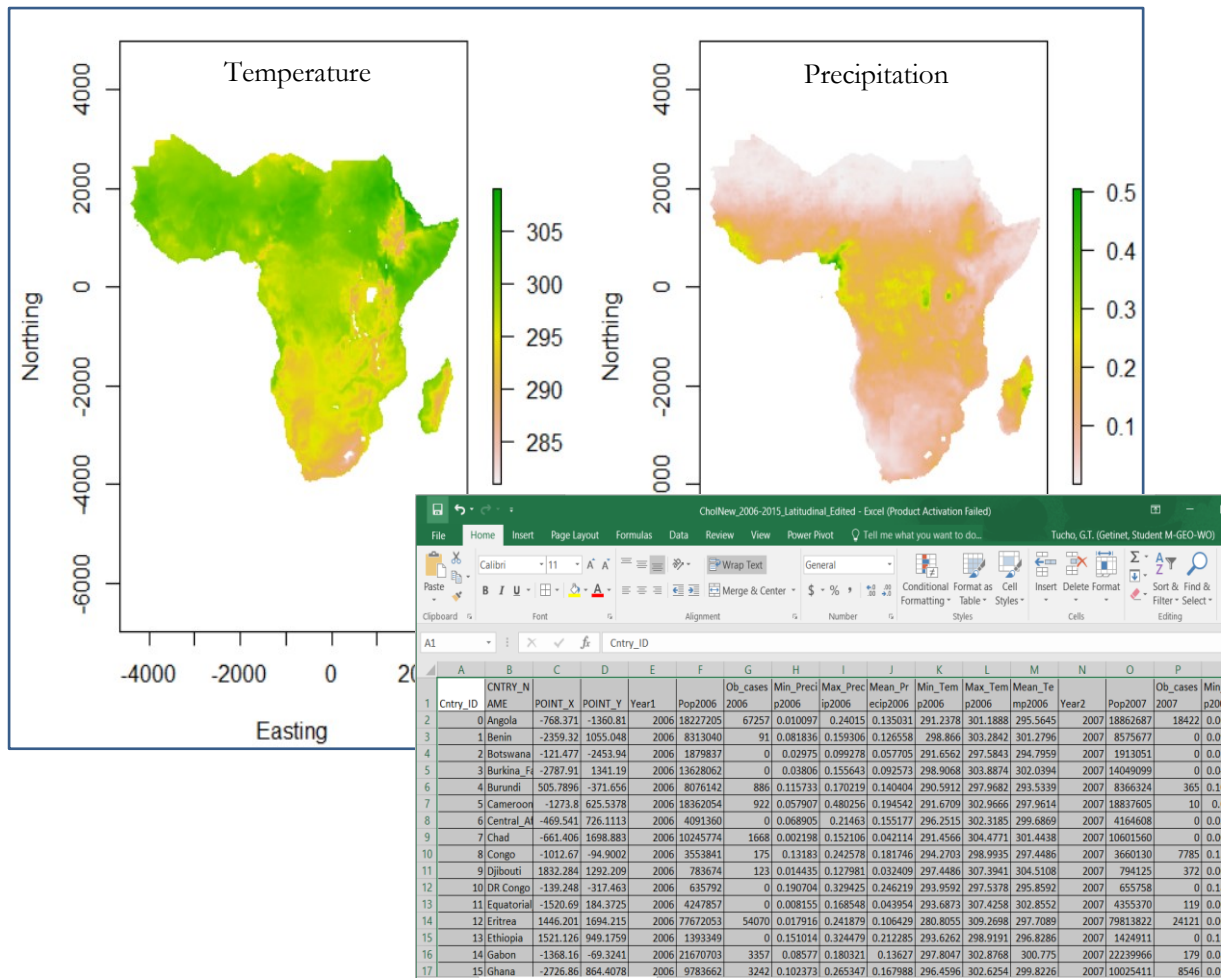


Figure 2. Processed dataset for the study

4. CHAPTER 4

Methods and Data

The study adopted different methodologies to investigate the impacts of climate variability and environmental variables on the spatial trends of cholera incidence across the region. Below is the flowchart of the overall methodological framework of the study (Figure 3).

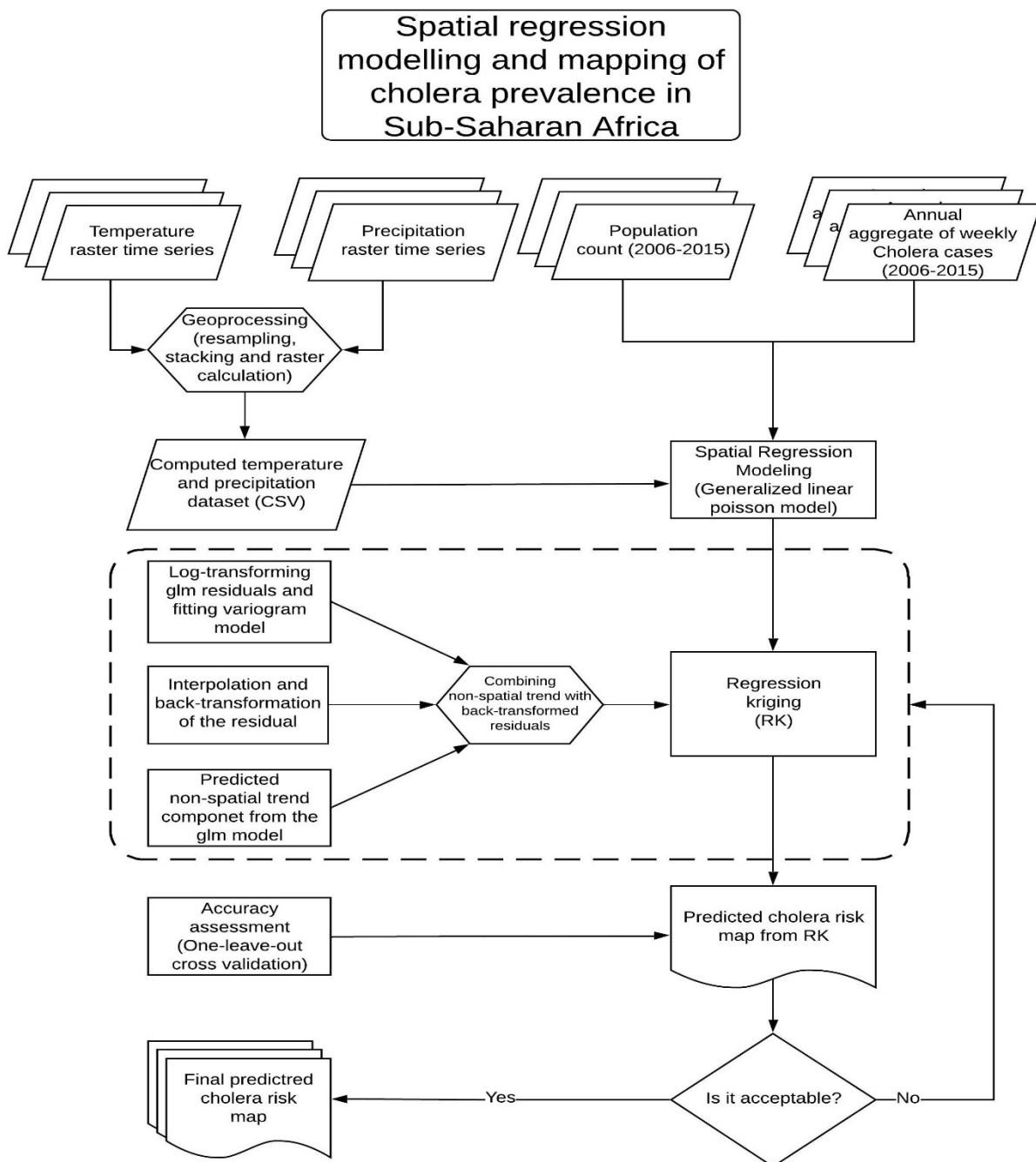


Figure 3. Flowchart of the applied methodology

4.1. Correlation test

Pearson's product moment correlation test is a pairwise comparison test that helps to investigate the linear relationship of two continuous variables. Pearson's correlation test was performed to identify the covariates with a strong correlation to the target response variable, the risk of cholera incidence in this case. All covariates were seen to be weakly correlated to the target variable; however, they were still statistically significant to influence the target variable with p-values less than $p < 2e-16$ (Table 3). Hence, the mean annual temperature and mean annual precipitation were used as explanatory variables into the regression models (Table 1).

	Ob_cases2006	Min_Precip2006	Max_Precip2006	Mean_Precip2006	Min_Temp2006	Max_Temp2006	Mean_Temp2006
Ob_cases2006	1	-0.215136776	0.101296692	0.007765867	-0.296888312	0.261125471	-0.06812684
Min_Precip2006	-0.215136776	1	0.399600386	0.863856485	0.228055678	-0.504945029	-0.202511992
Max_Precip2006	0.101296692	0.399600386	1	0.7444907	-0.25482985	-0.119292547	-0.186682048
Mean_Precip2006	0.007765867	0.863856485	0.7444907	1	0.023644954	-0.426377035	-0.281124533
Min_Temp2006	-0.296888312	0.228055678	-0.25482985	0.023644954	1	0.189445992	0.650153734
Max_Temp2006	0.261125471	-0.504945029	-0.119292547	-0.426377035	0.189445992	1	0.816206635
Mean_Temp2006	-0.06812684	-0.202511992	-0.186682048	-0.281124533	0.650153734	0.816206635	1
	Ob_cases2007	Min_Precip2007	Max_Precip2007	Mean_Precip2007	Min_Temp2007	Max_Temp2007	Mean_Temp2007
Ob_cases2007	1	-0.21082858	0.18804774	-0.035707274	-0.262185404	0.228722148	0.022167668
Min_Precip2007	-0.21082858	1	0.404528677	0.857406614	0.231623295	-0.525403397	-0.19932976
Max_Precip2007	0.18804774	0.404528677	1	0.75861055	-0.207587884	-0.120412288	-0.162724527
Mean_Precip2007	-0.035707274	0.857406614	0.75861055	1	0.044010078	-0.411560629	-0.237119263
Min_Temp2007	-0.262185404	0.231623295	-0.207587884	0.044010078	1	0.208181058	0.669430423
Max_Temp2007	0.228722148	-0.525403397	-0.120412288	-0.411560629	0.208181058	1	0.80815175
Mean_Temp2007	0.022167668	-0.19932976	-0.162724527	-0.237119263	0.669430423	0.80815175	1
	Ob_cases2008	Min_Precip2008	Max_Precip2008	Mean_Precip2008	Min_Temp2008	Max_Temp2008	Mean_Temp2008
Ob_cases2008	1	-0.104790375	-0.197913002	-0.13851894	-0.200636778	-0.083766495	-0.139317884
Min_Precip2008	-0.104790375	1	0.406875269	0.78099057	0.266130898	-0.48886804	-0.179100394
Max_Precip2008	-0.197913002	0.406875269	1	0.666943162	-0.14414515	-0.077096267	-0.155772519
Mean_Precip2008	-0.13851894	0.78099057	0.666943162	1	0.034912365	-0.232576595	-0.205823413
Min_Temp2008	-0.200636778	0.266130898	-0.14414515	0.034912365	1	0.150061053	0.646708738
Max_Temp2008	-0.083766495	-0.48886804	-0.077096267	-0.232576595	0.150061053	1	0.797688559
Mean_Temp2008	-0.139317884	-0.179100394	-0.155772519	-0.205823413	0.646708738	0.797688559	1
	Ob_cases2009	Min_Precip2009	Max_Precip2009	Mean_Precip2009	Min_Temp2009	Max_Temp2009	Mean_Temp2009
Ob_cases2009	1	-0.21194045	0.07398563	-0.151382602	-0.437424429	0.017616482	-0.21900426
Min_Precip2009	-0.21194045	1	0.372608014	0.86606575	0.250110492	-0.54174513	-0.217476502
Max_Precip2009	0.07398563	0.372608014	1	0.705253685	-0.273294701	-0.176469912	-0.263957438
Mean_Precip2009	-0.151382602	0.86606575	0.705253685	1	0.041037797	-0.480368685	-0.302540241
Min_Temp2009	-0.437424429	0.250110492	-0.273294701	0.041037797	1	0.224940452	0.670537261
Max_Temp2009	0.017616482	-0.54174513	-0.176469912	-0.480368685	0.224940452	1	0.82558819
Mean_Temp2009	-0.21900426	-0.217476502	-0.263957438	-0.302540241	0.670537261	0.82558819	1
	Ob_cases2010	Min_Precip2010	Max_Precip2010	Mean_Precip2010	Min_Temp2010	Max_Temp2010	Mean_Temp2010
Ob_cases2010	1	-0.078845437	0.448192785	0.076539575	-0.064444205	0.058918554	0.027333848
Min_Precip2010	-0.078845437	1	0.350268514	0.864411015	0.271788515	-0.488819971	-0.155891188
Max_Precip2010	0.448192785	0.350268514	1	0.666074693	-0.15637082	-0.116060115	-0.214588135
Mean_Precip2010	0.076539575	0.864411015	0.666074693	1	0.181657902	-0.408033925	-0.209942322
Min_Temp2010	-0.064444205	0.271788515	-0.15637082	0.181657902	1	0.242371823	0.675517827
Max_Temp2010	0.058918554	-0.488819971	-0.116060115	-0.408033925	0.242371823	1	0.821323606
Mean_Temp2010	0.027333848	-0.155891188	-0.214588135	-0.209942322	0.675517827	0.821323606	1

	Ob_cases2011	Min_Precip2011	Max_Precip2011	Mean_Precip2011	Min_Temp2011	Max_Temp2011	Mean_Temp2011
Ob_cases2011	1	-0.085067818	-0.056777182	-0.105722589	0.034642559	0.219752235	0.194564213
Min_Precip2011	-0.085067818	1	0.935172502	0.712078024	-0.219453608	-0.108396854	-0.209980734
Max_Precip2011	-0.056777182	0.935172502	1	0.417237048	-0.22959155	0.104640691	-0.089783416
Mean_Precip2011	-0.105722589	0.712078024	0.417237048	1	-0.107971281	-0.485552438	-0.360800996
Min_Temp2011	0.034642559	-0.219453608	-0.22959155	-0.107971281	1	0.304649509	0.713454611
Max_Temp2011	0.219752235	-0.108396854	0.104640691	-0.485552438	0.304649509	1	0.835673592
Mean_Temp2011	0.194564213	-0.209980734	-0.089783416	-0.360800996	0.713454611	0.835673592	1
	Ob_cases2012	Min_Precip2012	Max_Precip2012	Mean_Precip2012	Min_Temp2012	Max_Temp2012	Mean_Temp2012
Ob_cases2012	1	0.184895643	0.112062127	0.235450003	-0.00830525	0.041437641	0.087669502
Min_Precip2012	0.184895643	1	0.286162391	0.873006089	0.221569551	-0.452826977	-0.169396958
Max_Precip2012	0.112062127	0.286162391	1	0.574065122	-0.299833519	0.05333194	-0.134996766
Mean_Precip2012	0.235450003	0.873006089	0.574065122	1	0.069483687	-0.353604877	-0.198479389
Min_Temp2012	-0.00830525	0.221569551	-0.299833519	0.069483687	1	0.28608974	0.704358649
Max_Temp2012	0.041437641	-0.452826977	0.05333194	-0.353604877	0.28608974	1	0.831173503
Mean_Temp2012	0.087669502	-0.169396958	-0.134996766	-0.198479389	0.704358649	0.831173503	1
	Ob_cases2013	Min_Precip2013	Max_Precip2013	Mean_Precip2013	Min_Temp2013	Max_Temp2013	Mean_Temp2013
Ob_cases2013	1	-0.042324521	0.12225619	0.093283416	-0.179359116	-0.019609571	-0.036956993
Min_Precip2013	-0.042324521	1	0.487414541	0.909890608	0.164962426	-0.490710815	-0.235997285
Max_Precip2013	0.12225619	0.487414541	1	0.712243938	-0.191449676	-0.100408437	-0.170831905
Mean_Precip2013	0.093283416	0.909890608	0.712243938	1	0.066223935	-0.377496891	-0.228061584
Min_Temp2013	-0.179359116	0.164962426	-0.191449676	0.066223935	1	0.307842041	0.713464581
Max_Temp2013	-0.019609571	-0.490710815	-0.100408437	-0.377496891	0.307842041	1	0.840482221
Mean_Temp2013	-0.036956993	-0.235997285	-0.170831905	-0.228061584	0.713464581	0.840482221	1
	Ob_cases2014	Min_Precip2014	Max_Precip2014	Mean_Precip2014	Min_Temp2014	Max_Temp2014	Mean_Temp2014
Ob_cases2014	1	-0.012282842	0.186407295	0.089828924	-0.015364984	0.098162301	0.135997977
Min_Precip2014	-0.012282842	1	0.468177172	0.911430555	0.224323679	-0.446998808	-0.186745532
Max_Precip2014	0.186407295	0.468177172	1	0.684849726	-0.21905009	-0.158725142	-0.224571419
Mean_Precip2014	0.089828924	0.911430555	0.684849726	1	0.076936824	-0.391470082	-0.237795159
Min_Temp2014	-0.015364984	0.224323679	-0.21905009	0.076936824	1	0.288993134	0.698816356
Max_Temp2014	0.098162301	-0.446998808	-0.158725142	-0.391470082	0.288993134	1	0.845047646
Mean_Temp2014	0.135997977	-0.186745532	-0.224571419	-0.237795159	0.698816356	0.845047646	1
	Ob_cases2015	Min_Precip2015	Max_Precip2015	Mean_Precip2015	Min_Temp2015	Max_Temp2015	Mean_Temp2015
Ob_cases2015	1	0.30516629	0.384115847	0.096975479	-0.378796436	0.129338272	-0.005110706
Min_Precip2015	0.30516629	1	0.945219077	0.839536368	-0.204345208	-0.097959106	-0.187417862
Max_Precip2015	0.384115847	0.945219077	1	0.616191611	-0.316195946	0.083263864	-0.122727998
Mean_Precip2015	0.096975479	0.839536368	0.616191611	1	0.033234126	-0.37492656	-0.247922704
Min_Temp2015	-0.378796436	-0.204345208	-0.316195946	0.033234126	1	0.245819374	0.682468804
Max_Temp2015	0.129338272	-0.097959106	0.083263864	-0.37492656	0.245819374	1	0.82996361
Mean_Temp2015	-0.005110706	-0.187417862	-0.122727998	-0.247922704	0.682468804	0.82996361	1

Table 1. Pearson's product moment correlation matrix

4.2. Fitting the generalized linear model (GLM)

Modifying the original count data to satisfy the assumption of normality could end up in providing a misleading conclusion (Lo & Andrews, 2015). A generalized linear model (GLM) allows a response variable which has a distribution other than the normal distribution. Hence, the GLM model was used to fit the varying cholera case counts with the assumption of Poisson

distribution in which mean and variance are equal. Then the model residuals were extracted and log-transformed to compute the sample variance. The empirical semi-variogram models were fitted to the computed sample variogram to estimate the variogram parameters. The best-fitted variogram models with optimal parameters were included in the regression kriging prediction to map the risk of cholera incidence across the region.

4.3. Semi-variogram modelling

The covariance structure of spatial variability between pairs of observation points within the study region was estimated through parameterized semi-variogram modelling as a function of spatial separation or the lag distance (h) between each corresponding pair of points. The three main variogram models that were used to investigate the spatial structure in cholera incidence are Exponential, Spherical and Linear variogram models. To fit these models, the experimental sample variogram were computed using the log-transformed GLM residuals. The three important variogram parameters (Sill, Range, and Nugget) were computed from the fitted variogram models. The models with optimal parameters were chosen and incorporated into the ordinary kriging to interpolate the log-transformed residuals. The log-transformed residuals were summed up to the trend component derived from the GLM to map the risk of cholera incidence across the study region.

4.4. Regression Kriging Interpolation and Mapping

The GLM model residuals were extracted and log-transformed, the experimental variogram models were fitted to the log-transformed residuals to assess covariance structures of the observed cases at each location within the study region. Then residuals are interpolated using Ordinary Kriging. The interpolated residuals were then back-transformed and summed up to the non-spatial mean component predicted with the GLM model. Then the isopleth maps of the risk of cholera incidence across the region were produced for ten years (2006-2015). From these maps, locations of the highest cholera risks were identified. That is, the hottest the colour in the map the higher the risk of cholera incidence on the ground. Details of the results are found under the result and discussion part (Chapter 5).

4.5. Cross-Validation

The leave-one-out cross-validation was applied to test the predictive performance of the regression kriging model. The model performance is measured using the parameters derived from the cross-validation process; mean error (ME) and root mean squared error (RMSE). Hence, it produced quantitative results based on which the relatively best performing models were selected and applied.

5. CHAPTER 5

Result and Discussion

For this study, the annual cholera epidemiological time-series dataset from all countries of the study region, sub-Saharan Africa were collected from the WHO online data repository from 2006 to 2015. The annual temperature and annual precipitation raster/grids were downloaded from ([https://giovanni.gsfc.nasa.gov/giovanni/...*](https://giovanni.gsfc.nasa.gov/giovanni/...)) and were processed in ArcGIS to compute the minimum, maximum and mean annual temperature and precipitation dataset that were incorporated into the spatial regression model as explanatory variables to investigate the spatial variability of the risk of cholera incidence across the region. Results are discussed under the following respective topics.

5.1. Descriptive statistics of the study data

The summary statistics of cholera incidence dataset in Sub-Saharan Africa from 2006 to 2015 showed that the mean continuously changed through the years showing that cholera incidence had randomly distributed, meaning that there were a lot of outliers which showed peak of the disease incidence and missing data in the dataset recorded in all respective sub-regions, which is a normal behaviour of the reality on the ground as there is no ideal normality in uninterrupted natural world. Even though each country of Sub-Saharan Africa had experienced a varying degrees of risks depending on the existing climatic and environmental conditions in each individual country, the mean values of the disease events in all countries in each year revealed that cholera incidence exhibited a remarkable declining trends from 2006 to 2015. However, in 2011 and 2014 that the mean values regained a rise up to 4600.78 (CI 518.56, 8683.01) and 2567.73 (CI 185.57, 4949.89) respectively. On the other hand, these falling patterns in the mean values of cholera cases over time as opposed to the current punitive global and regional climate variability and environmental degradation indicate that there might be other related factors that pushed to decrease the risk such as the improved socio-economic and cultural practice of the society across the region, improved public health infrastructures, the global and regional governmental interventions to mitigate the disease risk and/or related that, due to data unavailability, were not considered in this particular study. The summary table of cholera incidence in Sub-Saharan Africa from 2006 – 2015 is presented as the following table (Table 2).

	Ob_case s2006	Ob_case s2007	Ob_case s2008	Ob_case s2009	Ob_case s2010	Ob_case s2011	Ob_case s2012	Ob_case s2013	Ob_case s2014	Ob_case s2015
Mean	6385.05	4994.68	4394.51	5634.24	2545.05	4600.78	2868.68	1376.12	2567.73	1615.20
Standard Error	2269.89	1663.31	1663.37	1940.10	1134.33	2082.77	1122.73	695.96	1215.39	642.38
Median	870.00	179.00	972.00	159.00	32.00	117.00	187.00	23.00	0.00	0.00
St. Deviation	14534.37	10650.38	10650.75	12422.70	7263.24	13336.22	7189.00	4456.30	7782.27	4113.25
Sample Variance	2112477 89.25	1134305 84.37	1134385 54.36	1543235 46.69	5275463 3.00	1778547 69.43	5168166 5.52	1985857 5.76	6056377 7.05	1691884 8.96
Kurtosis	9.57	5.99	19.53	16.15	29.16	23.21	10.06	28.43	11.67	9.47
Skewness	3.08	2.57	4.16	3.66	5.14	4.50	3.18	5.08	3.50	3.07
Range	67257.00	41643.00	60055.00	68153.00	44456.0	77636.00	33661.0	26944.0	35996.0	19182.0
Minimum	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Maximum	67257.00	41643.00	60055.00	68153.00	44456.00	77636.00	33661.00	26944.00	35996.00	19182.00
Sum	261787.0 0	204782.0 0	180175.0 0	231004.0 0	104347.0 0	188632.0 0	117616.0 0	56421.00 0	105277.0 0	66223.00 0
Count	41.00	41.00	41.00	41.00	41.00	41.00	41.00	41.00	41.00	41.00
95% CI: Lower	1936.07	1734.60	1134.31	1831.65	321.77	518.56	668.13	12.05	185.57	356.13
Upper	10834.03	8254.77	7654.71	9436.84	4768.33	8683.01	5069.24	2740.20	4949.89	874.26

Table 2. Descriptive statistics of cholera incidence in Sub-Saharan Africa, 2006-2015

5.2. Standardized mortality ratio (SMR)

Standardized mortality ratio is an essential measure of severity of the relative risk which is computed as the ratio of the observed number of disease cases in each stratum to the expected disease risk within each stratum as long as there is a disease incidence within the study region. The SMR greater than one means that the risk is more severe than expected which in turn shows that there is a severe risk of death in a given area, whereas $SMR < 1$ means that there is a lower risk of death than the observed one. The crude SMR results of cholera incidence in sub-Saharan Africa from 2006 to 2015 have shown significant discrepancies among all countries of the study region ranging from zero to 41 per country per year. Among the 41 countries of the study region, there have been five out-standing countries with relatively higher SMR. Guinea (2008) was seen to have the highest SMR, while, Djibouti (2010), Senegal (2011), Zambia (2006) and Angola (2006) have been identified to stand second, third, fourth and the fifth countries respectively (Figure 4).

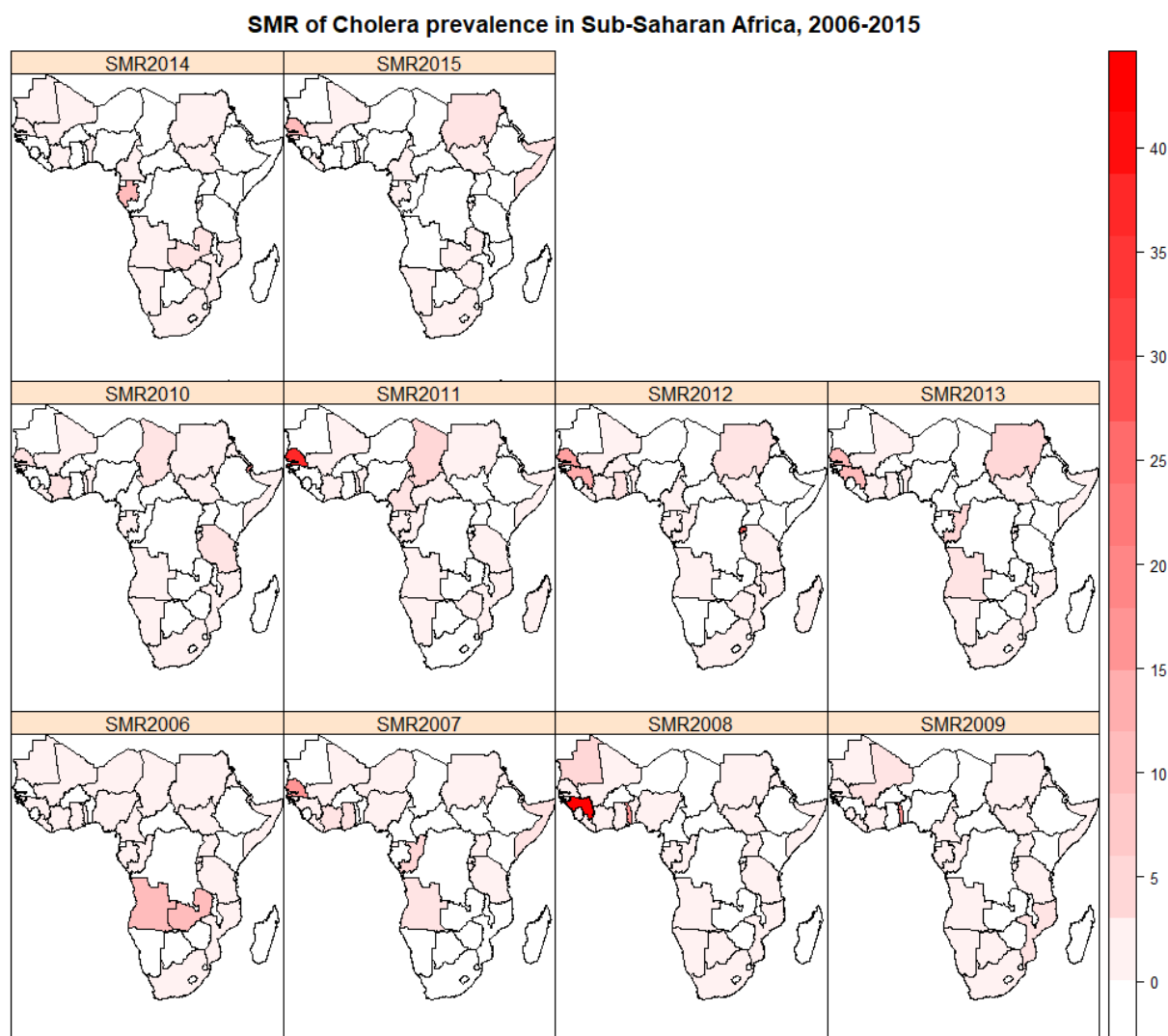


Figure 4. Relative risk map by SMR

5.3. Spatial regression modelling; the generalized linear model (GLM)

The regression coefficients are extracted from the GLM models of cholera case counts in Sub-Saharan Africa from 2006 to 2015. The estimated regression coefficients indicated that there was weaker correlation between the target response variable (cholera count) and the explanatory variables (mean annual temperature and mean annual precipitation). However, the variables are equally significant in all the years with p-values less than $2e-16$. The two covariates were incorporated into the GLM model to estimate the model coefficients (Table 2).

Parameters	Models							
	2006				2007			
	Estimate	Std.Error	Z-value	P-value	Estimate	Std.Error	Z-value	P-value
(Intercept)	23.1218502	0.1818921	127.12	<2e-16	5.8003071	0.2175662	26.66	<2e-16
Mean Temperature	-0.0480211	0.0006074	-79.06	<2e-16	0.0095040	0.0007254	13.10	<2e-16
Mean Precipitation	-0.3563166	0.0336178	-10.6	<2e-16	-1.0293669	0.0346272	-29.73	<2e-16
	2008				2009			
(Intercept)	29.3251900	0.2284583	128.4	<2e-16	44.833202	0.177662	252.4	<2e-16
Mean Temperature	-0.0954171	0.0007675	-124.3	<2e-16	-0.146769	0.000596	-246.2	<2e-16
Mean Precipitation	-8.3367658	0.0451918	-184.5	<2e-16	-11.128632	0.051345	-216.7	<2e-16
	2010				2011			
(Intercept)	-28.029977	0.379976	-73.77	<2e-16	-1.197e+02	4.589e-01	-260.95	<2e-16
Mean Temperature	0.091240	0.001267	72.02	<2e-16	3.991e-01	1.514e-03	263.63	<2e-16
Mean Precipitation	6.386054	0.071086	89.84	<2e-16	8.805e-01	6.487e-02	13.57	<2e-16
	2012				2013			
(Intercept)	-36.136840	0.365647	-98.83	<2e-16	12.95547	0.39259	33.00	<2e-16
Mean Temperature	0.142616	0.001212	117.69	<2e-16	-0.02079	0.00131	-15.88	<2e-16
Mean Precipitation	10.732142	0.049217	218.06	<2e-16	4.01999	0.06032	66.65	<2e-16
	2014				2015			
(Intercept)	-1.039e+02	5.441e-01	-191.0	<2e-16	-18.64890	0.509588	-36.60	<2e-16
Mean Temperature	3.411e-01	1.799e-03	189.6	<2e-16	0.059465	0.001693	35.13	<2e-16
Mean Precipitation	1.278e+01	6.000e-02	213.0	<2e-16	7.920230	0.092742	85.40	<2e-16

Table 3. Summary of the model (GLM)

P-values of all individual covariates are equally significant with "*". Hence, the mean annual temperature and mean annual precipitation are incorporated into the model as explanatory variables*

5.4. Prediction of Cholera risk in Sub-Saharan Africa, 2006

The exponential variogram model fitted to the sample variogram computed from the log-transformed GLM model (Model2006) residuals with sill = 0.020225, range = 1250.317 and nugget = 0.0 values depicted a remarkable spatial autocorrelation among the observation locations (the centroids of the 41 countries of Sub-Saharan Africa). From the risk isopleth map of cholera incidence, the hottest areas indicated that there was relatively higher risk of cholera

incidence in the area. Accordingly, Angola, Eritrea, Sudan and Uganda were found to be the highest cholera prevalent countries in the year 2006 (Figure 5 (right)). The results from cross-validation have been given in the table below (Table 4).

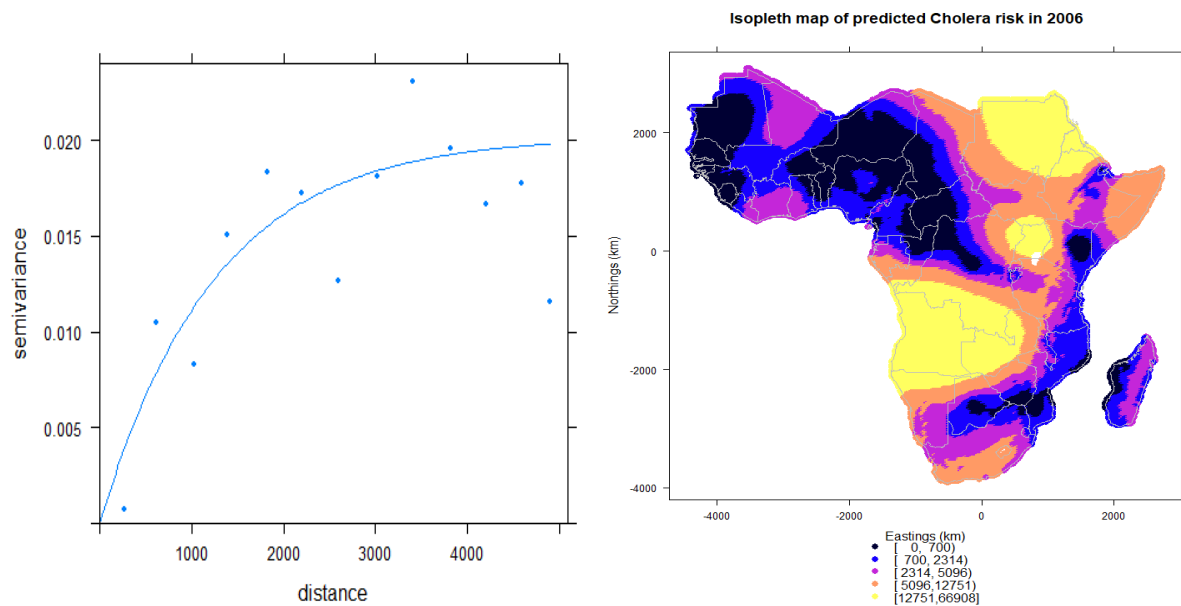


Figure 5. Exponential model variogram with cutoff=4000 width=500 (left), Predicted risk (right)

Prediction method: Regression kriging			
Variogram model: Exponential			
SSEr	Sill	Range	Nugget
3.2e-09	0.02022467	1250.317	0.0

Table 4. Variogram model parameters, 2006

5.5. Prediction of Cholera incidence in Sub-Saharan Africa in 2007

The exponential variogram model with the optimal model parameters; sill = 0.008592, range = 581.8773 and the nugget = 0 fitted to the GLM model (Model2007) log-transformed residuals. The exponential variogram model captured the inherent spatial autocorrelation among the cholera cases at different observation locations (centroids of the 41 countries of the study region). According to the predicted risk map, Senegal, Sudan, Somalia, and Angola were seen to have experienced higher cholera risk in 2007. From the isopleth maps from the two years; 2006 and 2007, the risks of cholera incidence exhibited nearly similar spatial patterns in the region except in 2007; there were additional locations (countries) with higher cholera incidence (Figure 6 (right)).

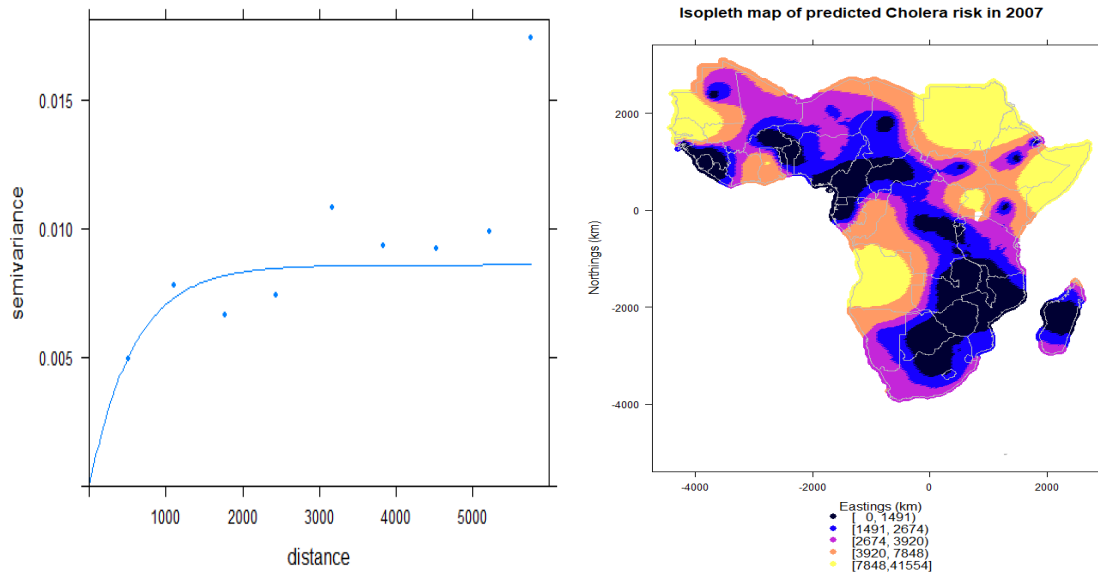


Figure 6. Exponential model variogram (left), Predicted risk (right)

Prediction method: Regression kriging			
Variogram model: Exponential cutoff=6000 width=700			
SSEr	sill	Range	Nugget
2.57e-10	0.008592	581.8773	0.00

Table 5. Variogram model parameters, 2007

5.6. Prediction of Cholera incidence in Sub-Saharan Africa in 2008

The linear model variogram showed that there are spatial discontinuity at the origin due to the nugget effect which is a small scale variability among the observed cholera cases at a distance less than the smallest possible separation (h) between observation points within the study region. This means that there is no spatial autocorrelation at that particular observation point implying that the variables are purely random. Hence, the possible predictor might be the non-spatial mean component only in this case. The isopleth map of the risk of cholera in the region also evidenced that the predicted risk is not smoothed prediction. Hence the predicted risk has a lot of noise (Figure 7). The variogram model parameters and cross-validation results are given in the table.

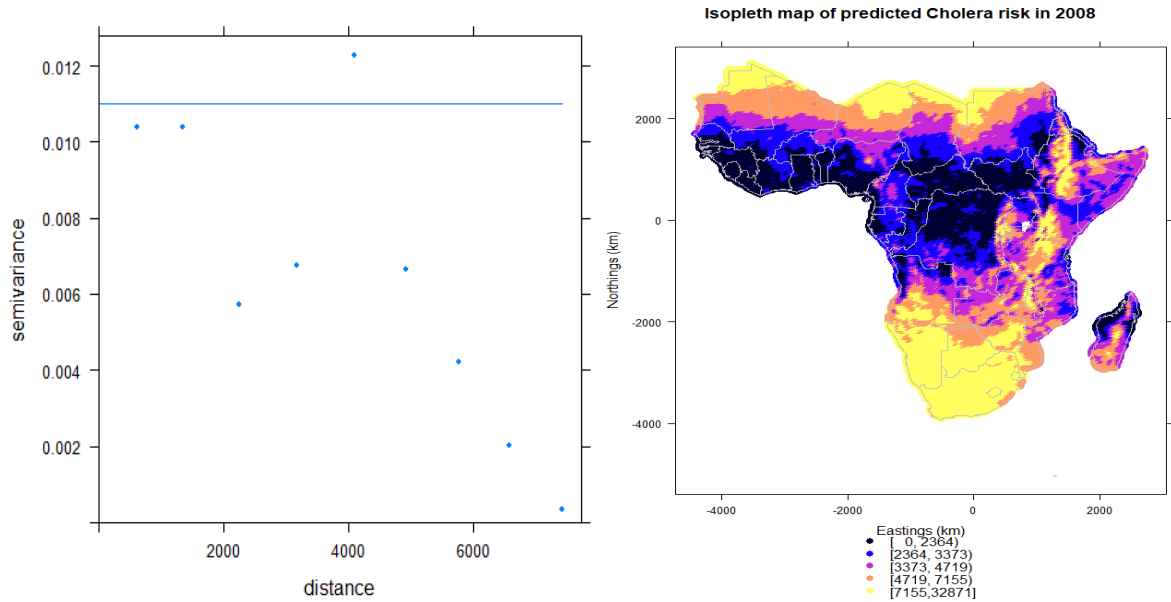


Figure 7. Exponential model variogram (left), Predicted risk (right)

Prediction method: Regression kriging			
Variogram model: Linear			
Cutoff = 10000 Width = 900			
SSEr	Sill	Range	Nugget
0.000115	0.00	2589.445	0.010995

Table 6. Variogram model parameters, 2008

5.7. Prediction of Cholera incidence in Sub-Saharan Africa in 2009

The experimental variogram was computed for cholera incidence, and the linear model variogram depicted that there is a weaker but positive linear spatial autocorrelation among the observation locations within the study area. The linear variogram model shows that there is a trend in the data in which the spatial variability tends to increase linearly with the spatial separation between observation locations. Not only that, but the nugget value being different from zero also revealed that there is a spatial discontinuity somewhere at or near the origin which implies that there is a weaker spatial autocorrelation in the dataset and hence, the predicted risk was most by the non-spatial trend component (large scale trend) plus the weaker spatial random effect (Figure 8) and (Table 7).

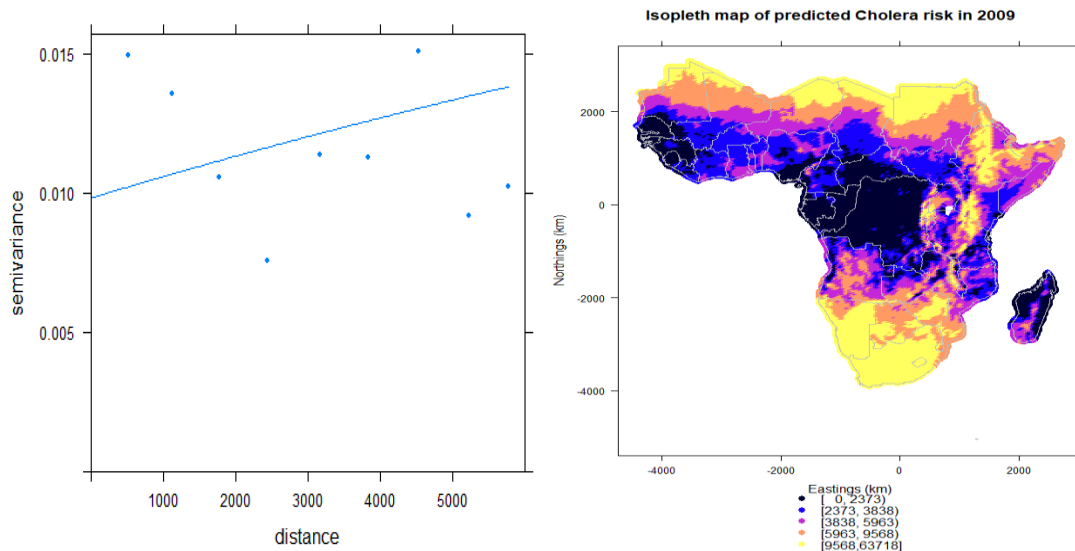


Figure 8. Linear model with cutoff=6000 width=400 (left), Predicted risk (right)

Prediction method: Regression kriging			
Variogram model: Linear Cutoff = 6000 Width = 400			
SSEr	Sill	Range	Nugget
4.7e-08	0.017566	4872.862	0.009824

Table 7. Variogram model parameters 2009

5.8. Prediction of Cholera incidence in Sub-Saharan Africa in 2010

The exponential variogram model fitted the sample variogram computed from the log-transformed residuals from the GLM model depicted a remarkable spatial correlation among the observation locations. The isopleth map of cholera risk (right) showed a defined patterns of the disease incidence across the study region which would lead to the meaningful interpretation of the disease clusters in the region (Sub-Saharan Africa) (Figure 9) and (Table 8).

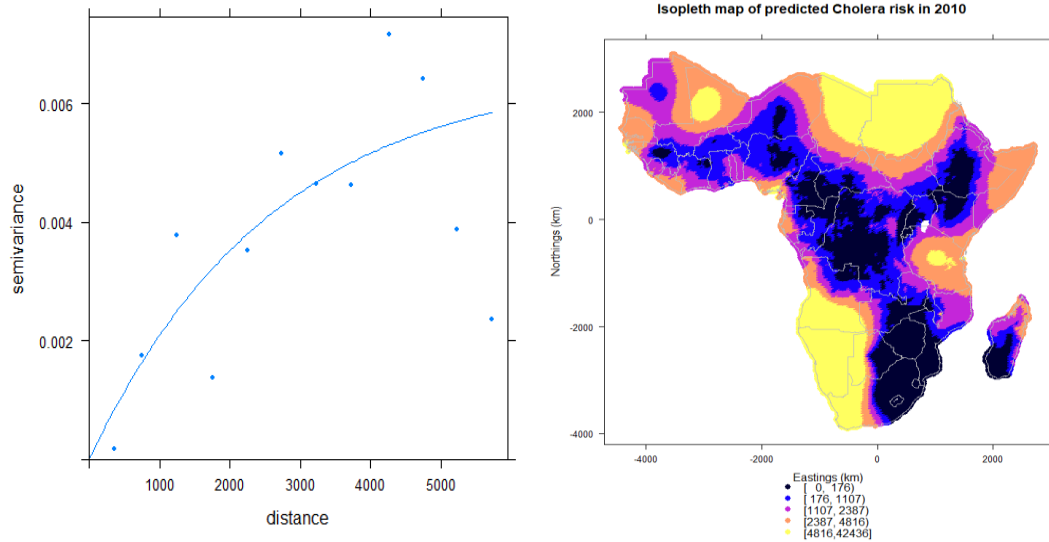


Figure 9. Exponential variogram cutoff=6000 width=500 (left), Predicted risk (right)

Prediction method: Regression kriging			
Variogram model: Exponential Cutoff = 6000 Width = 500			
SSEr	Sill	Range	Nugget
2.93e-10	0.0065867	2610.631	0.00

Table 8. Variogram model parameters, 2010

5.9. Prediction of Cholera incidence in Sub-Saharan Africa in 2011

The linear model variogram depicted a linear increase in spatial variability or trend in the data. Whereas the nugget effect is a clear indication of a spatial discontinuity at or near the origin. Hence, there would be less or no spatial autocorrelation among the observed location points. Thus, the prediction was handled by the non-spatial mean (large scale trend) component predicted from the GML model (Figure 10).

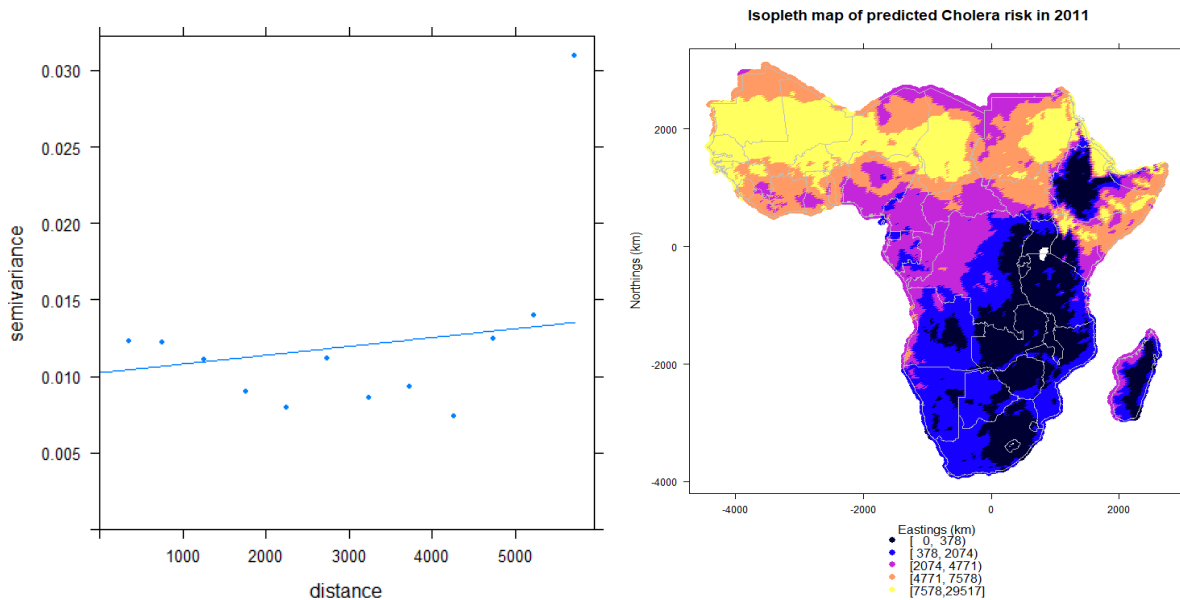


Figure 10. Linear model variogram (left), Predicted risk (right)

Prediction method: Regression kriging			
Variogram model: Linear cutoff=6000 width=500			
SSEr	Sill	Range	Nugget
1.63e-09	0.003321	5751	0.0102

Table 9. Variogram model parameters, 2011

5.10. Prediction of Cholera incidence in Sub-Saharan Africa in 2012

The existence of the nugget effect in linear model variogram means that there is no further continuity in the observed locations which indicates that there is weak or no meaningful spatial autocorrelation in the data. The observed variability may be due to the nugget effect. Thus, the non-spatial trend component would be used to predict the risk in this case (Figure11).

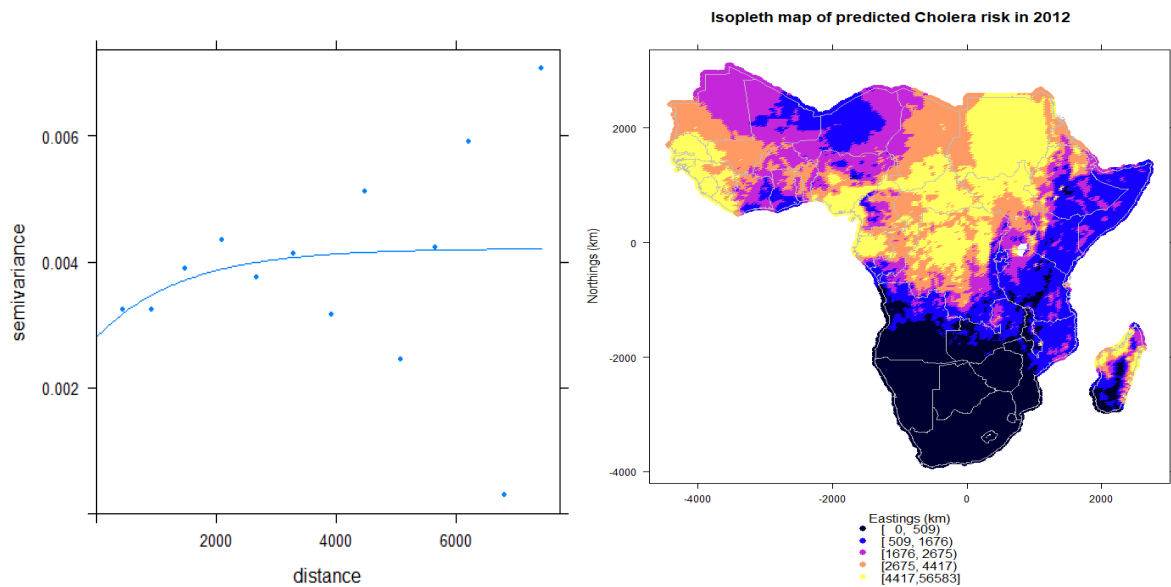


Figure 11. Linear model variogram (left), Predicted risk (right)

Prediction method: Regression kriging			
Variogram model: Exponential Cutoff = 100000 Width = 600			
SSEr	Sill	Range	Nugget
3.26e-11	0.001412084	1445.25	0.002802023

Table 10. Variogram model parameters, 2012

5.11. Prediction of Cholera incidence in Sub-Saharan Africa in 2013

The exponential variogram model fitted to the experimental sample variogram showed that there is a remarkable spatial autocorrelation among the observation locations. Hence, the generated isopleth map clearly shows that there is a clear spatial pattern in cholera incidence across the region. The RMSE is relatively lower when compared to the rest of cholera incidence which shows the strength of the predictive capability of the kriging prediction. (Figure12).

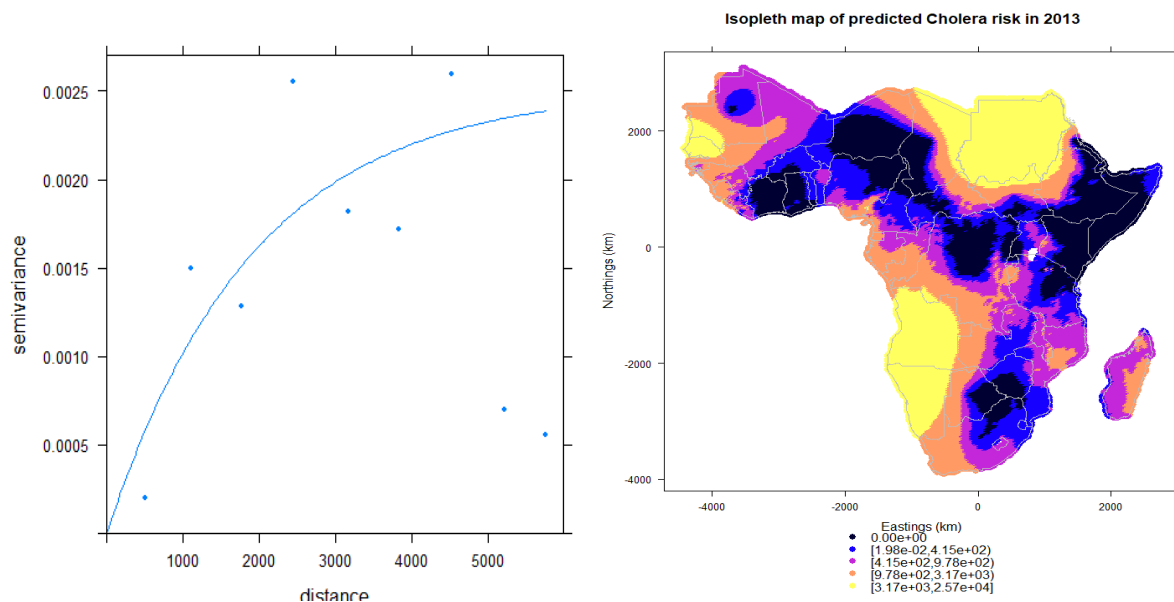


Figure 12. Exponential variogram (left), Predicted risk (right)

Prediction method: Regression kriging			
Variogram model: Exponential Cutoff = 6000 Width = 700			
SSer	Sill	Range	Nugget
6.46e-08	0.002513	1922.535	0.00

Table 11. Variogram model parameters, 2013

5.12. Prediction of Cholera incidence in Sub-Saharan Africa in 2014

The fitted exponential variogram model showed that there is a remarkable spatial autocorrelation, while the nugget effect shows that there is a spatial discontinuity at or near the origin the semi-variogram model plotted as the following. The isopleth map of cholera incidence somehow looks blurred, but one can identify there is a similar trend among the closer observation point than the furthest ones (Figure 13).

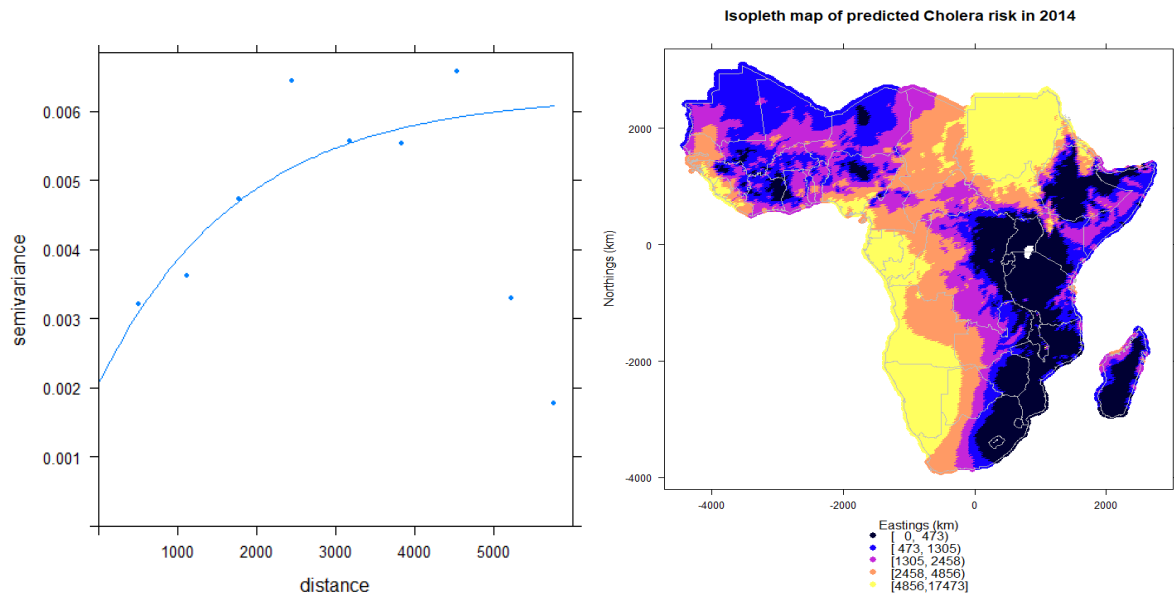


Figure 13. Exponential variogram (left), Predicted risk (right)

Prediction method: Regression kriging			
Variogram model: Exponential Cutoff = 6000 Width = 700			
SSEr	Sill	Range	Nugget
8.53e-11	0.004183254	1776.422	0.0020594

Table 12. Variogram model parameters, 2014

5.13. Prediction of Cholera incidence in Sub-Saharan Africa in 2015

The exponential variogram model fitted to the sample variogram which was computed from the GLM model residuals also showed that there was remarkable autocorrelation among location of the observed cholera incidence in the region. The optimal model parameters from the variogram model and the cross-validation results are given in the table (Table 13). The spatial patterns observed in the produced isopleth maps resembled the spatial patterns observed in the crude SMR map which was mapped from the raw count of the disease (Figure 14).

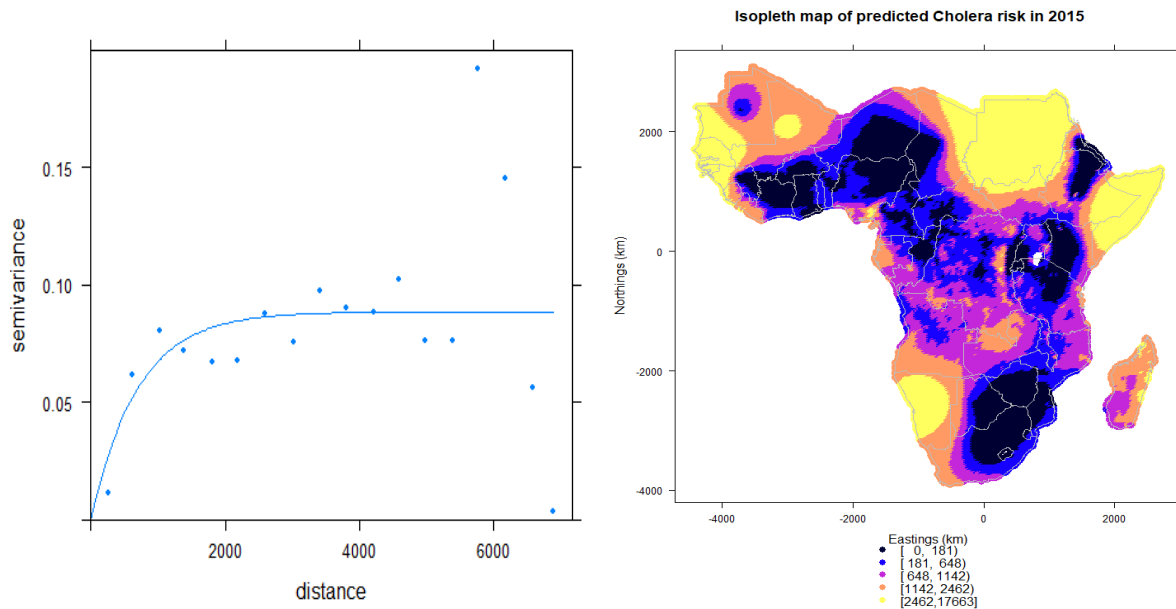


Figure 14. Exponential variogram (left), Predicted risk map (right)

Prediction method: Regression kriging			
Variogram model: Exponential			
Cutoff = 6000 Width = 700			
SSER	Sill	Range	Nugget
7.34e-08	0.08866104	700.4872	0.00

Table 13. Variogram model parameters, 2015

5.14. Cross-validation of the regression kriging prediction

The study utilized cholera case sample dataset from 41 observation locations (centroids of each 41 country) which were so scarce when compared to the study area size. Hence, the cross-validation (leave-one-out) technique was applied in which one observation point was taken out and was predicted using the remaining observation points around it in order to test and justify the predictive capability of regression kriging prediction of the risk of cholera incidence. The regression kriging predictions with smaller ME and RMSE values were taken to the final prediction of the disease risk across the study region (Table 14).

2006			2007			2008		
ME	MSE	RMSE	ME	MSE	RMSE	ME	MSE	RMSE
450.31	233598006	15283.91	96.930	157043636	12531.71	-55.512	120428674	10974
2009			2010			2011		
ME	MSE	RMSE	ME	MSE	RMSE	ME	MSE	RMSE
-118.0	158262964	12580.26	152.353	65924325	8119.38	-10.587	193059645	13894.59
2012			2013			2014		
ME	MSE	RMSE	ME	MSE	RMSE	ME	MSE	RMSE
32.110	54494920	7382.067	81.343	25375627	5037.423	-31.877	67672370	8226.32
2015								
ME	MSE	RMSE						
29.94	16657092	4081.31						

Table 14. Cross-validation of the regression kriging prediction

6. CHAPTER 6

Conclusion and Recommendation

6.1. Conclusion

Interpolation of the annual regional cholera dataset over a continuous surface to investigate the spatial trend of the disease incidence was the primary objective of the study. This study utilized the ten years (2006 – 2015) dataset of the annual epidemiological cholera case collected from the WHO online database, the annual temperature and precipitation raster downloaded and preprocessed to convert them into the interoperable and compatible format for the study. The mean annual temperature and mean annual precipitation values were extracted from the raster datasets. In the study, two statistical methods were adopted; the generalized linear model (GLM) for modelling the disease spatial relationship with the potential covariates and the Regression Kriging (RK) interpolation for disease spatial prediction and risk mapping. The GLM modelling revealed that there was significant relationship between cholera incidence and climatic and environmental covariates across Sub-Saharan Africa. The study tried to answer the research questions as follows.

- The Pearson's product moment correlation test was performed to test the strength of the correlation between cholera incidence and the climatic and environmental factors; they were significantly correlated to the disease incidence across the region with p-values $<2e-16$ (Table 3).
- The disease maps were produced using the geostatistics analytical modelling; the generalized linear model for spatial regression, variogram modeling for the investigation of the disease spatial structure among every pair of observation location, regression Kriging to map the risk of cholera incidence across Sub-Saharan Africa.
- The research findings revealed that there were strong spatial autocorrelations among the observed risks of cholera incidence across the region in the years; 2006, 2007, 2010, 2013, 2014 and 2015. Thus, the predicted risk maps were able to show the well clustered cholera incidence across the region (Figures: 5, 6, 9, 12, 13, 14)

- However, in 2008, 2009, 2011 and 2012 there were weaker spatial autocorrelations among the observed cholera incidence across the region and hence, the predicted risk maps could capture less clusters (Figures: 7, 8, 10, 11)
- As it was evidenced from the isopleth maps of the years; 2006, 2007, 2010, 2013, 2014 and 2015, there were significant clusters throughout the region. However, in the rest of the years; 2008, 2009, 2011 and 2012 the maps showed that there were less clusters in cholera incidence across the region.

6.2. Recommendation

The following are some points to recommend for further research

- To further extend the research work with additional explanatory variables such as socio-economic variables; household income, education, gender, and access to public infrastructure to fully understand the spatial dynamics of cholera across the region
- It would be better to obtain an adequate sample size to capture a more meaningful spatial distribution of the disease prevalence across the region.
- To use the fine spatial resolution data analysis of disease risk to produce detailed prediction of the risk map (e.g. sub-region level)

REFERENCE

- Ainsworth, L. M., & Dean, C. B. (2006). Approximate inference for disease mapping. *Computational Statistics and Data Analysis*, 50(10), 2552–2570. <https://doi.org/10.1016/j.csda.2005.05.001>
- Ali, M., Nelson, A. R., Lopez, A. L., & Sack, D. A. (2015). Updated Global Burden of Cholera in Endemic Countries. *PLOS Neglected Tropical Diseases*, 9(6), e0003832. <https://doi.org/10.1371/journal.pntd.0003832>
- Bandyopadhyay, S., Kanji, S., & Wang, L. (2012). The impact of rainfall and temperature variation on diarrheal incidence in Sub-Saharan Africa. *Applied Geography*, 33(1), 63–72. <https://doi.org/10.1016/j.apgeog.2011.07.017>
- Berke, O. (2004). Exploratory disease mapping: kriging the spatial risk function from regional count data. <https://doi.org/10.1186/1476-072X-3-18>
- Constantin de Magny, G., & Colwell, R. R. (2009). Cholera and climate: a demonstrated relationship. *Transactions of the American Clinical and Climatological Association*, 120(May 2014), 119–128.
- Deen, J. L., von Seidlein, L., Sur, D., Agtini, M., Lucas, M. E. S., Lopez, L., ... Clemens, J. D. (2008). The high burden of cholera in children: Comparison of incidence from endemic areas in Asia and Africa. *PLoS Neglected Tropical Diseases*, 2(2). <https://doi.org/10.1371/journal.pntd.0000173>
- Johnson, S., & Collection., R. D. S. (2007). *The ghost map : the story of London's most terrifying epidemic-- and how it changed science, cities, and the modern world.*
- Koelle, K. (2009). The impact of climate on the disease dynamics of cholera. <https://doi.org/10.1111/j.1469-0691.2008.02686.x>
- Kutner, M. H. (2005). *Applied Linear Statistical Models*. Retrieved from http://books.google.fr/books?id=0xqCAAACA AJ&dq=intitle:Applied+linear+statistical+models+djvu&hl=&cd=1&source=gbs_api
- Lance A. Waller, C. A. G. (2004). *Applied Spatial Statistics for Public Health Data*. *Journal of the American Statistical Association* (Vol. 100). <https://doi.org/10.1198/jasa.2005.s15>
- Lawson, A. (2010). Statistical Methods for Disease Clustering. *Annals of Epidemiology*, 20(12), 964. <https://doi.org/10.1016/j.annepidem.2010.07.101>
- Lawson, A., Banerjee, S., Haining, R., & Ugarte, L. (2016). *Handbook of Spatial Epidemiology*.
- Lessler, J., Moore, S. M., Luquero, F. J., McKay, H. S., Grais, R., Henkens, M., ... Azman, A. S. (2018a). Mapping the burden of cholera in sub-Saharan Africa and implications for control: an analysis of data across geographical scales. *Lancet (London, England)*, 391(10133), 1908–1915. [https://doi.org/10.1016/S0140-6736\(17\)33050-7](https://doi.org/10.1016/S0140-6736(17)33050-7)
- Lessler, J., Moore, S. M., Luquero, F. J., McKay, H. S., Grais, R., Henkens, M., ... Azman, A. S.

- (2018b). Mapping the burden of cholera in sub-Saharan Africa and implications for control: an analysis of data across geographical scales. *The Lancet*, *391*(10133), 1908–1915.
[https://doi.org/10.1016/S0140-6736\(17\)33050-7](https://doi.org/10.1016/S0140-6736(17)33050-7)
- Lo, S., & Andrews, S. (2015). To transform or not to transform: using generalized linear mixed models to analyse reaction time data. *Frontiers in Psychology*, *6*(August), 1–16.
<https://doi.org/10.3389/fpsyg.2015.01171>
- Lobitz, B., Beck, L., Huq, A., Wood, B., Fuchs, G., Faruque, a S., & Colwell, R. (2000). Climate and infectious disease: use of remote sensing for detection of *Vibrio cholerae* by indirect measurement. *Proceedings of the National Academy of Sciences of the United States of America*, *97*(4), 1438–1443. <https://doi.org/10.1073/pnas.97.4.1438>
- Luquero, Francisco J. (2009). Time-series analysis of cholera in Guinea-Bissau (1996–2008), *33*(December), 1996–2008.
- Magny, G. C. De, Guégan, J., Petit, M., & Cazelles, B. (2007). *BMC Infectious Diseases*, *9*, 5–8.
<https://doi.org/10.1186/1471-2334-7-20>
- Mengel, M. A., Delrieu, I., Heyerdahl, L., & Gessner, B. D. (2014). Cholera Outbreaks in Africa (pp. 117–144). Springer, Berlin, Heidelberg. https://doi.org/10.1007/82_2014_369
- Moraga, P. (2017). SpatialEpiApp : A Shiny web application for the analysis of spatial and spatio-temporal disease data. *Spatial and Spatio-Temporal Epidemiology*, *23*, 47–57.
<https://doi.org/10.1016/j.sste.2017.08.001>
- Noel A. C. Cressie, 1993. (1993). *Statistics for Spatial Data* (Revised Ed). New York: JOHN WILEY & SONS, INC.
- Osei, F. B., Duker, A. A., & Stein, A. (2011). Hierarchical Bayesian modeling of the space-time diffusion patterns of cholera epidemic in Kumasi, Ghana. *Statistica Neerlandica*, *65*(1), 84–100.
<https://doi.org/10.1111/j.1467-9574.2010.00475.x>
- Osei, F. B., Duker, A. A., & Stein, A. (2012). Bayesian structured additive regression modeling of epidemic data: application to cholera. *BMC Medical Research Methodology*, *12*(1), 118.
<https://doi.org/10.1186/1471-2288-12-118>
- Reyburn, R., Kim, D. R., Emch, M., Khatib, A., Von Seidlein, L., & Ali, M. (2011). Climate variability and the outbreaks of cholera in Zanzibar, East Africa: A time series analysis. *American Journal of Tropical Medicine and Hygiene*, *84*(6), 862–869.
<https://doi.org/10.4269/ajtmh.2011.10-0277>
- Robinson, M., & Dietrich, S. (2016). An Introduction to Spatial Autocorrelation and Kriging Tobler and Spatial Relationships.
- Wakefield, J. (2007). Disease mapping and spatial regression with count data. *Biostatistics*, *8*(2), 158–183. <https://doi.org/10.1093/biostatistics/kxl008>
- Waller, L. A., & Gotway, C. A. (2004). *Applied spatial statistics for public health data*. John Wiley

& Sons. Retrieved from

[https://books.google.nl/books?hl=en&lr=&id=OuQwgShUdGAC&oi=fnd&pg=PR7&dq=spatial+statistics+journal&ots=6q5JnLg_aJ&sig=hdtsBkCiQzHbuG7KSzNEC061s8I#v=onepage&q=spatial statistics journal&f=false](https://books.google.nl/books?hl=en&lr=&id=OuQwgShUdGAC&oi=fnd&pg=PR7&dq=spatial+statistics+journal&ots=6q5JnLg_aJ&sig=hdtsBkCiQzHbuG7KSzNEC061s8I#v=onepage&q=spatial+statistics+journal&f=false)