# FLEXIBLE SPATIAL MODELLING USING COPULA WITH AN APPLICATION TO WATER, SANITATION AND HYGIENE (WASH) INDICATORS IN NAMIBIA

A THESIS SUBMITTED IN PARTIAL FULFILMENT

OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE IN APPLIED STATISTICS AND DEMOGRAPHY

OF

THE UNIVERSITY OF NAMIBIA

BY

ANASTASIA JOHANNES

200618130

APRIL 2019

SUPERVISOR: Professor Lawrence N. Kazembe

## ABSTRACT

Spatial modelling is one approach that can serve as a basic tool in identifying areas that need intervention. Understanding the spatial distribution of indicators is of vital importance, as it help with national planning and decision-making. Different models such as Conditional Autoregressive (CAR) can be used in spatial modelling. However, these models pose analytic and statistical inferential challenges especially when non-continuous and multivariate response are involved. Therefore, a flexible approach such as copula can be used. Copula provides a flexible way to model multivariate outcomes by decomposing information from their joint distribution into that of its marginal distributions and defines the dependence structure. The approach is exemplified to analyze spatial distribution of Water, Sanitation and Hygiene (WASH) indicators.

This study used a quantitative cross sectional study design using secondary data from the 2013 Namibia Demographic Health Survey (NDHS). The outcome variables used were water, sanitation and hygiene, while the covariates were; sex of household head, household size, age of the head of household, type of residence, wealth index, literacy, exposure to media, marital status, respondent occupations, working status and region. Two copula-based spatial modelling approaches were applied to model the discrete outcome variables. First, a Gaussian Copula Marginal Regression (GCMR) was used to estimate the spatial distribution of WASH indicators on geostatistical data. The fitting of GCMR models was done through the method of Maximum Likelihood Estimation (MLE). Secondly, a Generalized Joint Regression Modelling (GJRM) was used to model the joint spatial dependency of WASH indicators on areal data. The Markov Random Field (MRF) smoother was applied to the variable called region, while for continuous and discrete variables such as age and household size, the smooth functions were represented using the spline approach.

Results from the best GCMR model fitted showed that significant spatial disparities for improved drinking water, sanitation and hygiene exist at the regional level across Namibia. In addition, covariate such as sex, wealth index, education level, marital status, literacy, working status and residence were significant with respect to improved drinking water, sanitation and hygiene. Household size was significant to improved drinking water and hygiene, while age was significant to improved sanitation and hygiene. Improved drinking water was found to be low in Kavango, Ohangwena, Omusati and Kunene regions, while improved sanitation was low in Ohangwena and Zambezi regions. However, improved hygiene was much of a concern in Zambezi, Ohangwena, Oshana and Omusati regions.

For the joint model, the results indicated that residence and wealth index were significant to improved drinking water, hygiene and sanitation. It has been observed that there was a high positive correlation between the WASH indicators. In addition, the study has noticed that more regions were found to have a little spatial variation in improved drinking water.

Applying copula provided a flexible approach to the study in identifying areas that need major intervention for future improved drinking water, sanitation and hygiene. The presented maps and analysis approach demonstrated a mechanism for monitoring access

to improved drinking water, sanitation and hygiene. It can also assist policy makers especially those involved in planning to develop comprehensive programmes and develop strategies for areas that were found to have low improved drinking water, sanitation and hygiene.

**Keywords:** Gaussian copula marginal regression, generalized joint regression modelling

# Table of Contents

# LIST OF TABLES

## LIST OF FIGURES

## LIST OF ABBREVIATIONS

| | |
|---|---|
| AIC | Akaike Information Criterion |
| BIC | Bayesian Information Criterion |
| CAR | Conditional Autoregressive |
| CDFs | Cumulative Distribution Functions |
| CE | Continuous Extension |
| CI | Confidence Interval |
| CML | Composite Marginal Likelihood |
| copCAR | copula Conditional Autoregressive |
| DIC | Deviance Information Criterion |
| DHS | Demographic and Health Survey |
| DT | Distribution Transform |
| EA | Enumeration Area |
| GCMR | Gaussian Copula Marginal Regression |
| GIS | Geographic Information System |
| GJRM | Generalized Joint Regression Model |
| GLMM | Generalized Linear Mixed Model |
| GMRF | Gaussian Markov Random Field |
| GPS | Global Positioning System |
| HH | Household |
| HRTWS | Human Right to Safe Drinking Water and Sanitation |

| | |
|---|---|
| IEC | Information Education Communication |
| MCMC | Markov Chain Monte Carlo |
| MDGs | Millennium Development Goals |
| MLE | Maximum Likelihood Estimation |
| MoHSS | Ministry of Health and Social Services |
| MRF | Markov Random Field |
| NDHS | Namibia Demographic and Health Survey |
| NDP5 | The Fifth National Development Plan |
| NPC | National Planning Commission |
| OR | Odds Ratio |
| PMLE | Pseudo Maximum Likelihood Estimation |
| PSUs | Primary Sampling Units |
| SAR | Spatial Autoregressive |
| SDGs | Sustainable Development Goals |
| UN | United Nations |
| UNICEF | United Nations Children’s Fund |
| UREC | University Research Ethics Committee |
| VIP | Ventilated Improved Pit |
| WASH | (improved drinking) Water, Sanitation and Hygiene |
| WHO | World Health Organization |

# ACKNOWLEDGEMENTS

Thank you, Lord Jesus Christ for giving me time, energy, knowledge, wisdom and perseverance to complete my research. This work has been possible only because you have been by my side. Thank you for taking more of me, and giving me more of you.

I would like to express my deepest gratitude to my supervisor Prof. Lawrence Kazembe for his guidance, advice and inspiration throughout my research. His continuous support has really helped me, especially during the data analysis and thesis write up. I would also like to express my gratitude to everyone that I have consulted during my research and to those who have been of assistance.

To my family and friends, thank you for always supporting and encouraging me to keep pushing harder and not to give up. Your support is invaluable.

# DEDICATION

I dedicate this work to my late mother, Beatha K. Iipinge who always believed that I could do better if I try my best. I also dedicate this work to my three (3) amazing aunties (Maria M. Iipinge, Salome N. Iipinge and Elizabeth Iipinge) who always supported and believed in me, and for taking the role of mother, father and aunt in my life.

## DECLARATIONS

I, Anastasia Johannes, hereby declare that this study is my own work and is a true reflection of my research, and that this work, or any part thereof has not been submitted for a degree at any other institution.



I, Anastasia Johannes, grant The University of Namibia the right to reproduce this thesis in whole or in part, in any manner or format, which The University of Namibia may deem fit.

| | | |
|---|---|---|
| Anastasia Johannes | ……………… | …………………. |
| Name of Student | Signature | Date |

# CHAPTER 1: INTRODUCTION

## 1.1 Background of the study

Improved drinking Water, Sanitation and Hygiene (WASH) have been called one of the best remedial advances since 1840 and has been classified as the overlooked establishment of wellbeing (Magalhaes, Barnett, & Clements, 2011). According to Hutton and Chase (2016), WASH are basics for improved standard of living, which include the protection of health and the environment, improved educational outcomes, dignity and gender equality. Furthermore, poor populations have brought down progress to improved WASH services and poorly connected practices. Henceforth, drinking water and sanitation is part of the Sustainable Development Goals (SDGs), Goal 6, which focuses on accomplishing general access to WASH by 2030 (Hutton & Chase, 2016, p. 11). Additionally, the Human Right to Safe Drinking Water and Sanitation (HRTWS) was approved in 2010 under a United Nations (UN) Resolution to guarantee sheltered, moderate, adequate, available, and accessible drinking water and sanitation administration for all. Furthermore, one of the objectives of the Namibia vision 2030 is to have access to portable drinking water for all people (National Planning Commission (NPC), 2004).

Although progress has been made to achieve improved WASH, still many developing countries in the global South lag behind. Roughly, 80% of the general population utilizing water from unimproved sources are amassed in three regions: sub-Saharan Africa, Eastern Asia and Southern Asia. In sub-Saharan Africa, progress was made from 49% coverage in 1990 to 56% in 2004. For sanitation, general levels of utilization of improved facilities are far lower than for drinking water. In 2004, 59% of the world population had access to any type of improved sanitation facility at home, from 49% in 1990 (WHO, UNICEF, 2006). Notwithstanding such enhancement, national and sub-national inequalities in WASH scope endure, including rural-urban and regional disparities. In Namibia, the situation is similar. According to the Namibia Demographic and Health Survey (NDHS) (2014), 87% of households have access to improved drinking water, while only 34% have access to improved sanitation. The survey further indicated that inadequate sanitation is poor in rural areas with only 17% having access to improved sanitation, as compared to urban areas with 49%. Furthermore, 98% of households in urban areas have access to improved drinking water than 76% of households in rural areas. However, improved drinking water scarcity continued to be a concern for the government and the population at large.

According to the United Nations Children's Fund (UNICEF) (2016), 88% of deaths caused by diarrhoea are because of the absence of access to safe drinking water, poor sanitation and hygiene.

Various studies have demonstrated that one of the most inequalities in access to improved drinking water and sanitation relates to location, ordinarily measured in terms of urban versus rural areas (Yu, Bain, Mansour, & Wright, 2014). An analysis that was done in 19 Latin American and Caribbean countries, showed that location adds to the gap in access to improved water for 17 countries (Yu et al., 2014) and this applies to sanitation.

Spatial modelling of WASH indicators is of importance as it can help with the investigation of distribution and patterns of WASH indicators for targeted interventions.

Distinctive models, for example, the Conditional Autoregressive (CAR) models have been used in various spatial modelling studies utilizing the Markov Chain Monte Carlo (MCMC) algorithm. CAR models are for the most part used to provide results measured at areal level. When undertaking CAR modelling of areal data, the adjacency matric that portrays the neighbourhood of the data is characterized. However, CAR model proposed by Besag, York and Mollie (1991) exhibit a single global level of spatial smoothness determined by geographical adjacency and thus not flexible enough to capture complex structure. In addition, CAR model often assume normality, which may not be achieved in this case. According to Hughes and Haran (2013), CAR model poses collinearity problems that leads to poor estimation performance for the fixed effects.

Therefore, this study proposed a joint model for discrete outcomes using copulas as an alternative to CAR model. The significance of copulas lies in the fact that by a method of change, it can be used to generate joint distribution functions for random variables with arbitrary non-uniform marginal (Winkelmann, 2011).

Copulas has attracted much consideration in spatial modelling in the course of recent years (Kazianka & Pilz, 2009). This is because of their adaptability in portraying conditionals and the likelihood of separating marginal and joint effects (Perrone & Müller, 2016). According to Perrone and Müller (2016), copula has been applied in many application areas of applied statistics such as insurances, econometrics, medicine, marketing, spatial extreme events, time series analysis and even sports.

Bhat and Sener (2009) used a copula-based approach for spatial dependence modelling based on a spatial logit structure rather than a spatial probit structure, to model teenagers' physical activity participation levels. They showed that copula method can be used to consider locational information of neighbouring observations as a way to control for spatial dependence.

Copula modelling has also been applied using a Bayesian approach (Smith, 2011). Previous Bayesian work on copula modelling includes that of Huard, Evin, and Favre (2006), who recommended a technique selected between various bivariate copulas and Silva and Lopes (2008) who use Markov Chain Monte Carlo (MCMC) methods to estimate low dimensional parametric copula functions. Pitt, Chan, and Kohn (2006), Hoff (2007) and Danaher and Smith (2011) estimate Gaussian copula regression models using MCMC methods. However, adopting a Gaussian copula does not mean the data are normally distributed. Smith, Gan, and Kohn (2010) extended the work of Pitt et al. (2006) to copulas derived by inversion from skewed t-distributions constructed by hidden conditioning.

Regardless of their boundless use in other quantitative fields, copula methods have not yet been utilized widely in spatial science. According to Kuethe, Hubbs, and Waldolf (2009), copula techniques in the spatial area have beforehand been recommended in a predetermined number of other geographic sciences, for example in geostatistics and geophysics. Kazianka and Pilz (2009) use copula-based models for spatial interpolation in a study of inadvertent arrivals of radioactive materials because of their flexible tool to perform spatial interpolation for non-Gaussian, multimodal and extreme values.

A different approach has been used to model spatial dependence across observational units in discrete choice models (Fleming, 2004). "A large portion of these methodologies are centred on the binary spatial probit model, which is predicated on a multivariate normality assumption to portray the spatial dependence structure" (Bhat, Sener, & Eluru, 2010, p.1). However, this study will instead apply the Gaussian copula approach, to model spatial distribution of improved drinking Water, Sanitation and Hygiene (WASH) indicators in Namibia with respective to regional level.

## 1.2 Statement of the problem

The availability and accessibility of basic household amenities are important in assessing the general welfare and socio-economic condition of a population. An absence of improved drinking water and sanitation facilities combined with poor hygiene impacts contrarily on the sickness weight of numerous children and adults.

Spatial modelling is mostly used to analyse survey data to attain predictive mapping across regions of interest. The most popular strategy for such modelling has been to model spatial distribution. Therefore, spatial modelling is utilised to identify areas that have less access to improved drinking water, sanitation facilities and good hygiene for better planning and guide intervention strategies.

Different models can be used for spatial modelling of WASH indicators such as the CAR, automodel and Generalized Linear Mixed Model (GLMM). Nonetheless, all these models pose numerous analytical, statistical and computational challenges. The CAR model influences random effects to show single global level of spatial smoothness established by geographical adjacency, however, they are not flexible enough to capture the complex structure of WASH indicators found in the residual spatial autocorrelation (Lee, Rushworth, & Sahu, 2014).

Therefore, understanding spatial modelling of WASH indicators requires a flexible approach for spatial analysis, such as copula. There have been recent developments in copula such as the Gaussian copula that provide flexible framework to model dependent responses of any kind (Masarotto & Varin, 2012), semiparametric techniques that have been fuelled as a result of hidden peripheral dispersion and pair-copula which permits to form flexible models through vines (Fermanian, 2017). Copula brings in a perspective that defines marginal distributions through a joint distribution of random variables. In addition, copula investigates techniques that deal with areal and geostatistical data and these are the data that will be used in this study. Copula has not been widely applied in the area of WASH indicators, therefore there is a need for a flexible approach to analyse the WASH indicators.

## 1.3 Objectives of the study

The main objective of this study was to fit a copula-based approach to model spatial distribution of WASH indicators in Namibia to better guide programs of intervention and allocate resources effectively.

The specific objectives were:

1. To estimate the spatial distribution of WASH indicators using copula on geostatistical data.
2. To model the joint spatial dependency of WASH indicators by applying copula approach on areal aggregated data.
3. To quantify the determinants of WASH indicators.

## 1.4 Significance of the study

Over the recent years, there has been development of model spatial data using copula. Copulas shape a valuable idea for modelling many multivariate distributions. Knowledge of the WASH indicators distribution is essential for authorities that seek to use limited resources by identifying areas that need to advance WASH by analysing their geo-reference data. WASH are essential ingredients to ensure human wellbeing. Therefore, access to improved drinking water, adequate sanitation and good hygiene were included in the Fifth National Development Plan (NDP5) and the Harambee Prosperity Plan. Thus, this study provided a premise in planning spending execution to guarantee that resources are assigned appropriately and develop Information Education Communication (IEC) materials to fortify behaviour change around hygiene. Enhancing these services will convey monetary additions while also helping to construct adaptability given expanding atmosphere changeability.

Unsafe drinking water, poor sanitation and inadequate hygiene are a portion of the main sources of death among youngsters younger than five on the planet (Eid, 2015) and Namibia is no exception. Hence, the results from this study would guide regional councils, municipalities, and local authorities in Namibia in their project implementations, as to which areas to target in order to improve WASH. This will then be able to decrease the occurrences of  diseases that are associated with WASH, such as diarrhoea and cholera. In addition, there is a lack of information with regards to the spatial modelling of WASH using copula in Namibia.

## 1.5 Delimitation of the study

This study covered the whole of Namibia's representation of the NDHS. It will focus on using copula-based approach to model WASH because copula allows for modelling of spatial dependence across an observational unit with discrete variables.

## 1.6 Organisation of the Thesis

The thesis is organized as follows. Chapter 2 is the methodological review of spatial modelling and copula. Chapter 3 deals with how the study data was collected and how the analysis were carried out. Chapters 4 & 5 narrate the results of the objectives. Discussion of the results, conclusions and recommendations are presented in Chapter 6.

# CHAPTER 2: METHODOLOGICAL REVIEW

## 2.1 Introduction

Spatial modelling is one approach that is utilized to indicate the potential distribution of WASH indicators. Such distribution models has turned into an essential tool for planning and basic leadership. There has been an increase in the study of spatial distribution due to the low cost of Geographic Information System (GIS), advance statistical methodology and geographic referenced health and environmental data. Advances in technology now permit not only disease mapping but also the application of spatial statistical methods, such as cluster analysis (Rosenberg et al., 1999; Kulldorff et al., 2005). Despite several studies using spatial modelling, fewer studies have analyzed the spatial distribution of WASH indicators.

Distinctive models such as geostatistics approach, Gaussian Markov Random Field (GMRF), hybrid approach, Spatial Autoregressive (SAR) and Conditional Autoregressive (CAR) models are used as part of spatial modelling. Geostatistics, which depend on the theory of regionalized is progressively favored in light of the fact that it enables one to capitalize on the spatial correlation between neighbouring observations to foresee quality at unsampled locations (Goovaerts, 1999). GMRFs are also known as Conditional Autoregressive (CARs) following the work of Besag (1974, 1975). According to Rue and Martino (2007), GMRFs have a broad use in statistics, with imperative applications in structural time-series analysis, analysis of longitudinal and survival data, graphical models and spatial statistics.

In the CAR model, spatial dependence is expressed through the mean term by setting the expected value of observations in a region to be a function of the adjacent areas' means (Besag, 1974). Customarily, geostatistical models are viewed as more adaptable in catching the spatial dependence when an extensive number of perceptions can be measured at various locales. However, the disadvantage to these models is the computational weight required and in addition, the constraint of the accessible models for non-stationary covariance structures (Besag, 1974).

Spatial Autoregressive (SAR) by Cliff and Ord (1973) is a standard tool for analysing data with spatial correlation (Qu & Lee, 2015). It is applied to cases where outcomes of a spatial unit at one location depend on those of its neighbours.

However, spatial modelling often relies on the Gaussian assumption, which is not satisfied for environment process (Kazianka & Pilz, 2009). Kuethe et al. (2009) pointed out that spatial modelling are often difficult to estimate in applied work due to complex nonlinearities and non-typical conduct of data. Blanchet and Davison (2011) alluded that it has been difficult in modelling events of the spatial process that lack flexible models and appropriate inferential tools.

Therefore, an approach referred to as "*copula*" has recently emerged as an alternative to spatial modelling as it allows to model spatial dependency because of its flexibility. Copula is a multivariate dependence structure for joint distribution of random variables that are parted from the marginal distribution of individual random variables (Trivedi & Zimmer, 2005). It also deals with random fields having discrete marginal distribution.

First, copula links the marginal distribution together to form the joint distribution. Second, copula define nonparametric measures of dependence of pairs of random variables.

Copula can be applied in different application. Bárdossy (2006) used copula to describe spatial variability for ground water quality. Kuethe et al. (2009) used copula approach to model continuous point – generating process leading to spatial point patterns. Kazembe (2017) used copula to model timing and frequency of antenatal care utilization.

## 2.2 Spatial data

Spatial data are otherwise called geospatial data or geographic information. Information or data recognizes the geographic area of features and limits on earth, for example, natural or built highlights, oceans and more. Spatial data is normally put away as directions and topology and is information that can be mapped (Schabenberger & Gotway, 2005).

### 2.2.1 Types of Spatial data

#### 2.2.1.1 Lattice data

Lattice data of location concentrate on the point of the nearest neighbour, second-nearest neighbour and so on (Cressie, 2015). This data either can either be a time-series, longitudinal or clustered data. The lattice data are linked to surveys, census and health statistics, where it enables the provision of information for the distribution of socio-economic characteristics of area based on geographical space. These data are denoted as $W^{(k)} = (w_{ij}^{(k)})$ where $w_{ij}^{(k)}$= 1 if $i$ and $j$ are k-th nearest neighbours and 0 otherwise.

#### 2.2.1.2 Point data

Point data refers to discrete events that can be located with less or more precision in space and in time. They are usually assumed to follow a probability distribution and often described as a Poisson process. However, the spatial data is not independent and points far from each other are similar (DiMaggio, 2014).

According to DiMaggio (2014), point data can be analysed in different ways. Either by summary statistics, assessment of randomness or model fitting. Summary statistics resemble closest analyses. The assessment of randomness involves determining if a set of points can be accepted as a uniform Poisson process, while the model fit of point data can be used for prediction and clarification.

#### 2.2.1.3 Geostatistical data

Geostatistical data refers to the sub-branch of spatial statistics in which the data comprise of a limited example of measured esteems identifying with a fundamental spatially persistent phenomenon (Diggle & Riberio, 2006). These data are denoted as $(x_i : y_i) : i = 1, \dots., n,$ where $x_i$ identifies a spatial location (typically in a two-dimensional space, although one-dimensional and three-dimensional examples can occur) and $y_i$ is a scalar value associated with the location $x_i$. Each $y_i$ is a realization of a random variable $Y_i$ whose distribution is dependent on the value at the location $x_i$ of an underlying spatially continuous stochastic process S$(x)$ which is not directly observable.

### 2.2.2 Modelling Spatial data

Spatial data have information about the attribute of interest and its location in space. The spatial datasets can be applied to modelling and regression (Paciorek, 2013). Conditional autoregressive (CAR) models are mostly used to analyse areal-aggregated spatial data.

According to Paciorek (2013), CAR represents the spatial dependence structure such that areas that share boundaries are considered neighbours. This is alluded to by Tobler's law (1970) that nearer things are similar.

### 2.2.3 Spatial model structure for Areal and point data

Let $\mu_i = E(Y|X_i, g)$ be related via a link function, $h(.)$ to a linear predictor. The structure for areal data is defined as:

$$h(\mu_i) = X_i^T \beta + F_i g \tag{1}$$

where $F_i$ is the $ith$ row of a mapping matrix, $X_i$ is the spatial predictor for the $ith$ areal unit having regression coefficient $\beta$, while $F$ and $g(.)$ be the unknown latent spatial process represented as a piecewise constant surface on a fine rectangular grid of

$$g \sim N(0, ({}_k H)^{-}),$$

where $H$ is a CAR precision matrix and $k$ a precision parameter, recognizing the potential singularity of $H$ by using the generalized inverse notation (Paciorek, 2013, p. 949).

For point data, Paciorek (2013) showed that, $F_i$ will be a sparse vector with a single 1 that matches the location of the observations to the grid cell in which it falls.

According to De Oliveira (2012), CAR models have been widely used for the analysis of spatial data in different areas, such as demography, economy, epidemiology and geography, as models for latent and observed variables. When undertaking CAR modelling of data at an areal level, it is important to describe a nearness grid that depicts those neighbourhood structures of the information being analysed. There are quite a few methods in conformity with doing this, which include defining neighbors in according to the distances within centroids, declaring pair areas according to lie neighbors if they share a boundary (Earnest et al., 2007).

A geographic region that is divided into sub regions indexed by integers $1,2, \ldots, n$ is taken into consideration. Where by the sub regions is assumed to be endowed with a neighbourhood system, $\{N_i : i = 1, \ldots, n\}$ where $N_i$ denotes the sub regions that, in a well-defined sense, are neighbours of sub region $i$. The neighbourhood system is the main determinant for the dependence structure of the CAR model, which must satisfy that for any $i, j = 1, \ldots, n,$ where $j \in N_i$ if and only if $i \in N_j$ and $i \notin N_i$. A commonly used example in this application is the neighbourhood system defined in terms of geographic adjacency

$N_i = \{j\text{:} \, sub\ regions\ i\ and\ j\ share\ a\ boundary\},\ \ i = 1, \ldots., n.$ (De Oliveira, 2012).

## 2.3 Dependency Modelling using Copula

The idea we focused on in this study was that of copula, greatly known for some time within the statistics literature. The word copula first appeared in the statistics literature in 1959, Hoeffding has noted Nelsen (1999), with similar ideas and the results in 1940.

According to Trivedi and Zimmer (2005), interest in copulas arises from several perspectives. First, econometricians often possess more information about marginal distributions of related variables than their joint distribution. The copula approach is a useful method for deriving joint distributions given the marginal distributions, especially when the variables are not normally distributed. Second, in a bivariate context, copulas can be used to define nonparametric measures of dependence for pairs of random variables. When fairly general and/or asymmetric modes of dependence are relevant, such as those that go beyond correlation or linear association, then copulas play a special role in developing additional concepts and measures. Finally, copulas are useful extensions and generalizations of approaches for modelling joint distributions and dependence.

### 2.3.1 Defining Copula

Bohdalova and Nanasiova (2006) stated that, an $n$-dimensional copula is a multivariate Cumulative Distribution Function (CDF), $\boldsymbol{C}$, with uniformly distributed margins on [0,1] (i.e.$\boldsymbol{U}(0,1)$) and has the following properties:

1. $\boldsymbol{C}: [0, 1]^n \rightarrow [0,1]$;

2. $\boldsymbol{C}$ is grounded and $n$-increasing;

3. $\boldsymbol{C}$ has margins $C_i$ which satisfy $C_i(u) = C(1, \dots, 1, u, 1, \dots, 1) = u$ for all $u \in [0,1]$.

If $F_1, \dots, F_n$ are univariate distribution functions, then $C(F_1(x_1), \dots, F_n(x_n))$ is a multivariate CDF with margins $F_1, \dots, F_n$ since $U_i = F_i(X_i), i = 1, \dots, n$ is a uniform random variable. Copula functions are a useful tool to construct and simulate multivariate distributions (Bohdalova & Nanasiova, 2006).

### 2.3.2 The Sklar's Theorem

The Sklar's theorem (Sklar, 1959) is of importance with regard to copula and is used in all the application of copula (Baek, 2010). It is denoted as follows.

Let F be an n-dimensional distribution function with margins $F_1, \dots, F_n$. Then there exists an n-copula C such that for all x ∈ $R^n$,

$$F(x_i, \dots, x_n) = C\big(F_1(x_1), \dots, F_n(x_n)\big) \qquad (2)$$

If $F_1, \dots, F_n$ are all continuous, then C is unique. Conversely, if C is an *n*-copula and $F_1, \dots, F_n$ are distribution functions, then the function F defined above is an n-dimensional distribution function with margins $F_1, \dots, F_n$ (Nelsen, 2006).

### 2.3.3 Continuous Variable

Modelling dependence for continuous variables, through copula is defined as follows. Let $F(y_1, y_2)$ denote the joint distribution function of continuous random variables $U_1$ and $U_2$ with marginal distributions $F_1(y_1)$ and $F_2(y_2)$, respectively $y_1, y_2 \epsilon (-\infty, \infty) = \Re$. Denote by $F_i^{-1}(v_i) = \inf\{y \epsilon \Re, F_i(y) \geq v_i\}$ the corresponding inverse functions, i = 1,

2. The importance of copulas is justified by Sklar's theorem (Sklar, 1959), which states that there exists a unique copula function

$$C(v_1, v_2) = F\big(F_1^{-1}(v_1), F_2^{-1}(v_2)\big), \quad v_1, v_2 \epsilon\ (0, 1) \qquad (3)$$

that connects $F(y_1, y_2)$ to $F_1(y_1)$ and $F_2(y_2)$ via

$$F(y_1, y_2) = C\big(F_1(y_1), F_2(y_2)\big), y_1, y_2\ \epsilon\ (-\infty, \infty)$$

Hence, the information in the joint distribution $F(y_1, y_2)$ is decomposed into those of marginal distributions and that of copula function $C(v_1, v_2)$, which captures the dependence structure between $Y_1$ and $Y_2$ (Kolev & Paiva, 2009).

### 2.3.4 Discrete Variable

Applying copulas to discrete data is not as straightforward as in the case of continuous data (Genest & Neslehova, 2007). Therefore it is important to specify copulas with a simple form (Baek, 2010). For example, it is not possible to interpret the copula parameter in terms of dependence.

Let G be the distribution function of a discrete random variable X and $P\ (X = x_r) = p_r, r = 1, 2, \ldots$ with $\Sigma_{r=1}^{\infty} p_r = 1$. Then $\mathrm{P}(G(X) \leq u) \leq u, u\ \epsilon (0,1]$ and the distribution of $G(X)$ is given by

$$p(G(X)) = (p_1 + \cdots + p_k) = p_k, \quad k = 1, 2, \ldots, \qquad (4)$$

which is discrete, and therefore far away from the uniform distribution on (0, 1), (Kolev & Paiva, 2009)

## 2.4 Spatial Modelling using Copulas

In the spatial statistical framework, copula can be used to describe the joint multivariate distribution corresponding to the variables with the points of the studied area (Bardossy, 2006) and presented as a method that aims to describe all multivariate distribution of the random field. However, Bardossy (2006) method only parameterises the dependency structure of the copula. There is another method that was developed by Kazianka and Pilz (2009) where $F_n$ denotes the univariate distribution of the random process, and where η are the corresponding parameters.

Kazianka and Pilz (2009) stated that the natural assumption for a spatial copula is symmetry, where it implies that the dependency between locations $x_1$ and $x_2$ is the same as the dependence between $x_2$ and $x_1$. This means that the locations data are correlated.

## 2.5 The Gaussian Spatial Copula Model

In the class of random fields of copula, the Gaussian random field is of importance where all multivariate distribution follows a Gaussian distribution, with the mean vector of 0 and variance matrix $\sum$. The dependency structure of the Gaussian copula is parameterized by assuming that its correlation function follows one of the classical geostatistical models with parameter θ, for example, the Matern model (Kazianka & Pilz, 2009).

## 2.6 Gaussian copula framework

The vector of continuous, discrete or categorical dependent random variables can be expressed as $Y = (Y_1, \ldots, Y_n)^T$ and the corresponding realizations of $y = (y_1, \ldots y_n)^T$. Dependence can be of different forms, either in georeferenced, correlated or clustered data. In a situation where the primary scientific objective is to evaluate how the distribution of $Y_i$ varies according to changes in a vector of p covariates $x_i = (x_{i1}, \dots, x_{ip})^T$, dependence will be regarded as a secondary, but significant aspect (Masarotto & Varin, 2012).

According to Masarotto & Varin (2012, p.1519), let $p_i(y_i; \lambda) = p(y_i|x_i; \lambda)$ be the density function of $Y_i$ given $x_i$ so that covariates are allowed to affect not only the mean of $Y_i$ but also the univariate marginal distribution. With the absent of other assumptions about the nature of the dependence among the responses, inference on the marginal parameters λ can be conducted with the pseudo-likelihood constructed under working independence assumptions. Consider;

$$\mathcal{L}_{ind}(\lambda, y) = \prod_{i=1}^{n} p_i(y_i; \lambda). \qquad (5)$$

If the marginal $p_i(y_i; \lambda)$ are correctly specified, then the maximum independence likelihood estimator $\lambda_{ind} = \arg_{max\lambda} \mathcal{L}_{ind}(\lambda; y)$ consistently estimates λ with no specification of the joint distribution of Y given the model matrix X=$(x_1, \dots, x_n)^T$. Despite this important robustness property, there are various causes of concern with the above estimator. Firstly, it may suffer from considerable loss of efficiency when dependence is appreciable. Secondly, standard likelihood theory does not apply and corrected standard errors should be based on sandwich-type formulas (Masarotto & Varin, 2012, p.1519). Finally, predictions that do not take into account dependence may be of poor quality.

It is for these reasons that complementing the regression model $p_i(y_i;\ \lambda)$ with a dependence structure is important for attaining more precise inferential conclusions and predictions. The ideal model would contain all the possible joint distributions of $Y$ with univariate marginals $p_i(y_i;\ \lambda),$ for $i = 1, \dots, n$ (Masarotto & Varin, 2012).

### 2.7 Copula drawbacks and resolutions

Recently, spatial copula has mostly been used on geostatistical data (Kazianka & Pilz, 2010; Musgrove, Hughes, & Eberly, 2016) and areal data, where Hughes (2015) introduces the copula Conditional Autoregressive (copCAR) model to construct a Gaussian copula and applied the copula directly to areal outcome (Musgrove et al., 2016). Even though copCAR has many advantages over the CAR mixed model, it faces some drawbacks too.

First, it is not easy to do the copCAR analysis when the outcomes are discrete because the copCAR likelihood is computationally intractable unless the sample size is small. However, this challenge can be overcome by using a Continuous Extension (CE) approach (Madsen, 2009), Distribution Transform (DT) approach (Kazianka and Pilz, 2010) or a Composite Marginal Likelihood (CML) approach (Varin, 2008) approach. The CE approach transforms the discrete observations to continuous outcomes by convolving them with independent standard uniforms (Goren & Hughes, 2017). In the DT approach, the distributional transform stochastically "smoothes" the jumps of a discrete distribution function. On the other hands, the CML approach optimize a composite marginal likelihood formed as the product of pairwise likelihoods of adjacent observation (Goren & Hughes, 2017).

Secondly, copCAR applies the copula directly to the outcomes but the Gaussian Markov Random Field (GMRF) conditional independence does not carry over to the observations if the observations are discrete. However, by applying the copula to a continuous latent field, it can be overcome because it allows the conditional independence to carry over to the outcomes irrespective of their marginal distribution (Musgrove et al., 2016). Thirdly, even though copCAR can only analyse one indicator at a time, this can be overcome by using Generalized Joint Regression Modelling (GJRM).

## 2.8 Copula Functions

According to Trivedi and Zimmer (2005), copula functions are a useful tool for implementing efficient algorithms to simulate probability distributions in a more realistic way. Furthermore, they allow for the modelling of the dependence structure independently from the marginal distributions. Multivariate distribution with different margins can be constructed and the dependence structure given from the copula function. There are several copula functions that can be used in modelling and some are summarised in Table 1.

Table 1: A summary of selected copula functions (Trivedi & Zimmer, 2005)

| Copula type | Function $C(u_1, u_2)$ | θ-domain |
|---|---|---|
| *t*-copula | $\int_{-\infty}^{t_{\theta 1}^{-1}(u_1)} \int_{-\infty}^{t_{\theta_2}^{-1}(u_2)} \frac{1}{2\pi(1-\theta_2^2)^{\frac{1}{2}}} * \left\{1+\frac{s^2-2\theta_2 st+t^2}{v(1-\theta_2^2)}\right\}^{(\theta_1+2)/2} dsdt,$ | $\theta_1, \theta_2 \epsilon$ [-1,1] |
| Frank | $-\theta^{-1}\log\left\{1+\frac{(e^{-\theta u1})(e^{-\theta u2}-1)}{e^{\theta}-1}\right\}$ | θ ϵ (−∞, ∞). |
| Clayton | $\left(u_1^{-\theta}+u_2^{-\theta}-1\right)^{-1/\theta}$ | θ ϵ (0, ∞). |
| Gumbel | $\exp\left(-\left(\bar{u}_1^{\theta}+\bar{u}_2^{\theta}\right)^{\frac{1}{\theta}}\right)$ | θ ϵ [1, ∞) |
| Gaussian | $\Phi_G(\Phi^{-1}(\mu_1), \Phi^{-1}(\mu_2);\ \theta)$ | θ ϵ [-1, 1] |

The [-1, 1] of *t*-copula permit a typical correlation capturing both negative and positive correlation with range of -1 and 1. The clayton and Gumble copulas capture only positive correlation and permit tail dependences, while the Frank copula allow both positive and negative dependence. As for the Gaussian, when the dependence parameter approaches −1 and 1, it attains the Frechet lower and upper bound.

### 2.9 Dependence Concepts

Copulas provide a way to study and measure dependence between random variables. The joint distribution function describes the dependency between random variable. Dependence is a matter of association between X and Y along any measurable function, i.e. the more X and Y tend to cluster around the graph of a function, either $y = f(x)\ /x = g(y)$ , the more they are dependent (Kort, 2007).

### 2.9.1 Tail Dependence

According to Joe (2015), tail dependence measures the strength of dependence in lower and upper joint tail of a multivariate distribution.

Let X ~ F and Y ~ G be two random variables with copula C, define the coefficients of tail dependency

$$\lambda_L = \lim_{u \downarrow 0} \mathbb{P}[F(X) < u | G(Y) < u] = \lim_{u \downarrow 0} \frac{C(u,u)}{u} \tag{6}$$

$$\lambda_U = \lim_{u \uparrow 1} \mathbb{P}[F(X) > u | G(Y) > u] = \lim_{u \uparrow 1} \frac{1 - 2u + (u,u)}{1 - u}$$

where $\lambda_L$ is the lower tail and $\lambda_U$ is the upper tail

C is said to have lower (upper) tail dependence *iff*. $\lambda_L \neq 0$ ($\lambda_U \neq 0$) .

The interpretation of the coefficients of tail dependency is that it measures the probability of two random variables both taking extreme values (Kort, 2007).

### 2.9.2 Kendall's tau and Spearman's rho

According to Embrechts et al. (2001) Kendall's tau and Spearman's rho are important measures of dependence (concordance). They are good alternatives to linear correlation coefficient as a measure of dependence because the linear can be inappropriate and misleading.

Kendall's tau for a random vector $(X, Y)^T$ is defined as

$$_T(X,Y) = \mathbb{P}\{(X - \tilde{X})(Y - \tilde{Y}) > 0\} - \mathbb{P}\{(X - \tilde{X})(Y - \tilde{Y}) < 0\} \tag{7}$$

where ( $\tilde{X}, \tilde{Y}$) is an independent copy of $(X, Y)^T$.

Hence the Kendall's tau for $(X, Y)^T$ is the probability of concordance minus the probability of discordance.

Let $(X, Y)^T$ be a vector of continuous random variables with copula C.

Then Kendall's tau for $(X, Y)^T$ is given by

$$_T(X, Y) = Q(C, C) = 4 \iint_{[0,1)^2} C(u, v)\, dC(u, v) - 1$$

The integral above is the expected value of the random variable $C(U, V)$ where $U, V \sim U(0, 1)$ with joint distribution function C i.e $_T(X, Y) = 4\mathbb{E}\big(C(U, V)\big) - 1$

The Spearman's rho for the random vector $(X, Y)^T$ is given by

$$_{ps}(X, Y) = 3\big(\mathbb{P}\{(X - \tilde{X})(Y - Y') > 0\} - \mathbb{P}\{(X - \tilde{X})(Y - Y') < 0\}\big) \qquad (8)$$

where $\tilde{X}$ and $Y'$ are independent.

Let $(X, Y\ )^T$ be a vector of continuous random variable with copula C. The Spearman's rho for $(X, Y\ )^T$ is given by

$$_{ps}(X, Y) = 3Q(C, \Pi) = 12\{\iint_{[0,1]^2} uv dC(u, v) - 3 = 12 \iint_{[0,1]} C(u, v)\, dudv - 3$$

Hence, if $X \sim F$ and $Y \sim G$, and we let $U = F(X)$ and $V = G(Y)$, then

$$_{ps}(X, Y) = 12 \iint_{[0,1)^2} uv dC(u, v) - 3 = 12\mathbb{E}(UV) - 3$$

$$= \frac{\mathbb{E}(UV) - 1/4}{1/12} = \frac{Cov(U, V)}{\sqrt{Var(U)}\ \sqrt{Var(V)}}$$

$$= P\big(F(X), G(Y)\big)$$

## 2.10 Estimation

Maximum Likelihood Estimation (MLE) can be used to estimate copula parameters (Kazianka & Pilz, 2010). The correlation structure θ, copula parameter $\lambda$ and the parameter of the marginal distribution $\alpha$ are the parameters that can be estimated in the spatial copula model.

The likelihood function is given by;

$$L(\Theta, Z(x)) = c_{\theta,\lambda}\left(F_\alpha(Z_1), \dots, F_\alpha(Z_N)\right) * \prod_{i=1}^{n} f_\alpha(Z_i) \quad (9)$$

where: $\Theta = (\theta, \lambda, \alpha)$

$c_{\theta,\lambda}$ is the copula density

$f_\alpha$ is the marginal density

$F_\alpha$ is its distribution function.

The maximum likelihood is easily applied to Gaussian and Student *t*. In the case where the copula is no-central, it may be difficult for the density for higher dimension to be calculated. However, Kazianka and Pilz (2010) propose to use the bivariate copula densities to perform the maximum likelihood based on the assumption of independence of different pairs of observations.

There are other maximum likelihood estimation that also exists, such as, the canonical or Pseudo Maximum Likelihood Estimation (PMLE). This estimation can also be used for copula's parameters. The pseudo-observations calculation ensures the removal of a marginal characteristic of the variables and only information on the dependence structure remains (Jane, Valle, Simmonds, & Raby, 2016). The inference function for margins is used in a situation when MLE is complicated to use, and two procedures are followed. "First, the parameters of each marginal distribution are estimated separately via MLE. Second, it is subsequently substituted into the likelihood function which is then maximised to obtain estimates of the dependence parameters" (Jane et al., 2016, p. 6).

### 2.11 Posterior predictive distribution

Modelling spatial data is not only used to recognize critical covariates, but also for producing maps of the result by foreseeing it at unsampled areas (Ayele, Zewotir, & Mwambi, 2013). This forecast is known as kriging. Kriging is an interpolation based on a regression against observed values of surrounding data points. This spatial prediction can be illustrated as follows.

Let $Y_0$ be a vector of the binary response at a new unobserved location, $m_{xi}$, for $i = 1, \dots n_0$. Following the maximum likelihood approach, the distribute of $Y_0$ is given by

$$P(Y_0|\beta, U, \sigma^2, \phi) = \int \ (P(Y_0|\beta, U_0)P(U_0|U_1, \sigma^2, \phi)dU_0 \qquad (10)$$

where $\beta$, $\sigma^2$and $\phi$ are the maximum likelihood estimates of the relating parameters, $P(Y_0|\beta, U_0)$ is the Bernoulii-likelihood at the new area and $P(U_0|U_1, \phi)$ is the distribution of the spatial random effect $U_0$ at the new area, given $U$ at the observed sites and it is assumed to follow a normal distribution

$$P(Y_0|U, \sigma^2, \phi) = N(\textstyle\sum_{01} \sum_{11}^{1} U, \sum_{00} - \sum_{01} \sum_{11}^{1} \sum_{10}) \qquad (11)$$

The mean of the Gaussian distribution in (11) is the kriging estimator areas (Ayele et al., 2013).

### 2.12 Model Selection Criterion

Radice, Marra, and Wojtyś (2016) stated that copula models with a single dependence parameter can be thought of as non-nested models except for Gaussian and Student-t distribution when $\nu \rightarrow \infty$. One approach of choosing between copula models is either by using the Akaike Information Criterion (AIC) or Bayesian information criterion (BIC). It is important to carry out model selection to ensure goodness of fit and adjust or penalize model complexity. There are different selection criterion. Some are discussed as follows.

#### 2.12.1 Akaike Information Criterion (AIC)

The AIC (1973, 1974) was developed by Akaike as estimators of the expected Kullback-Lieber discrepancy between the model generating the data and a fitted candidate model (Cui & George, 2008).

It is given by:

$$AIC = -2\ln(l) + 2k \qquad (12)$$

where $k$ is the number of predictors and $l$ is the maximized likelihood value, $2k$ includes extra predictors and -2 rewards the fit between the model and the data. The model with the minimum AIC value is the best model to use.

### 2.12.2 Bayesian Information Criterion (BIC)

The BIC model was introduced by Schwarz (1978) and is based on the empirical log-likelihood. It does not require the specifications of prior probabilities, therefore, favoured in situations where the priors are difficult to set (Cavanaugh & Neath, 1999). It is similar to the AIC and both statistics penalize model complexity.

The BIC with a $k$ parameter model is of the form:

$$BIC = -2\ln(l) + k\ln(n) \tag{13}$$

where,

$l$= the maximum value of the likelihood function of the estimated model,

$k$= the number of free parameters to be estimated and

$n$= the number of observations or the sample size.

In model selection application, the optimal fitted model is identified by the minimum value of BIC.

### 2.12.3 Deviance Information Criterion (DIC)

Zhu and Carlin (2000) defines DIC as a generalization of the AIC and the BIC and is widely used in model selection where MCMC simulation is used.

The deviance is denoted as:

$$D(\theta) = -2\log\big(P(y|\theta)\big) + 2\log\big(f(y)\big) \tag{14}$$

where $y$ are the data, $\theta$ are the unknown parameters of the model, $p(y|\theta)$ is likelihood function and $f(y) = p(y|u(\theta) = y)$. The effective number of parameters of the model is computed as

$$P(D) = \overline{D}(\theta) - D(\bar{\theta})$$

where $\overline{\theta}$ is the expectation of θ, $P(D)$ is the effective number of parameters and $\overline{D}$ is the posterior mean of the deviance. Thus DIC is then calculated as

$$DIC = P(D) - \overline{D}\ (\theta)$$

In model selection, the models with smaller DIC value are preferred over models with a larger DIC.

# CHAPTER 3: RESEARCH METHODOLOGY

## 3.1 Introduction

The study cogitates using some of the methods reviewed in Chapter 2 to model WASH indicators in Namibia. The spatial modelling using copula was done at regional level using geostatistical and areal data.

## 3.2 Research Design

The study used secondary data acquired through a cross-sectional design done on the 2013 NDHS data by MoHSS. A stratified sampling method was employed at a two-stage cluster sampling. The first stage was the Primary Sampling Units (PSUs) and the second stage was at the households that were selected in every urban and rural areas. In addition, the sample was intended to give evaluations of the most key factors for the 13 administrative regions in Namibia (MoHSS & Macro International Inc, 2014).

## 3.3 Sample

The study made use of the household (HH) data that were the samples in the NDHS. A total of 11 004 households were selected for the sample, of which 10 165 were found to be occupied during data collection. Of the occupied households, 9 849 were successfully interviewed, giving a response rate of 92%.

## 3.4 Research instruments

Three questionnaires were administered in the 2013 NDHS: Household Questionnaire, Woman's Questionnaire, and Man's Questionnaire. These questionnaires were adjusted from the standard Demographic and Health Survey (DHS) centre questionnaires to reflect the population and medical problems pertinent to Namibia at a progression of gatherings with different partners from government services and offices, nongovernmental associations, and universal givers. However, this study mostly made use of the HH questionnaire, but covariates such as literacy, exposure to the media, respondent occupation and working status were obtained from the woman's questionnaire.

### 3.5 Data collection

The study used secondary data of the 2013 NDHS from the Ministry of Health and Social Services (MoHSS). 28 groups, each comprising of a team supervisor, a field editor, three female interviewers, one male interviewer, and health technician, collected the 2013 NHDS data. National and regional supervisors through close supervision and observing amid hands-on work kept up quality assurance. The questionnaires were altered by the field editors in the field and confirmed by the team supervisor before being transported to the MoHSS head office. In addition, national and regional supervisors ensured guaranteed quality control through altering of questionnaires and perception of interviewers. Common mistakes and practical solutions were conveyed through composed notes and talked about with all team members (MoHSS & Macro International Inc, 2014).

NDHS respondents were asked what their source of drinking water was, toilet facilities, and what they used to wash their hands. The study adopted the UNICEF definition for WASH. The response variables, $y_i$, in this study for source of drinking water took the value $y_i$= 1 if the HH has improved drinking water source, and $y_i$= 0 if the HH has non-improved drinking water source. For sanitation, $y_i$=1 if the HH uses an improved shared and not shared toilet facility and $y_i$=0 if they use a non-improved toilet facility, while for hygiene, $y_i$= 1 if the HH use water and soap to wash their hands and $y_i$= 0 if they only use water or do not wash their hands.

According to UNICEF (2016), an improved drinking water source refers to the way it is constructed and when used properly, it adequately protects the source from outside contamination. These sources are piped into a dwelling, piped to yard or plot, public tap/standpipe, tube well or borehole, protected well, protected spring, rainwater and bottle water. A non-improved water source is constructed in such a way that water is not protected from outside contamination, such as an unprotected well, an unprotected spring and tanker or cart with a drum.

Improved sanitation mean facilities that hygienically separate human excretion from human contact (UNICEF, 2016). These facilities are flush/pour flush to piped sewer system, flush/pour flush to a septic tank, flush/pour flush to pit latrine, Ventilated Improved Pit (VIP) latrine, pit latrine with slab, composting toilet. The non-improved sanitation is the flush/pour flush not to sewer/septic tank latrine, pit latrine without slab/open pit, bucket, hanging toilet/hanging latrine and no facility/bush/field. Meanwhile, improved hygiene is the availability of water and soap for handwashing.

### 3.6 Data Management

Data management in relation to mining and generation of variables from the household questionnaire file and woman's data set were done in IBM SPSS (version 20). Outcome water was generated from the variable source of drinking water, sanitation was generated from the type of toilet facility while hygiene was generated from a place where household members wash their hands, the presence of water at hand washing place and presence of soap or a detergent.

Several explanatory variables were recoded to fewer categories, such as the wealth index and occupation. The attribute of data cleaning was to reduce the number of variables in the dataset, in order to retain variables of interest that are vital for the analysis of the study.

In order to carry out model fitting for the Gaussian Copula Marginal Regression (GCMR), fewer cases were needed for smooth analysis. The variables were aggregated to 550 cases with the variable HV001 (cluster number) in the DHS. The Global Positioning System (GPS) coordinates that were obtained during the survey data collection were merged with the 550 cases. For the outcome variables, case numbers were obtained and for the covariates, they were transformed to percentages except, for the age and household size where the mean values were obtained. The data was then exported to **R** statistical software (R Core Team, 2016) for advanced analysis and to ArcGIS to create smooth maps.

### 3.7 Description of Key Variables

Variables for this study were; improved drinking water, sanitation and hygiene measured at household level. The covariates used were; sex of the head of household, household size, age of the head of household, type of place of residence, education level, marital status, exposure to media (frequency of reading newspaper or magazine, frequency of listening to radio and frequency of watching television), wealth index, literacy, respondent occupation and the respondent working status were considered.

Table 2 shows the description of the variables that were used in the study. The choice of the covariates was based on studies by Tiwari and Nayak (2013), Yang et al.(2013), Koskei, Koskei, Koske, and Koech (2013) and Adam, Boateng, and Amoyaw (2015) as they were found to predict access to improved drinking water, sanitation and hygiene. In addition, they were also chosen based on their statistical significance (Irianti, Prasetyoputra, & Sasimartoyo, 2016).

Table 2: Description of key variables

| Variable | Description |
|---|---|
| **Response variables** | |
| (i) Water | Source of drinking water<br>1=improved drinking water source,<br>0=non-improved drinking water source |
| (ii) Sanitation | Sanitation facilities<br>1= improved shared and not shared toilet facility<br>0= non-improved toilet facility |
| (iii) Hygiene | Hand washing<br>1= Place where household members wash their hands and presence of water and soap<br>0= No place for washing hands, no water or no soap |
| **Covariates** | |
| (i) Sex of household head | 1 = Male<br>2= Female |
| (ii) Household size | Number of people living in the house, fitted as continuous |
| (iii) Age of household head | The age of the head of the household, fitted as continuous |
| (iv) Type of place of residence | 1= Urban, 2 = Rural |

| | | |
|---|---|---|
| (v) | Household Wealth index | 1= Poor, 2= Middle, 3= Rich |
| (vi) | Women literacy | 0= Cannot read at all<br>1= Able to read only part of sentence<br>2= Able to read whole language<br>3= No card with required language<br>4= Blind/visually impaired |
| (vii) | Frequency of reading newspaper or magazine | 0= Not at all<br>1= Less than once a week<br>2-At least once a week |
| (viii) | Frequency of listening to radio | 0= Not at all<br>1= Less than once a week<br>2-At least once a week |
| (ix) | Frequency of watching television | 0= Not at all<br>1= Less than once a week<br>2-At least once a week |
| (x) | Highest educational level attained by household head | 0= No education<br>1= Primary<br>2= Secondary<br>3= Higher |
| (xi) | Current Marital Status of household head | 0= Never married<br>1= Married<br>2= Living together<br>3= Widowed<br>4= Divorced |
| (xii) | Respondent Occupation | 0= Not working/unemployed/student/retired<br>1= Unskilled<br>2= Skilled<br>3= Low Skilled |

| | |
|---|---|
| | 4= High Skilled |
| (xiii) Respondent working status | 0= No<br>1= Yes |
| **Spatial Effect** | |
| (i) Region | Administrative boundaries. Data were collected from 13 regions during the 2013 NDHS. |
| (ii) Enumeration Area (EA) | 550 point coordinates of each sampled cluster used for geostatistical modelling |

**3.8 Data Analysis**

To perform a flexible spatial modelling at the regional level, copula was applied to account for spatial dependency and spatial distribution for improved drinking water, sanitation and hygiene indicators.

Two copula approaches were utilized. First, different models were fitted on the geostatistical data using the Gaussian Copula Marginal Regression (GCMR). Here, the spatial model assumed geostatistical framework and the best fitted model was used to create smooth maps that estimated spatial distribution of the WASH indicators. Second, a joint analysis of WASH indicators was carried out to account for dependence and three models were fitted for complexity. Spatial copula models were used to describe data at different geographic levels.

For the GCMR, data were aggregated to cluster numbers assigned to regions for the purpose of carrying out the spatial analysis. For regions, it is known that spatial correlation exist because nearer things have similar values. The covariates used in the study were modelled as fixed effects and the spatial effects were utilized to represent spatial variation. A flexible copula based approach was adopted to estimate the joint discrete. Modelling the WASH indicators included non-linear effects, spatial effects and fixed effects. The joint regression model used in this study will contributed to the methodology of joining discrete variables.

### 3.9 Research Ethics

Approval to use the secondary data for the 2013 NDHS was granted by the MoHSS, see Appendix B. The data were stored carefully to ensure that no other person uses them for their own interest and were only used and analyzed to answer the objectives of this study. The University Research Ethics Committee (UREC) granted ethical clearance and the necessary research permission letter, see Appendix A.

# CHAPTER 4: SPATIAL DISTRIBUTION OF WASH INDICATORS USING COPULA ON GEOSTATISTICAL DATA IN NAMIBIA

## 4.1 Introduction

Chapter 4 gives a brief review of WASH and presents outcomes of spatial distribution for WASH indicators models fitted using GCMR and quantifies the determinants of the indicators. Predicted and prediction standard errors of WASH indicators were mapped to create coverage surface.

Access to improved drinking water, sanitation and hygiene (WASH) is a basic human right, which is essential to population's health, welfare and development (WHO & UNICEF, 2015). Hence, lack of access to WASH is a great concern globally. According to WHO & UNICEF (2015), globally, about 683 million people have no access to improved drinking water and 2.4 billion people have no adequate improved sanitation.

The United Nations (UN) has noted that insufficient WASH condition causes about 1.7 million death annually. Furthermore, poor WASH is the main cause of diarrhoeal diseases, which is the main leading cause of morbidity and mortality among children under the age of five years (Ezeh, Agho, Dibley, Hall, & Page, 2014) and the leading cause of death in Sub-Sahara Africa (Pruss-Ustun et al., 2014). Moreover, adequate WASH in health facilities is critical for maternal and newborn health. Communicable diseases can easily be prevented and controlled if there are safe water and sanitation in health facilities (UNICEF, 2015). According to Jasper, Le and Bartarm (2012), girls and women are mostly affected by poor WASH as they spend most of their time collecting water and thus leading to them not fully participating in school. In addition, their menstrual hygiene management can negatively be affected by lack of access to WASH. Inadequate sanitation, water and hygiene also prevent children to attend school and in the process affect their performances. Furthermore, children with disabilities are denied access to education as it gets difficult for them to access WASH when it is unavailable or inadequate (UNICEF, 2016).

Even though progress has been made over the Millennium Development Goals (MDGs) to increase access to WASH, much still need to be done. Besides that, there are also challenges in scaling up WASH. First, there has been an increase in the urban population. Urbanization poses a major challenge in providing safe drinking water and proper sanitation (Cohen, 2006). Water usage has risen dramatically in past years due to population growth and agricultural purpose, hence causing water scarcity and poor sanitation especially in the informal settlements. Second, there is an inadequate investment in the water and sanitation infrastructure (Moe & Rheingans, 2006). Even though there has been advocacy in water issues, investment in water and sanitation has been declining. Since 2012, investment in water and sanitation has declined from US$ 10.4 billion to US$ 8.2 million in 2015 (WHO, 2017). Even in areas where there has been investment, there is still lack of ownership from the communities of water and sanitation infrastructure; hence, they are vandalized in the process. Third, the destructive impact of climate change continues to be a threat to water and sanitation (WHO & UNICEF, 2015). Lastly, there is no monitoring and evaluation conducted on projects implemented for water and sanitation to determine whether they are successful and sustainable.

Mapping of WASH indicators is widely being used around the world for visualization of distribution in order to clearly identify areas that are served or not served (Garriga & Forguet, 2013). This study used a Gaussian copula approach to analyze and map WASH indicators. This is because copula separates the choice of dependence among variables from the choice of marginal distribution of each variable (Kuethe et al., 2009). In addition, copula can consolidate marginal distribution, which are hard to join in standard joint distribution, such as for, the general utilized multivariate normal or Poisson distribution.

## 4.2 Statistical Modelling

Despite diverse approaches such as automodel, CAR and GLMM that can be used to model WASH indicators; this study employed a GCMR model because it gives a scientific advantageous structure to deal with different types of dependence in regression models (Masarotto & Varin, 2012).

Point-referenced data were mapped with the interest to know whether the data display spatial correlation. The need to use these data is because the points near each other in space have improved drinking water, sanitation and hygiene that is similar with points that are far apart (Ayele et al., 2013). Liverani (2016) pointed out that this is because spatially correlated data cannot be considered as independent observation. If the analysis does not consider correlation structure of the data, the estimates obtained from the analysis may be inaccurate because of the underestimated standard errors. Point-referenced data were preferred in this case because of the interest of prediction.

### 4.2.1 Gaussian Copula Marginal Regression Models

According to Masarotto and Varin (2012), the GCMR model is of the form of

$$Y_i = g(x_i, \epsilon_i; \lambda), \quad i = 1, \ldots, n, \tag{15}$$

where $g(.)$ the corresponding is a function of $x_i$ and $\lambda$ is the parameter and $\epsilon_i$ is an unobserved stochastic variable, known as the *error* term. Among the possible $g(.)$ selection of the model, the following is suitable

$$Y_i = F_i^{-1}\{\Phi(\epsilon_i); \lambda\}, \quad i = 1, \ldots, n, \tag{16}$$

where $\Phi(.)$ is the cumulative distribution of $Y_i$ given $x_i$ (Masarotto & Varin, 2012).

Masarotto and Varin (2012) pointed out that, Gaussian copula provides a flexible general framework for modelling dependent response of any kind, such as spatial, time-series, cross-design, survival, longitudinal and clustered data. It allows flexible specification of the regression model and this can also be applied to logistic regression for spatially correlated binomial data that are point-referenced (Masarotto & Varin, 2017). The interpretation in marginal modelling is combined with the flexibility in the specification of the dependency structure by Gaussian copula.

Spatially correlated observations can be modelled by assuming that the errors $\epsilon$ are produced by a Gaussian random field with an appropriate spatial correlation work. A common class of spatial correlation functions is the Matérn of the form:

$$cor(\epsilon_i, \epsilon_j) = \frac{1}{2^{\alpha-1}\Gamma(\alpha)}\left(\frac{\|s_i - s_j\|_2}{\tau}\right)^2 K_b\left(\frac{\|s_i - s_j\|_2}{\tau}\right) \tag{17}$$

where $s_i$ is the coordinate of the $ith$ observation, $s_j$ is the data, $\|.\|_2$ is the Euclidean norm, and $K_b(.)$ is the modified Bessel function of order $\alpha$ (Masarotto & Varin, 2012). Masarotto and Varin (2012) pointed out that GCMR models differ from other marginal models when it comes to the form of bivariate and higher order dimensional joint distribution. For continuous cases, the mapping (eqn 16) between $\epsilon_i$ and $Y_i$ is one-to-one to obtain the marginal distribution of the response through standard transformation rules from the distribution of the corresponding errors.

Moreover, where there are linear regression models with normally distributed errors, the correlation between pairs of responses $Y_i$ and $Y_j$, with corresponding covariates $x_i$ and $x_j$, concurs with the correlation of the corresponding normal errors cor $(\epsilon_i, \epsilon_j)$. Otherwise the correlation between $Y_i$ and $Y_j$ is a nonlinear function of the correlation of $\epsilon_i$ and $\epsilon_j$ (Masarotto & Varin, 2012). Furthermore, Kendall $\tau$ and Spearman $\rho$ are measures of association that are regularly used in copula.

Maximum likelihood was used for fitting of the GCMR models. The use of likelihood ratio statistics and information criteria in relation to standard tools of hypothesis testing and model selection benefit the maximum likelihood. It is then advisable to treat the continuous and discrete cases individually. The continuous case, in terms of one-to-one relationship between response $Y$ and the errors $\epsilon$, yields a likelihood for $\theta$ as

$$\mathcal{L}(\theta; y) = \mathcal{L}_{ind}(\lambda; y)q(\epsilon; \theta), \tag{18}$$

where $\mathcal{L}_{ind}(\lambda; y)$ is the independence likelihood and $q(\epsilon; \theta)$ is the measure of evidence for dependence among the errors.

For discrete cases, the likelihood requires a computation of the $n$-dimension integral with $\vartheta_1$ and $\vartheta_n$ being the parameters to be estimated.

$$\mathcal{L}(\theta; y) = \int_{\vartheta_1(y_1;\lambda)} \dots \int_{\vartheta_n(y_n;\lambda)} p(\epsilon_1, \dots, \epsilon_n; \theta) d\epsilon_i \dots d\epsilon_n. \qquad (19)$$

**4.2.1.1 gcmr package in R**

The gcmr (version 1.01) package in **R** was developed by Masarotto and Varin in 2017, and currently depends on version 3.0.0 of **R**. The package is used to fit Gaussian copula marginal regression models. It can be found in a statistical software **R** (R Core Team, 2016). The main function of the gcmr is gcmr (), which takes an argument of:

*gcmr(formula, data, subset, offset, marginal, cormat, start, fixed, options = gcmr.options(...), model = TRUE, ...)*

The arguments in the function are used for model-frame specification (Masarotto & Varin, 2017). The formula is used to restrict analysis to a subset of the data, to set an offset, or to fix contrasts for factors. The *marginal* and *cormat* specify the marginal part of the model and the copula correlation matrix. The *start* is used for supplying the starting values, while the *fixed* is used for fixing the values of some parameters in the model building and the *options* used to set the fitting options.

The package implements maximum likelihood inference for Gaussian copula regression models (Masarotto & Varin, 2017). In addition, it allow a flexible description of the regression model and the dependence structure. Moreover, models are fitted with the method of maximum likelihood in the continuous case and maximum simulated likelihood in the discrete case. Different marginal distribution are found in the gcmr, such as binomial, gamma, Gaussian, Poisson and negative binomial.

### 4.3 GCMR model fitting

The gcmr package in R developed by Masarotto and Varin in 2017 was used to estimate the spatial distribution of separate WASH indicators using copula on geostatistical data. Cross tabulations were done between the response variables (water, sanitation and hygiene) and the covariates (sex of the head of household, household size, age, type of place of residence, education level, marital status, exposure to media and wealth index) that had 9849 cases with 550 EA in order to be aggregated and allow smooth analysis. The mean value was calculated for the size of the household and the age of the head of household were calculated. The spatial effects of the geographical area were characterized by different regions to account for the spatial variation of WASH indicators in different areas.

A matrix of the distance between the regions was constructed. The distances were expressed in kilometers through a scaling factor of 1,000. Scaling is helpful for avoiding potential numerical instabilities in the estimation of the spatial dependence parameter (Masarotto & Varin, 2017). Covariates were also scaled by a factor of 100, except for age, household size, residence and region. Spatial dependence was modelled with an exponential correlation matrix corresponding to the default value of the shape parameter (alpha = 0.5) in matern.cormat (D, alpha).

Fitting the GCMR models was done through the method of maximum likelihood estimation, using the Gaussian copula. Two separate models were fitted per WASH indicator, yielding six models in total. The first model of the indicators was fitted with fixed effects to find out how they influenced the indicators, while the second model was fitted with spatial effects, for spatial variation of WASH indicators. The AIC value per model was used to select the best fit model. The standard diagnostic plots were further used to see if the best fit model met the model conditions. The best fit model was later used to predict improved drinking water, sanitation and hygiene, including the prediction standard errors.

The equation for the Gaussian copula marginal regression model fitted for Tables 5, 7 and 9 was as follow:

$$E(Y_i|X_i) = log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon_i \tag{20}$$

where $p_i$ is the probability value , $\beta_0, \beta_1, \ldots, \beta_p$ are the coefficient , $X_1$,…., $X_p$ are the predictor variables and $\varepsilon_i$ is the random error term.

### 4.4 Distribution of percentages of households with improved drinking water, sanitation and hygiene

Point maps of households with improved drinking water, sanitation and hygiene were generated at regional level to visualize the distribution of improved water, sanitation and hygiene. The point maps of the WASH indicators provide a basis for comparison purposes of the GCMR model fitted value estimates.

The distribution of percentages of households with improved drinking water, sanitation and hygiene in Namibia at regional level is shown in Figure 1.

(a) (b) (c)

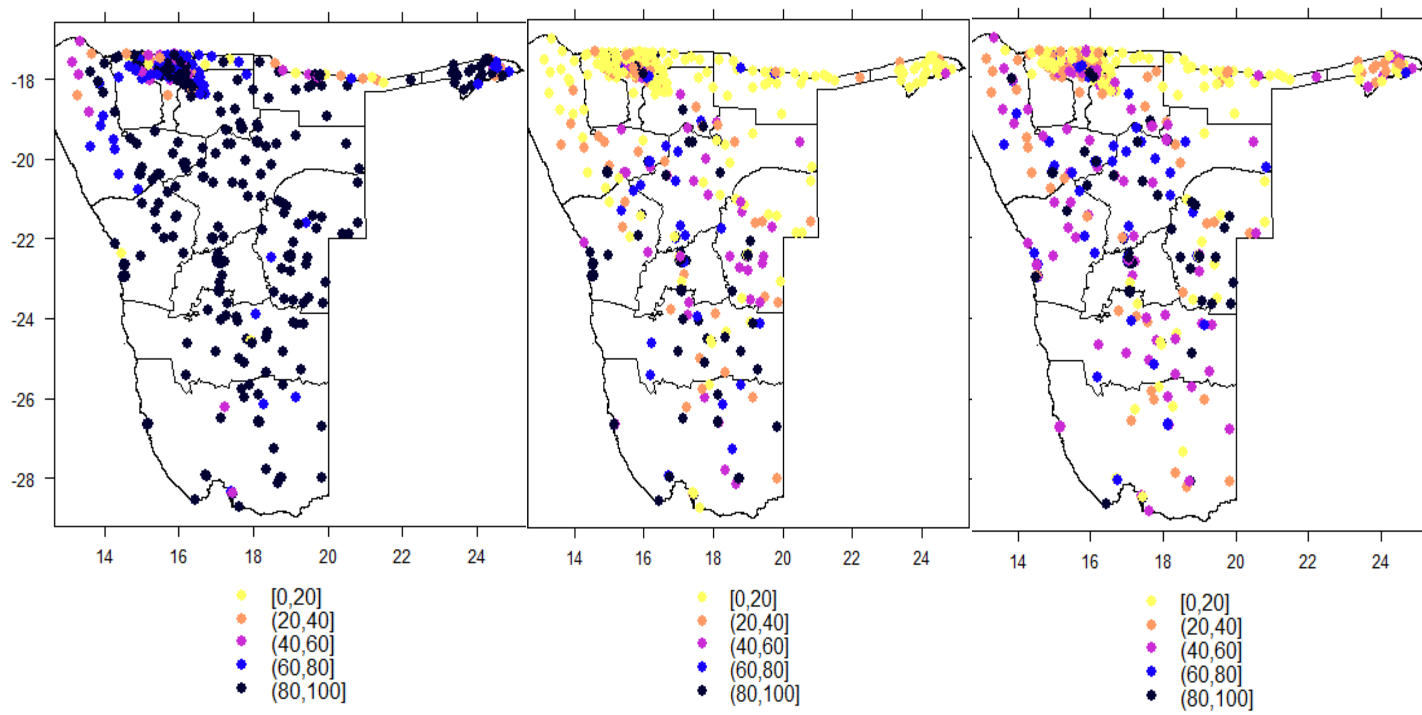



Figure 1: Point maps showing percentages of households with improved drinking water, sanitation and hygiene in Namibia. (See Appendix C for the names of the regions).

Figure 1(a) shows that many regions have households with improved drinking water, except for Kavango, Kunene and Ohangwena regions. For improved sanitation, Figure 1(b) indicates that the households with a moderate percentages of improved sanitation (60-80%) were in the Khomas, !Karas, Hardap, Erongo and Otjozondjupa regions.

Households with low percentages of improved sanitation were found in Kunene, Zambezi, Kavango, Ohangwena, Omusati and Oshana regions. The distribution of improved hygiene in Figure 1 (c) shows that households with low improved hygiene (0-20%) were found in !Karas, Hardap, Kavango, Zambezi, Ohangwena, Omusati, Oshikoto and Oshana regions. The point maps in Figure 1 presents results based on regional distribution. However, it is helpful to provide mean percentage per region as it gives an overall indication of household having access to improved drinking water, sanitation and hygiene. Table 3 presents results of the mean percentage, minimum percentage and maximum percentage of improved drinking water, sanitation and hygiene per region.

Table 3: Summaries of percentage coverage of improved drinking water, sanitation and hygiene

| Region | Improved drinking water | | | Improved sanitation | | | Improved hygiene | | |
|---|---|---|---|---|---|---|---|---|---|
| | Mean % | Min % | Max % | Mean % | Min % | Max % | Mean % | Min % | Max % |
| Zambezi | 87.71 | 25 | 100 | 20.69 | 0 | 100 | 36.53 | 5 | 85.71 |
| Erongo | 97.98 | 10 | 100 | 85.25 | 0 | 100 | 60.66 | 11.11 | 100 |
| Hardap | 92.74 | 12.50 | 100 | 64.69 | 0 | 100 | 54.59 | 0 | 95 |
| !Karas | 92.88 | 45 | 100 | 66.54 | 0 | 100 | 49.19 | 0 | 100 |
| Kavango | 70.79 | 0 | 100 | 22.17 | 0 | 100 | 23.23 | 0 | 85 |
| Khomas | 99.08 | 86.67 | 100 | 80.05 | 0 | 100 | 66.34 | 6.67 | 100 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Kunene | 84.55 | 23.08 | 100 | 33.39 | 0 | 100 | 57.49 | 17.65 | 100 |
| Ohangwena | 65.37 | 5.26 | 100 | 15.15 | 0 | 100 | 20.43 | 0 | 82.35 |
| Omaheke | 96.35 | 68.75 | 100 | 39.12 | 0 | 100 | 54.87 | 0 | 100 |
| Omusati | 68.69 | 10 | 100 | 19.14 | 0 | 100 | 25.8 | 0 | 85.71 |
| Oshana | 88.21 | 27.78 | 100 | 49.29 | 0 | 100 | 56.81 | 23.55 | 84.21 |
| Oshikoto | 83.91 | 35 | 100 | 30.09 | 0 | 100 | 41.12 | 0 | 100 |
| Otjozondjupa | 98.09 | 83.33 | 100 | 60.92 | 0 | 100 | 70.33 | 12.50 | 100 |

Min = Minimum, Max = Maximum

Table 3 shows that most households in the fourteen (14) regions have access to improved drinking water, with mean percentage value above 65%. Households with the lowest minimum percentage were found in Kavango region with 0%, while all regions were reported to have a maximum percentage of 100%. It is worth noting that the minimum percentage of households with improved drinking water was quite high in Khomas and Otjozondjupa regions with 86.67% and 83.33% respectively. With regard to improved sanitation, Ohangwena and Omusati regions had the lowest mean percentage of 15.15% and 19.14% respectively. In addition, the results shows that the minimum percentage of improved sanitation in all region was 0%, while the maximum was 100%.

Furthermore, households with the lowest mean percentage of improved hygiene were found in Ohangwena region with 20.43%, while the region with the highest mean percentage of improved hygiene was Otjozondjupa with 70.33%. Hardap, !Karas, Kavango, Ohangwena, Omaheke, Omusati and Oshikoto regions had a minimum percentage of households with improved hygiene of 0%. The results showed that Erongo,

!Karas, Khomas, Kunene, Omaheke, Oshikoto and Otjozondjupa regions had households with maximum percentage of improved hygiene of 100%.

### 4.5 GCMR model results

Under this section, we present results of different models fitted using the gcmr package in R for the WASH indicators using the 2013 NDHS data. Two models were fitted per indicator, one with fixed effects and another model with an additional covariate region. However, only the results of the best model per indicator is presented and used to estimate spatial distribution.

#### 4.5.1 Improved drinking water

Under this section, we present results of models fitted for improved drinking water. The results are presented in Tables 4 and 5 respectively.

##### 4.5.1.1 Model selection

Table 4 presents the AIC results obtained from the two models fitted for improved drinking water. The results show that the model fitted with fixed effects is the best model, with an AIC value of 3152.4. Looking at the difference in AIC of the two models, it can be concluded that the model fitted with an additional covariate region can strongly be differentiated from the model with fixed effects, with an AIC difference value of 1873.9.

Table 4: AIC results from improved drinking water models fitted using gcmr package

| Model | Likelihood | AIC |
|---|---|---|
| Improved drinking water with fixed effects | 1548.2 | 3152.4 |
| Improved drinking water with an additional covariate region | 2473.1 | 5026.2 |

### 4.5.1.2 Model with fixed effects results

From the results of the best model fitted in Table 4, the summaries of fixed effects that can influence improved drinking water are presented in Table 5. The standard errors obtained are for the estimates. The Odds Ratio (OR) were used to measure association between the outcome and the exposure.

Table 5: Results from the Gaussian copula marginal regression model for improved drinking water fitted using gcmr package.

| Coefficients | Estimates | Standard Errors | Odds Ratio (OR) | P-values |
|---|---|---|---|---|
| Sex (ref. Male) | | | | |
| Female | -1.25 | 0.01 | 0.29 | < 0.01*** |
| Wealth index (ref. Rich) | | | | |
| Poor | -2.33 | 0.02 | 0.09 | < 0.01*** |
| Middle | -0.63 | 0.0 | 0.53 | < 0.01*** |
| Education level (ref. Higher education) | | | | |
| No Education | 0.10 | 0.01 | 1.11 | 0.01*** |
| Primary | -0.48 | 0.01 | 0.62 | < 0.01*** |
| Secondary | -0.56 | 0.0 | 0.57 | < 0.01*** |
| Marital status (ref. Divorced) | | | | |

| | | | | |
|---|---|---|---|---|
| Never Married | -1.17 | 0.0 | 0.31 | < 0.01*** |
| Married | -0.79 | 0.01 | 0.45 | < 0.01*** |
| Widowed | -0.73 | 0.0 | 0.48 | < 0.01*** |
| Household size | -0.26 | 0.0 | 0.00 | < 0.01*** |
| Age | -0.02 | 0.01 | 0.09 | 0.01*** |
| Residence (ref. Rural) | | | | |
| Urban | 0.97 | 0.01 | 2.64 | < 0.01*** |
| Literacy (ref. Cannot read at all) | | | | |
| Able to read only part of sentence | -1.06 | 0.01 | 0.35 | < 0.01*** |
| Able to read whole language | 0.45 | 0.01 | 1.57 | < 0.01*** |
| No card with required language | 2.00 | 0.0 | 7.39 | < 0.01*** |
| Blind/visually impaired | -1.37 | 0.0 | 0.25 | < 0.01*** |
| Frequency of reading Newspaper/magazine (ref. Not at all) | | | | |
| Less than once a week | 0.28 | 0.01 | 1.32 | < 0.01*** |
| At least once a week | -0.46 | 0.03 | 0.63 | < 0.01*** |
| Frequency of listening to Radio (ref. not at all) | | | | |
| Less than once a week | 0.48 | 0.02 | 1.61 | < 0.01*** |
| At least once a week | 0.09 | 0.02 | 1.10 | 0.01*** |
| Frequency of watching TV (ref. Not at all) | | | | |
| Less than once a week | -0.07 | 0.0 | 0.93 | < 0.01*** |
| At least once a week | 0.68 | 0.03 | 1.97 | < 0.01*** |
| Occupation (ref. Unskilled) | | | | |
| Skilled | -0.60 | 0.0 | 0.55 | < 0.01*** |
| Low skilled | -0.54 | 0.01 | 0.58 | < 0.01*** |
| High skilled | 0.28 | 0.0 | 1.32 | < 0.01*** |
| Working status (ref. No) | | | | |
| Yes | -0.49 | 0.02 | 0.61 | < 0.01*** |
| Coefficients Gaussian copula | | | | |

| | | | | |
|---|---|---|---|---|
| Tau | 0.0 | 0.0 | 1.00 | <0.01*** |

Significant. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

OR= Odds Ratio, “ref”= reference, “P-values” = Probability values

Table 5 presents the significant covariates for improved drinking water. As shown in the table, most of the covariates are significant to improved drinking water at a significant level of p-value = <0.01, except for no education, age and listening to the radio at least once a week. With reference to male-headed households, improved drinking water is lower in female-headed household (OR=0.29, p-value= <0.01). With reference to rich households, improved drinking water is lower for poor (OR=0.09, p-value= < 0.01) and in middle households (OR=0.53, p-value= < 0.01). In addition, with reference to higher education, improved drinking water is lower in primary education (OR=0.62, p-value= <0.01) and in secondary education (OR=0.57, p-value= <0.01). Furthermore, with reference to divorce, improved drinking water is lower in never married (OR=0.31, p-value= <0.01), married (OR=0.45, p-value= <0.01) and widowed (OR=0.48, p-value= <0.01).

Based on unit increase in household size, the odds of improved drinking water decrease by 100% (OR=0.00, p-value= <0.01). In addition, the odds of improved drinking water is higher in urban areas than in rural areas (OR= 2.64, p-value= <0.01).

Moreover, with reference to households where the head cannot read at all, improved drinking water is higher in households where the head is able to read whole language (OR=1.57, p-value= <0.01) and lower in household where the head can only read part of the sentence (OR=0.35, p-value= <0.01) and those that are blind/ visually impaired (OR=0.25, p-value= <0.01). With reference to not reading newspapers/magazine at all, improved drinking water is higher where reading is less than once a week (OR=1.32, p-value= <0.01), and lower where reading is at least once a week (OR=0.63, p-value= <0.01). In addition, with reference to not watching TV at all, improved drinking water is higher where watching is less than once a week (OR=0.93, p-value= <0.01), and lower where watching is at least once a week (OR=1.97, p-value= <0.01). With reference to not listening to the radio at all, improved drinking water is higher in households who at least listen to the radio less than once a week (OR=1.61, p-value= <0.01).

Furthermore, with reference to unskilled occupation, improved drinking water is lower in skilled (OR=0.55, p-value= <0.01) and low skilled (OR=0.58, p-value= <0.01) households, while it is higher in high skilled households (OR=1.32, p-value=<0.01). The maximum simulated likelihood estimate of the dependence parameter tau is 0.01 km, a value that shows the presence of significant.

#### 4.5.1.3 Spatial distribution of improved drinking water from model estimates

Results of the maps for predicted improved drinking water and standard errors are given in Figure 2(a) and Figure 2(b) respectively. The maps were generated from the best model, fitted with fixed effects. Figure 2(a) shows the predicted improved drinking water for Namibia obtained from the model fitted with fixed effects. The darker blue on the map indicated that the area is predicted to have high improved drinking water, while the light blue indicate areas where predicted improved drinking water is low. From Figure 2(a), Khomas, Otjozondjupa, Omaheke and Erongo regions are predicted to have higher improved drinking water, while Kavango, Ohangwena, Omusati and Kunene regions are predicted to have low improved drinking water. Other regions such as Hardap and !Karas regions are predicted to have average improved drinking water. See Appendix C for the names of the regions.

(a) (b)

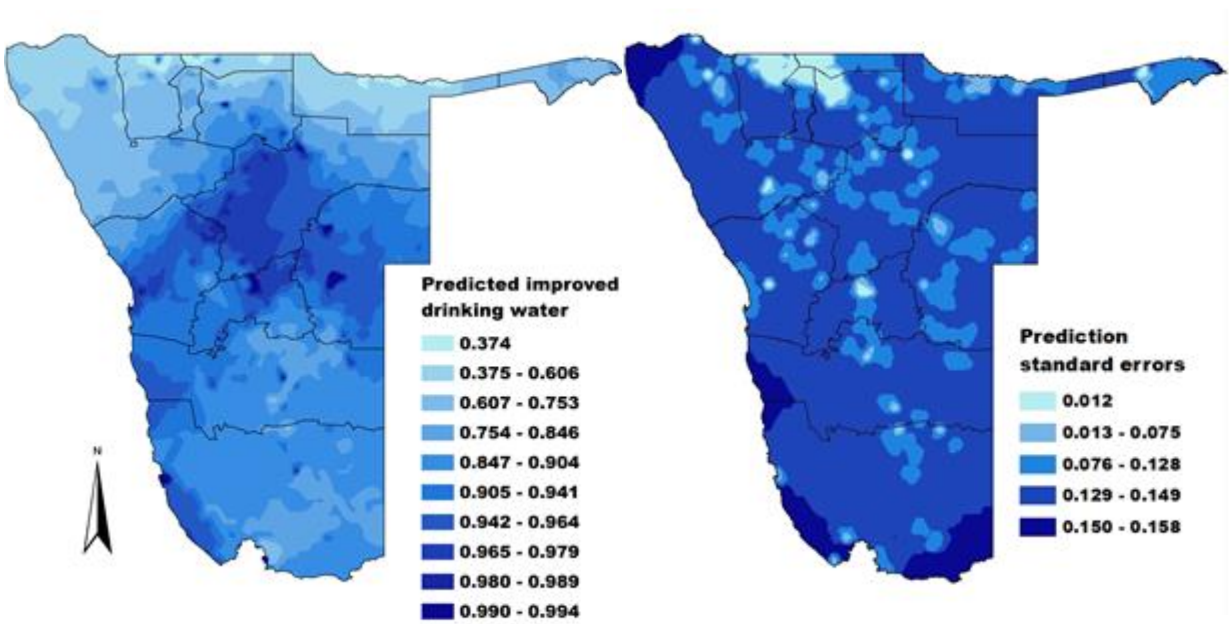



Figure 2: (a) Map showing predicted improved drinking water based on fitted values of the predictive distribution. (b) Prediction standard errors obtained from the GCMR best model with fixed effects. Mapping was carried out in ArcGIS. See Appendix C for the names of the regions.

Figure 2(b) displays a map of the prediction standard errors of improved drinking water. As expected, it illustrates low prediction errors around data locations. High errors were observed in areas that were far from data locations as shown in Figure 1(a). For instance, in the north-west and south-east of !Karas region, the errors rapidly increased with increased distance from data locations. Similarly, for the west of Hardap and the south-west of Kunene regions. Low prediction errors were depicted mostly in the northern part of Namibia.

### 4.5.2 Improved sanitation

The results of the two models fitted for improved sanitation are given in Tables 6 and 7 respectively.

#### 4.5.2.1 Model selection

The AIC results from the two models fitted for improved sanitation are displayed in Table 6. The inclusion of region in the second model yield an increase in the AIC value, thus making the model with fixed effects the best model as shown in Table 6.

Table 6: AIC results from improved sanitation models fitted using gcmr package

| Model | Likelihood | AIC |
|---|---|---|
| Improved sanitation with fixed effects | 2214.7 | 4485.4 |
| Improved sanitation with an additional covariate region | 3009.3 | 6098.7 |

#### 4.5.2.2 Model with fixed effects results

The results of the model for improved sanitation fitted with fixed effects are displayed in Table 7.

Table 7: Results from the Gaussian copula marginal regression model for improved sanitation fitted using gcmr package

| Coefficients | Estimates | Standard Errors | Odds Ratio (OR) | P-values |
|---|---|---|---|---|
| Sex (ref. Male) | | | | |
| Female | -0.94 | 0.0 | 0.39 | < 0.01*** |
| Wealth index (ref. Rich) | | | | |
| Poor | -4.93 | 0.0 | 0.01 | < 0.01*** |
| Middle | -3.44 | 0.0 | 0.03 | < 0.01*** |
| Education level (ref. Higher education) | | | | |
| No Education | -2.98 | 0.0 | 0.05 | < 0.01*** |
| Primary | -2.16 | 0.0 | 0.16 | < 0.01*** |
| Secondary | -1.88 | 0.01 | 0.15 | < 0.01*** |
| Marital status (ref. Divorced) | | | | |
| Never Married | 0.66 | 0.07 | 1.94 | < 0.01*** |
| Married | -0.14 | 0.0 | 0.87 | < 0.01*** |
| Widowed/ | -0.14 | 0.0 | 0.87 | < 0.01*** |
| Household size | -0.03 | 0.02 | 0.97 | 0.08 . |
| Age | 0.04 | 0.0 | 1.04 | < 0.01*** |
| Residence (ref. Rural) | | | | |
| Urban | 0.26 | 0.01 | 1.29 | < 0.01*** |
| Literacy (ref. Cannot read at all) | | | | |
| Able to read only part of sentence | -0.49 | 0.0 | 0.61 | < 0.01*** |
| Able to read whole language | -0.50 | 0.01 | 0.61 | < 0.01*** |
| No card with required language | -0.11 | 0.0 | 0.89 | < 0.01*** |
| Blind/visually impaired | -2.39 | 0.0 | 0.09 | < 0.01*** |

| | | | | |
|---|---|---|---|---|
| Frequency of reading Newspaper/magazine (ref. Not at all) | | | | |
| Less than once a week | -0.15 | 0.0 | 0.86 | <0.01*** |
| At least once a week | 0.15 | 0.02 | 1.16 | <0.01*** |
| Frequency of listening to Radio (ref. Not at all) | | | | |
| Less than once a week | -0.11 | 0.0 | 0.89 | <0.01*** |
| At least once a week | -0.22 | 0.01 | 0.80 | <0.01*** |
| Frequency of watching Television (ref. Not at all) | | | | |
| Less than once a week | -0.02 | 0.0 | 0.98 | <0.01*** |
| At least once a week | 0.11 | 0.02 | 1.12 | 0.01*** |
| Occupation (ref. Unskilled) | | | | |
| Skilled | 0.11 | 0.0 | 1.11 | <0.01*** |
| Low skilled | -0.07 | 0.0 | 0.94 | <0.01*** |
| High skilled | 0.18 | 0.0 | 0.84 | <0.01*** |
| Working status (ref. No) | | | | |
| Yes | 0.16 | 0.01 | 1.17 | <0.01*** |
| Coefficients Gaussian copula | | | | |
| Tau | 0.0 | 0.0 | 1.00 | <0.01*** |

Significant. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

OR= Odds Ratio, “ref”= reference, “P-values” = Probability values

Table 7 shows that covariate household size and watching television once a week were not significant to improved sanitation. As shown in table 7, with reference to male-headed households, improved sanitation is lower in female-headed households (OR=0.39, p-value= <0.01). In addition, with reference to rich household, improved sanitation is lower in poor households (OR=0.01, p-value= <0.01) or middle households (OR=0.03, p-value= <0.01). Moreover, with reference to divorce, improved sanitation is higher in never married (OR=1.94, p-value= <0.01) and lower in married (OR=0.87, p-value= <0.01) or widowed (OR=0.87, p-value= <0.01). Furthermore, reference to higher education, improved sanitation is lower in no education (OR=0.05, p-value= <0.01), primary education (OR=0.16, p-value= <0.01) and secondary (OR=0.15, p-value= <0.01).

With unit increase in the age of head of household, the odds of improved sanitation increase by 4% (OR= 1.04, p-value= <0.01). In addition, the odds of improved sanitation is higher in urban areas than in rural areas (OR= 1.29, p-value= <0.01). Furthermore, with reference to a household where the head cannot read at all, improved sanitation is lower in households where the head is able to read whole language (OR=0.61, p-value= <0.01), can only read part of the sentence (OR=0.61, p-value= <0.01) or no card with required language (OR=0.89, p-value= <0.01) and those that are blind/ visually impaired (OR=0.09, p-value= <0.01).

The results also shows that, with reference to not reading newspapers/magazine at all, improved sanitation is lower in reading less than once a week (OR=0.86, p-value= <0.01), and higher in reading at least once a week (OR=1.16, p-value= <0.01). In addition, with reference to not watching television, improved sanitation is lower in households were people watch television less than once a week (OR=0.98, p-value= <0.01). Moreover, with reference to not listening to the radio at all, improved sanitation is lower in listening less than once a week (OR=0.89, p-value= <0.01), and reading at least once a week (OR=0.80, p-value= <0.01). Furthermore, with reference to unskilled, improved sanitation is lower in low skilled (OR=0.94, p-value= <0.01) and higher in skilled (OR=1.11, p-value= <0.01) and high skilled household (OR=0.84, p-value=0.01).

The odds of improved sanitation is higher in households were the respondent is working (OR=1.17, p-value= <0.01). The maximum simulated likelihood estimate of the dependence parameter tau is 0.01 km and thus indicates the presence of significance.

#### 4.5.2.3 Spatial distribution of improved sanitation from model estimates

The maps below were generated from the best model, fitted with fixed effects. The map of predicted improved sanitation is given in Figure 3(a). Predicted improved sanitation was found to be high partly in Khomas, Otjozondjupa and Hardap regions. Regions such as Zambezi, Kavango, Ohangwena, Omusati, Oshana and Kunene are predicted to have low improved sanitation. Part of !Karas, Hardap and Erongo regions is predicted to have an average improved sanitation. The prediction standard errors of improved sanitation are shown in Figure 3(b). As one would expect, high errors values were obtained closer to the data locations. Omusati, Oshikoto and Oshana regions had small errors due to the data points being close to each other. See Appendix C for the names of the regions.

(a) (b)

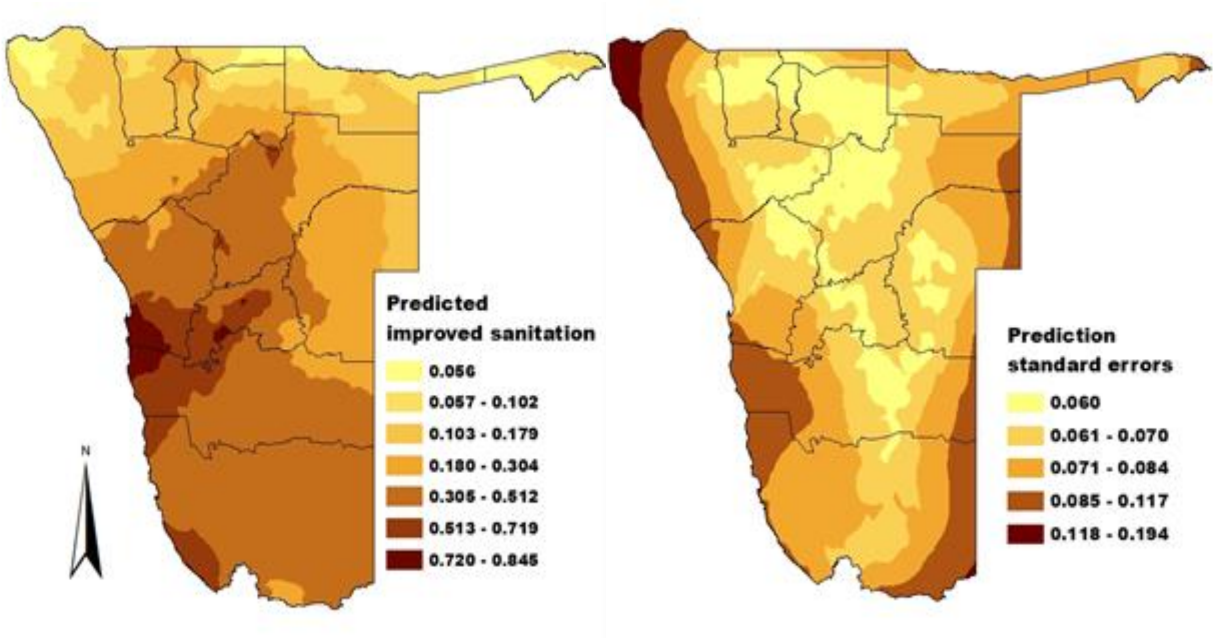



Figure 3: (a) Map showing predicted improved sanitation based on fitted values of the predictive distribution. (b) Prediction standard errors obtained from the GCMR best model with fixed effects. Mapping was carried out in ArcGIS. See Appendix C for the names of the regions.

### 4.5.3 Improved hygiene

This section presents results of two models fitted for improved hygiene given in Tables 8 and 9 respectively. In addition, smooth maps for the predicted improved hygiene and prediction standard errors are shown in Figure 4.

#### 4.5.3.1 Model selection

Table 8 shows the AIC results from the models fitted for improved hygiene. The inclusion of region in the second model of improved hygiene yield a drop in the AIC, thus making the model with spatial effects the best model as shown in Table 8.

Table 8: AIC results from improved hygiene models fitted using gcmr package

| Model | Likelihood | AIC |
|---|---|---|
| Improved hygiene with fixed effects | 4039.4 | 8134.9 |
| Improved hygiene with an additional covariate region | 3727.5 | 7534.9 |

#### 4.5.3.2 Model with an additional covariate region results

The results of the model for improved hygiene fitted with an additional covariate region are presented in Table 9.

Table 9: Results from the Gaussian copula marginal regression model for improved hygiene fitted using gcmr package

| Coefficients | Estimates | Standard Errors | Odds Ratio | P-values |
|---|---|---|---|---|
| Sex (ref. Male) | | | | |
| Female | 0.19 | 0.07 | 1.22 | 0.01** |
| Wealth index (ref. Rich) | | | | |
| Poor | -1.38 | 0.06 | 0.25 | < 0.01*** |
| Middle | -0.56 | 0.04 | 0.57 | < 0.01*** |
| Education level (ref. Higher education) | | | | |
| No Education | -1.70 | 0.06 | 0.18 | < 0.01*** |
| Primary | -2.06 | 0.06 | 0.13 | < 0.01*** |
| Secondary | -1.27 | 0.06 | 0.28 | < 0.01*** |
| Marital status (ref. Divorced) | | | | |
| Never Married | 0.51 | 0.05 | 1.67 | < 0.01*** |
| Married | 0.96 | 0.06 | 2.61 | < 0.01*** |
| Widowed | 0.69 | 0.04 | 2.01 | < 0.01*** |
| Household size | -0.14 | 0.01 | 0.87 | < 0.01*** |
| Age | 0.02 | 0.0 | 1.02 | < 0.01*** |
| Residence (ref. Rural) | | | | |
| Urban | 0.26 | 0.04 | 1.29 | 0.01*** |
| Literacy (ref. Cannot read at all) | | | | |
| Able to read only part of sentence | -0.18 | 0.09 | 0.84 | 0.04* |
| Able to read whole language | -0.21 | 0.07 | 0.81 | 0.01** |
| No card with required language | -1.06 | 0.03 | 0.35 | < 0.01*** |
| Blind/visually impaired | 1.19 | 0.01 | 3.29 | < 0.01*** |
| Frequency of reading Newspaper/magazine (ref. Not at all) | | | | |

| | | | | |
|---|---|---|---|---|
| Less than once a week | -0.0 | 0.06 | 0.99 | 0.95 |
| At least once a week | -0.09 | 0.06 | 0.91 | 0.15 |
| Frequency of listening to Radio (ref. Not at all) | | | | |
| Less than once a week | 0.16 | 0.07 | 1.18 | 0.01* |
| At least once a week | 0.09 | 0.05 | 1.09 | 0.08. |
| Frequency of watching Television (ref. Not at all) | | | | |
| Less than once a week | 0.03 | 0.07 | 1.03 | 0.69 |
| At least once a week | 0.27 | 0.05 | 1.30 | 0.01*** |
| Occupation (ref. Unskilled) | | | | |
| Skilled | 0.62 | 0.07 | 1.86 | < 0.01*** |
| Low skilled | -0.13 | 0.04 | 0.88 | 0.01** |
| High skilled | 0.39 | 0.07 | 1.47 | 0.01*** |
| Working status (ref. No) | | | | |
| Yes | 0.17 | 0.05 | 1.18 | 0.01*** |
| Region (ref. Zambezi) | | | | |
| Erongo | -0.11 | 0.09 | 0.89 | 0.24 |
| Hardap | 0.17 | 0.11 | 1.19 | 0.13 |
| !Karas | -0.18 | 0.07 | 0.84 | 0.01** |
| Kavango | -0.23 | 0.01 | 0.79 | < 0.01*** |
| Khomas | 0.24 | 0.08 | 1.27 | 0.01** |
| Kunene | 0.93 | 0.10 | 2.54 | < 0.01*** |
| Ohangwena | -0.29 | 0.10 | 0.75 | 0.01** |
| Omaheke | 0.70 | 0.09 | 2.02 | < 0.01*** |
| Omusati | -0.25 | 0.08 | 0.78 | 0.01** |
| Oshana | 0.70 | 0.08 | 2.02 | < 0.01*** |
| Oshikoto | 0.30 | 0.09 | 1.35 | 0.02** |
| Otjozondjupa | 1.09 | 0.11 | 2.99 | < 0.01*** |
| Coefficients Gaussian copula | | | | |

| | | | | |
|---|---|---|---|---|
| Tau | 0.01 | 0.01 | 1.00 | <0.01*** |

Significant. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

OR= Odds Ratio, ref"= reference, "P-values" = Probability values

Table 9 shows that, with reference to rich households, the odds of improved hygiene is lower in poor households (OR= 0.25, p-value= <0.01) and middle households (OR= 0.57, p-value= <0.01). In addition, reference to higher education, improved hygiene is lower in no education (OR= 0.18, p-value= <0.01), primary education (OR= 0.13, p-value= <0.01) and secondary (OR=0.28, p-value= <0.01). Furthermore, with reference to divorce, improved hygiene is higher in never married (OR= 1.67, p-value= <0.01), married (OR= 2.61, p-value= <0.01) and widowed (OR= 2.01, p-value= <0.01).

Based on the unit increase of the household size, the odds of improved hygiene decreased by 14% (OR= 0.87, p-value= <0.01). In addition, for a unit increase in the age of head of household, the odds of improved hygiene increase by 2% (OR= 1.02, p-value= <0.01). With reference to unskilled occupation, improved hygiene is higher in households with skilled (OR= 1.86, p-value= <0.01).

With reference to Zambezi, the odds of improved hygiene is lower in Kavango (OR= 0.79, p-value= <0.01) and higher in Kunene (OR= 2.54, p-value= <0.01), Omaheke (OR= 2.02, p-value= <0.01), Oshana (OR= 2.02, p-value= <0.01) and Otjozondjupa (OR= 2.99, p-value= <0.01). The summary confirms that covariate region contains relevant information about the spatial variation of improved hygiene. The parameter tau in the model with region shows that the spatial dependence is significant.

#### 4.5.3.3 Spatial distribution of improved hygiene from model estimates

The maps in Figure 4 were generated from the best model, fitted with an additional covariate region. The predicted map in Figure 4(a) shows that improved hygiene is mostly high in Otjozondjupa region, while Omaheke and Khomas regions are predicted to have moderate improved hygiene. Zambezi, Kavango, Oshikoto and Ohangwena regions are predicted to have the lowest improved hygiene. Moreover, !Karas, Hardap, Erongo and Kunene regions are predicted to have somewhat improved hygiene. Similarly for Figure 2(b) and Figure 3(b), high error values were obtained at areas far away from the data locations. See Appendix C for the names of the regions.

(a) (b)

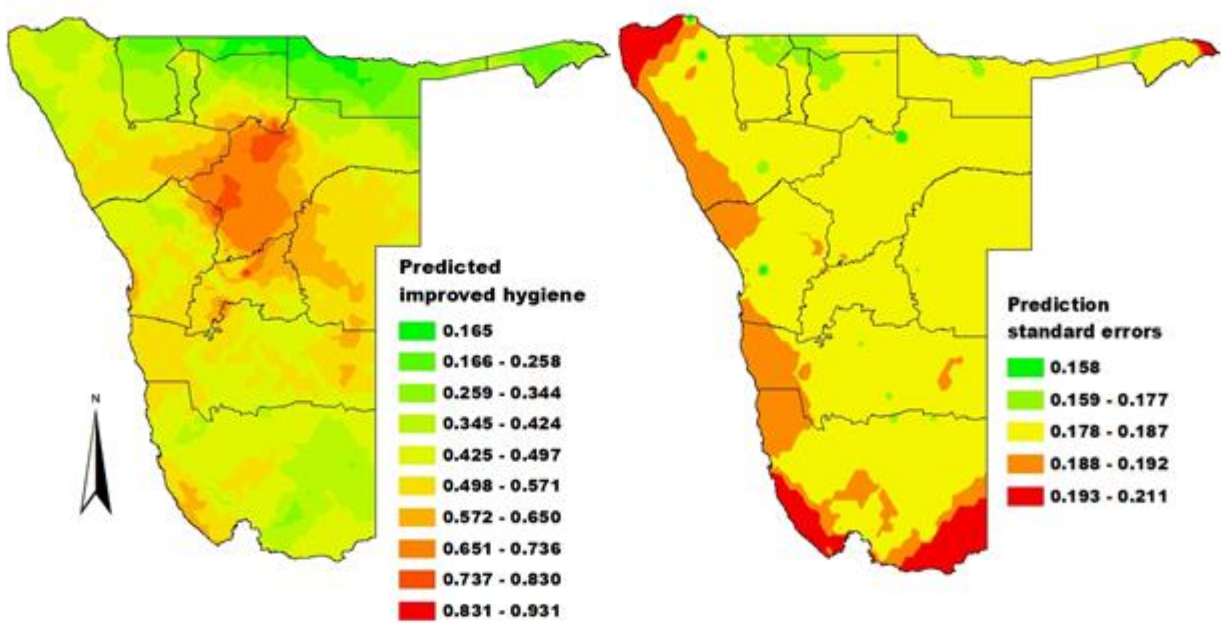



Figure 4: (a) Map showing predicted improved hygiene based on fitted values of the predictive distribution. (b) Prediction standard errors obtained from the GCMR best model with spatial effects. Mapping was carried out in ArcGIS. See Appendix C for the names of the regions.

**4.6 Conclusions**

Namibia is currently focusing efforts to improve access to safe drinking water, adequate sanitation and proper hygiene. The study presented determinants of improved drinking water, sanitation and hygiene and had revealed areas that have low improved drinking water, sanitation and hygiene through mapping. The results shows that the covariates examined in the study contribute significant to households having improved drinking water, sanitation and hygiene. To be precise, the study suggest that improved drinking water is influenced by sex of head of household, wealth index, marital status, residence, literacy, household size, exposure to media, occupation and working status. Covariates such as sex of head of household, wealth index, education level, marital status, age of head of household, residence, literacy, exposure to media, occupation and working status contribute to a household having improved sanitation. The study has also noted that wealth index, education level, marital status, household size, age of head of household, household size contribute greatly to a household having access to improved hygiene.

The study has noted that male headed households are more likely to have improved drinking water and sanitation compared to the female headed households. Therefore, there is a need to empower women in Namibia through financial awareness and education in order to be able to achieve the goal 6 of the sustainable development goals. The level of education of the household head significantly influence access to improved drinking water, sanitation and hygiene. Hence, it is crucial to provide basic education to all people as this will help them to manage their WASH and its application in the future.

The wealth index of a household is associated with the affordability of improved drinking water, proper sanitation and having better hygiene. Therefore, regional planning should focus on poor households when implementing strategies for communities to have access to improved drinking water, sanitation and hygiene. There is a need to also address the socio-economic factors that are associated with inadequate access to improved drinking water, sanitation and hygiene. This will help ensure equal access to the basic needs and play an important role in the prevention and reduction of the burden of water borne diseases.

Furthermore, the study found that more efforts are required to increase access to improved drinking water in Zambezi, Kavango, Omusati, Ohangwena, Oshana !Karas and Hardap and Kunene regions. For improved sanitation, efforts should be done in Zambezi, Omusati, Oshana, Ohangwena and Kunene regions, while for hygiene more efforts should be done in Zambezi, Kavango, Oshana and Omusati regions. The findings from the study could be used by the government of Namibia to identify areas of concern and efficiently allocate resources to them.

# CHAPTER 5: SIMULTANEOUS MODELLING OF WATER, SANITATION AND HYGIENE USING GENERALIZED JOINT REGRESSION MODEL

## 5.1 Introduction

Chapter 5 presents results of the joint regression model of WASH indicators, applying a copula approach. Different models have been fitted for goodness of fit and adjusted or penalized for model complexity.

According to Zimmer and Trivedi (2006), many modelling approaches either use fully parametric model or use few data semiparametric framework that requires partial specification of the interdependent equation. However, these approaches poses challenges because their joint distribution of variables of interest maybe unknown, thus making it difficult to model. In addition, their joint distribution are less tractable, thus marginal distribution of joint modelling provide an alternative to model without difficulties. Hence, a joint copula approach then allows to effectively capture dependence structure of data, without giving up appealing properties of the marginal. Moreover, joint model have the prospective to decrease estimate bias (Sudell, Kolamunnage-Dona, & Tudur-Smith, 2016) and enable the inclusion of covariates that might have errors.

Joint analysis of WASH indicators is of importance as improved drinking water and sanitation alone would not guarantee health, but rather adding hygiene is required to improve health (Garriga & Foguet, 2013), thus allowing to establish the association between these outcomes. In addition, WASH indicators are often handled jointly as without improved water or proper sanitation, hygiene can be an issue as well. According to Costello (2013), water is important for hygiene practice especially handwashing with soap. If a household lacks water, risky hygiene behavior such as not washing hands after critical time are likely to develop. Furthermore, Ntozini et al. (2015) pointed out that access to improved drinking water and sanitation are important determinants of behavioral response to improved hygiene and intervention.

This study used areal data for joint analysis of WASH indicators as opposed to point-referenced data used in chapter 4. This is because areal data can account for spatial dependence and spatial regression inference, hence applying smoothing to model for complexity. Areal data are known to be neighbors because they share a common boundary (Monogan, 2013), thus necessary for joint analysis. In terms of the Markov Random Field (MRF) that will be applied, the areal setting $Y_1, \ldots, Y_n$ and a neighborhood matrix use a different approach as compared to point references data for joint distribution $f(y_1, \ldots, y_n)$ where at least two or more random variables are considered as the case for this study (Spatial statistics, n.d.).

## 5.2 Generalized Joint Regression Model

In general, a trivariate binary model can be expressed as

$$y_{xi} = V_{xi}^{T} y_x + \sum_{v_x=1}^{\acute{N}x} S_{xv_x} \left( Z_{xv_{x\hat{\imath}}} \right) + \epsilon_{xi}, \ i = 1, \ldots, n, \forall_x = 1,2,3 \qquad (21)$$

where $n$ is the sample size, $y_{xi}$ is a latent continuous variable, $V_{xi}$ is the binary or categorical outcome, $y_x$ is a vector that represent the effect of the variables in $V_{xi}$ and

$S_{xv_x}(Z_{xv_{x\hat{\imath}}})$ is a smooth function of continuous covariate $(Z_{xv_{x\hat{\imath}}})$, $\forall v_x = 1, \dots, \acute{\mathrm{N}}_x$ with $\acute{\mathrm{N}}_x$ being the number of smooth terms ( Filippou, Marra, & Radice, 2017).

Regression models are commonly fitted with one response variable and a set of covariates (Filippou et al., 2017). However, multiple response variables can also be fitted. Joint regression models can be applied to continuous, discrete and mixed correlated outcomes. The most common choice is the Gaussian copula given by

$$C(u_1, u_2, u_3|\rho) = \Phi_3(\Phi^{-1}(u_1), \Phi^{-1}(u_2), \Phi^{-1}(u_3)) \tag{22}$$

The dependency in (22) has been captured using a trivariate Gaussian copula cumulative distribution function $C(u_1, u_2, u_3|\rho)$, where $u_1, u_2, u_3$ are the random variables, $\Phi$ is the standard normal distribution, and $\Phi_3$ is the standard trivariate normal distribution, with dependence parameter $\rho$.

It has been noted that constructing trivariate copula can be challenging, however Song, Li and Yuan (2009) pointed out that Gaussian copula is an exception because of its flexible dependence structure. Copula dependence and marginal parameters can easily be estimated through Gaussian method. Each parameters depends on the joint model predictor including different type of covariates effects such as linear, non-linear, random and spatial effects (Marra & Radice, 2017a). A joint Cumulative Distribution Function (CDF) of three discrete random variables $Y_1, Y_2, Y_3$ can be expressed as

$$F(y_1, y_2, y_3) = C(F_1(y_1|u_1, \sigma_1, v_1), F_2(y_1|u_2, \sigma_2, v_2), F_3(y_3|u_3, \sigma_3, v_3); \xi, \theta), \tag{23}$$

where $\vartheta = (u_1, \sigma_1, v_1, u_2, \sigma_2, v_2, u_3, \sigma_3, v_3, \xi, \theta)^T, F_1(y_1|u_{1,}\sigma_1, v_1), F_2(y_2|u_2, \sigma_2, v_2)$

and $F_3(y_3|u_3, \sigma_3, v_3)$ are marginal CDFs of $Y_1, Y_2, Y_3$ and $u_m, \sigma_m$ and $v_m$ for $m = 1,2,3$ are marginal distribution parameters, C is a defined 3-place copula function with dependence coefficient $\theta$ and $\xi$ is the number of degree of freedom (Marra & Radice, 2017a). In terms of joint modelling, a maximum likelihood approach is used as it can combine random effects.

### 5.3 Model fitting

This study applied a copula regression to joint three WASH indicators, quantify the determinants of improved water, sanitation and hygiene, and capture spatial dependency. A Gaussian copula was assumed. The MRF smoother was applied to model the spatial dependence based on the geographic location of survey respondents, as introduced in the mgcv package by Wood (2017). The variable region was fitted as an unstructured random effect. This method is common when the geographic areas or countries of interest are divided into discrete connecting geographic units or regions (Marra, Radice, Bärnighausen, Wood, & McGovern, 2017). Based on the 2013 NDHS data, there were 13 geographic regions and this allowed us to use the information contained in the neighbouring observation that is located in the same country. Spatial pattern suggests that areal observations close to each other have more similar values than those that are far from each other. Thus, $N_{ij} = \begin{cases} 1\,, & if\ i\ and\ j\ are\ neighbuors \\ 0 & otherwise \end{cases}$

For continuous variables such as age and size of household, the smooth function was applied using the regression spline approach by Eilers and Marx (1996). These splines are used to model non-linear effects for variables such as age and household size. Model fitting used a SemiParTRIV () function in the GJRM package by Marra and Radice (2017b) in R to fit the model. A SemiParTRIV fits a flexible trivariate binary model with several types of covariate effects. One joint model was fitted for WASH indicators. The model was fitted with spatial effects to capture regional spatial variation using a penalized likelihood estimation. The non-linear effects were included in the model to capture non-linearity for variables, household size & age. All the trivariate response used were discrete. The WASH indicators in the trivariate were assumed to be correlated as an unrelated regression model.

Considering a trivariate model and let $y$ be the binary indicator at location $i$. Thus the joint (trivariate) model is given by

$$y_i = W_\beta + s(non - linear) + s(spatial) \tag{24}$$

where $W_\beta$ are the fixed effects fitted and $s$ is the smooth function applied.

In this study, the equation of GJRM fitted for Table 10 was as follow:

$$\log p(improved\ hygiene) = y_{1i}^T\, \alpha_{11} + s_{11}(Members) + s_{12}(Age) + s_{13}(Region) + \ \varepsilon_i \tag{25}$$

$$\log p(improved\ water) = y_{2i}^T\, \alpha_{21} + s_{21}(Members) + s_{22}(Age) + s_{23}(Region) + \ \varepsilon_i$$

$$\log p(improved\ sanitation) = y_{3i}^T\, \alpha_{31} + s_{31}(Members) + s_{32}(Age) + s_{33}(Region) + \ \varepsilon_i$$

where $y_{1i,}\ \ y_{2i}$ and $y_{3i}$ are binary and categorical variables, $\alpha_{11}, \alpha_{21}$ and $\alpha_{31}$ represent the effects of the variables in $y_i$, while $s_u$ for $u = 1, 2, 3\ and\ k$ are smooth functions of members, age and region and $\varepsilon_i$ is the error terms. The variables included in $y_{1i,}\ \ y_{2i}$ and

$y_{3i}$ were: sex, marital status, wealth, literacy, newspaper, radio, TV, respondent occupation and working status.

### 5.4 Joint Model results

Under this section, we present results of a joint spatial model of WASH indicators that was fitted to capture spatial variation and assess dependence. In this section, age, household size and region were applied as smooth functions to the model spatial effects.

Table 10: Results from the GJRM with spatial, fixed and non-linear effects of region to capture spatial variation and spline to model non-linear effect for variables age and household size, for improved drinking water, hygiene and sanitation.

| Random variables | Improved Hygiene | | Improved drinking water | | Improved sanitation | |
|---|---|---|---|---|---|---|
| Coefficients | Estimates | P-values | Estimates | P-values | Estimates | P-values |
| Residence (ref. Urban) | | | | | | |
| Rural | -0.29 | <0.00*** | -0.53 | <0.00*** | -0.27 | 0.00*** |
| Sex (ref. Male) | | | | | | |
| Female | -0.00 | 0.90 | -0.09 | 0.04* | -0.04 | 0.30 |
| Marital status (ref. Never married) | | | | | | |
| Married | 0.02 | 0.52 | -0.02 | 0.63 | -0.03 | 0.52 |
| Widowed | -0.11 | 0.05* | -0.16 | 0.03* | -0.07 | 0.32 |
| Divorced | -0.14 | 0.05* | -0.15 | 0.11 | -0.04 | 0.65 |
| Wealth index (ref. Poor) | | | | | | |
| Middle | 0.38 | <0.00*** | 0.53 | <0.00*** | 1.09 | <0.00*** |
| Rich | 1.06 | <0.00*** | 1.15 | <0.00*** | 2.59 | <0.00*** |
| Literacy (ref. Cannot read at all) | | | | | | |

| | | | | | | |
|---|---|---|---|---|---|---|
| Able to read only part of sentence | 0.04 | 0.48 | 0.01 | 0.89 | 0.04 | 0.62 |
| Able to read whole language | 0.08 | 0.07 . | 0.09 | 0.14 | -0.06 | 0.31 |
| No card with required language | -0.01 | 0.96 | 0.04 | 0.79 | -0.09 | 0.47 |
| Blind/visually impaired | 0.68 | 0.14 | -033 | 0.56 | 1.19 | 0.09 . |
| Frequency of reading Newspaper/Magazine (ref. Not at all) | | | | | | |
| Less than once a week | 0.00 | 0.89 | -0.04 | 0.48 | 0.05 | 0.27 |
| At least once a week | -0.07 | 0.09 . | -0.03 | 0.57 | -0.02 | 0.76 |
| Frequency of listening to the Radio (ref. Not at all) | | | | | | |
| Less than once a week | -0.09 | 0.04* | -0.02 | 0.71 | -0.08 | 0.17 |
| At least once a week | -0.04 | 0.25 | -0.06 | 0.23 | 0.04 | 0.43 |
| Frequency of watching Television (ref. Not at all) | | | | | | |
| Less than once a week | -0.08 | 0.11 | -0.02 | 0.77 | -0.11 | 0.09 . |
| At least once a week | 0.017 | 0.65 | -0.03 | 0.59 | -0.02 | 0.73 |
| Occupation (ref. Unskilled) | | | | | | |
| Skilled | -0.13 | 0.19 | 0.01 | 0.94 | -0.41 | 0.00** |
| Low skilled | -0.14 | 0.11 | -0.02 | 0.88 | -0.41 | 0.00*** |
| High skilled | -0.08 | 0.42 | -0.13 | 0.35 | 0.44 | 0.00*** |
| Working status (ref. No) | | | | | | |
| Yes | 0.08 | 0.32 | -0.07 | 0.54 | 0.38 | 0.00*** |

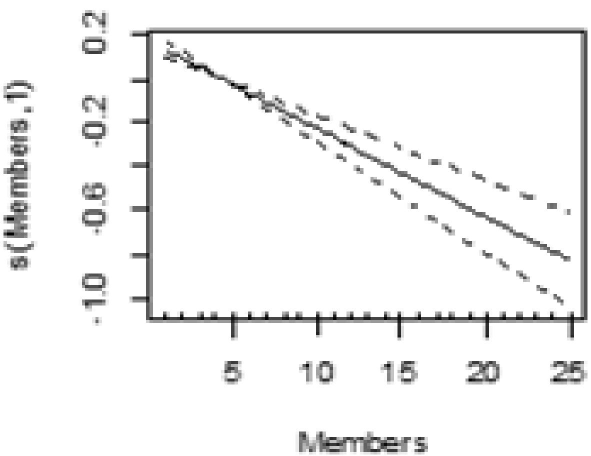
Significant. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

From Table 10, with regard to the covariates of improved hygiene, estimates indicate an association with residence, listening to the radio less than once a week and wealth index. However, there is no significance in covariates marital status, literacy, exposure to newspaper, exposure to TV, respondent occupation and working status. Figure 5 reports the smooth function estimate for improved hygiene. The effects of household members, age and region show different degree of non-linearity. The intervals of the smooth for household members and age ranged from -1.0 to 0.2, while for region they range between zero line. This suggests that household size and age are good predictors of improved hygiene and that region is not an important determinant of improved hygiene. There is a negative association with increase to household members as shown in Figure 5(a).

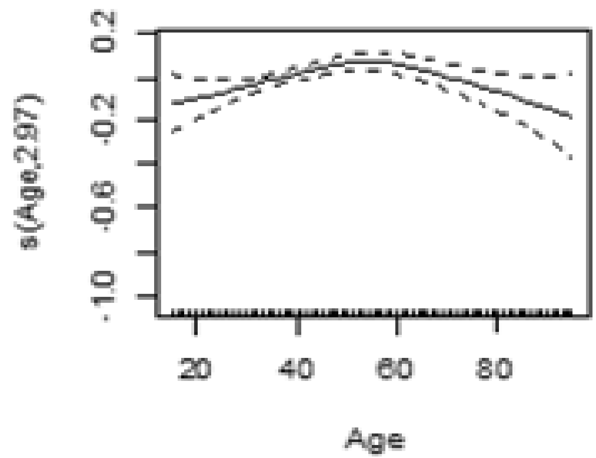
Figure 5(c) shows a high spatial variation of improved hygiene in Khomas, Omaheke, Erongo, Otjozondjupa and Kavango regions, while !Karas and Oshana regions have the lowest. Hardap, Kunene, Omusati and Oshikoto regions shows a moderate spatial variation. See Appendix C for the names of the regions.

The results in Table 10 shows that improved drinking water is associated with residence, sex and wealth, with sex negatively associated with household members as shown in Figure 6(a). The intervals of the smooth for household members and age ranged from -0.8 to 0.0, while for region ranged between -0.4 to 0.8. The p-value shows that region is an important determinant of improved drinking water.  Figure 6(c) shows a high spatial variation of improved drinking water in Zambezi, Oshana, Ohangwena, Omaheke and Omusati regions, with low spatial variation in Erongo, Khomas, Hardap, Kunene, Otjozondjupa and !Karas regions. Smooth estimate of association between improved drinking water and age are significant, p-value = 0.00. See Appendix C for the names of the regions.

(a)




(b)




(c)

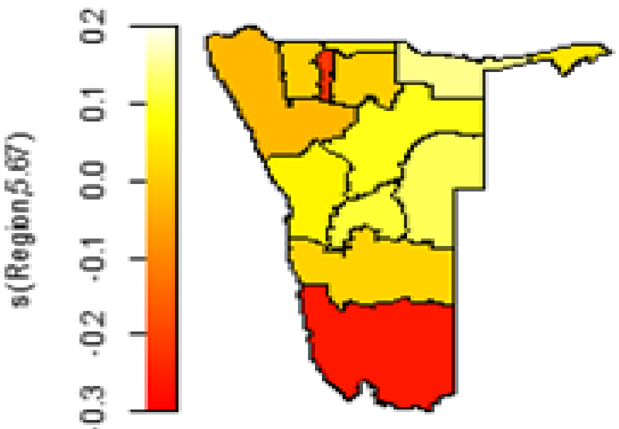



Figure 5: (a) The smooth function household size (b) The smooth function age, (c) The region spatial variation of improved hygiene. P-values for the smooth terms of household members, age and region were 0.00, 0.00 and 0.00 respectively. See Appendix C for the names of the regions.

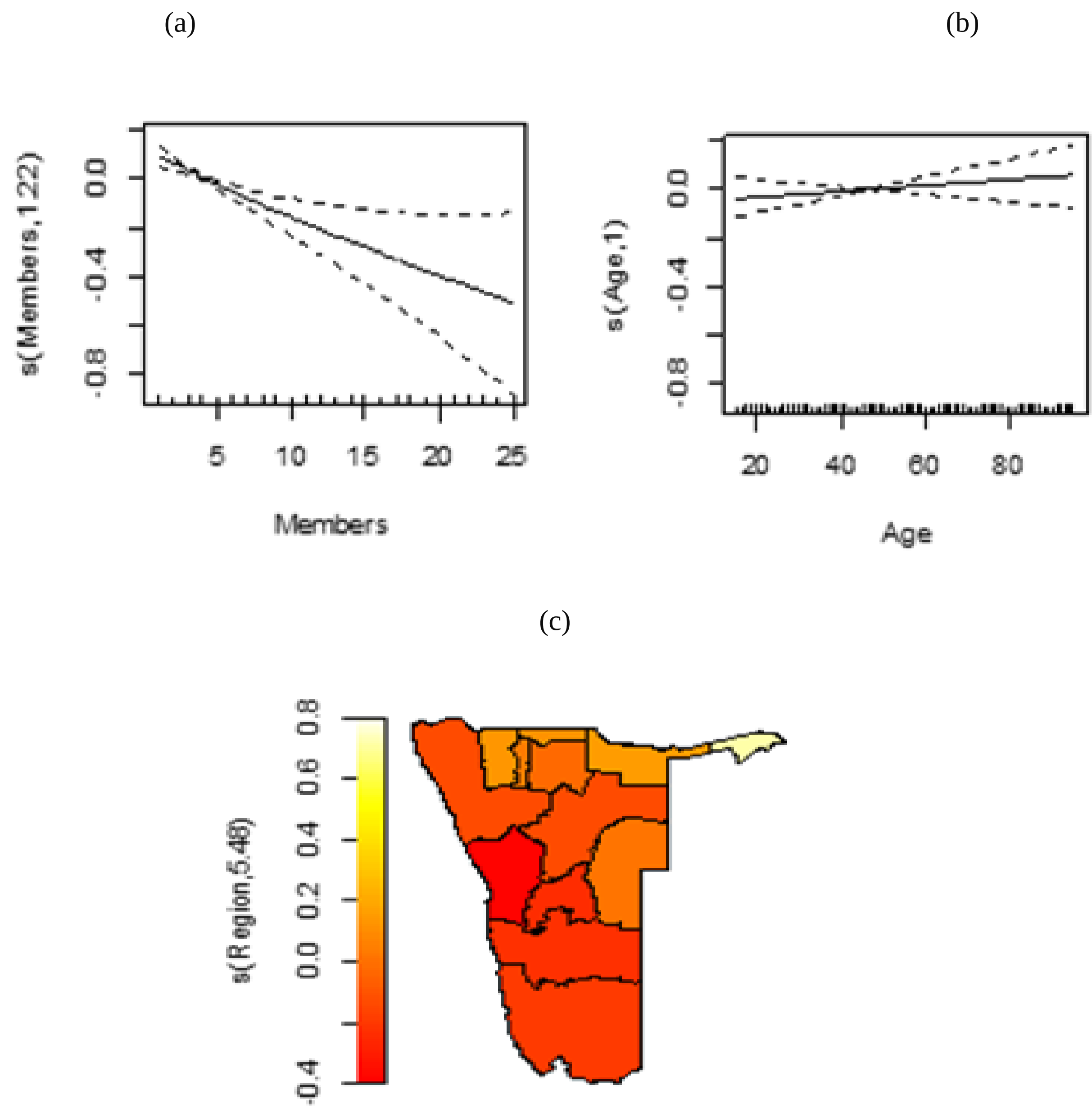
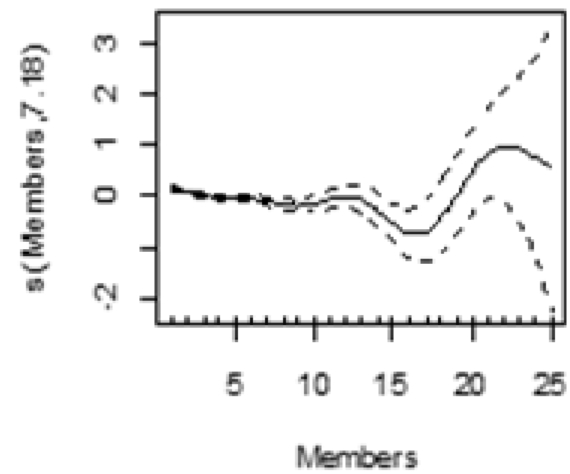



Figure 6: (a) The smooth function household size (b) The smooth function age, (c) The region spatial variation of improved drinking water. P-values for the smooth terms of household members, age and region were 0.00, 0.38 and <0.00 respectively. See Appendix C for the names of the regions.

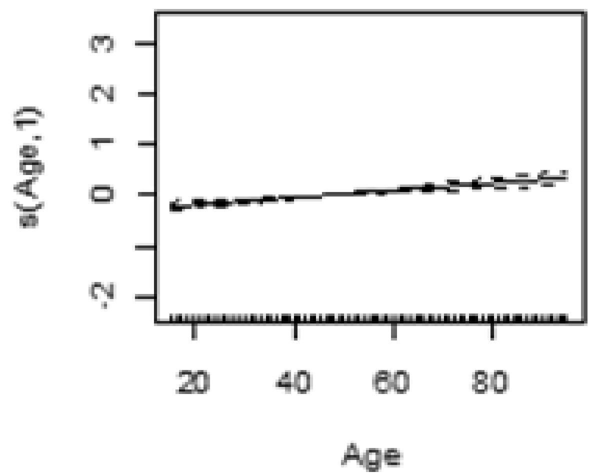
Furthermore, from Table 10, improved sanitation increased with wealth index and it was associated with occupation and working status. Marital status, literacy, sex and exposure to the media show no significant to improved sanitation. Poor sanitation was observed in rural areas. The smooth function estimate for improved sanitation are shown in Figure 7. The intervals of the smooth for household members and age ranged from -2 to 3, while for region, the range was between -0.6 to 0.6. This suggests that household size and age are good predictors of improved sanitation. The p-value showed that region was an important determinant of improved sanitation. It has been observed that the probability of improved sanitation decreases with an increase in household members.

High region variation was captured in !Karas, Oshikoto and Kunene regions, while a low variation was captured in Omusati and Oshana regions as shown in Figure 7(c). Khomas, Hardap, Erongo, Otjozondjupa, Ohangwena, Kavango regions had a moderate spatial variation. See Appendix C for the names of the regions.

(a)




(b)




(c)

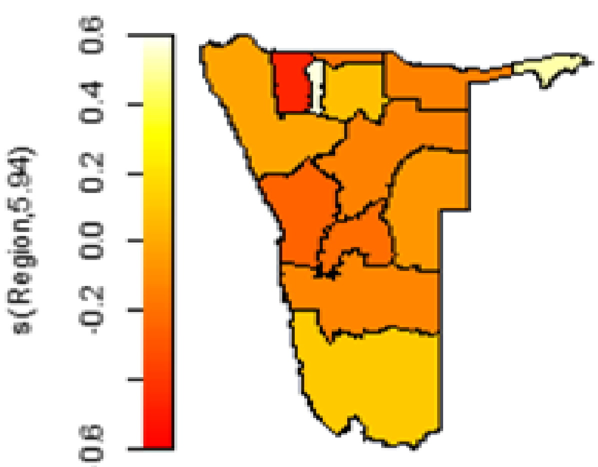



Figure 7: (a) The smooth function household size, (b) The smooth function age, (c) The region spatial variation of improved sanitation. P-values for the smooth terms of household members, age and region were 0.00, 0.00 and <0.00 respectively. See Appendix C for the names of the regions.

The estimated correlations and confidence Intervals (CIs) for model fitted with spatial, fixed and non-linear effects using equation (25) are reported in Table 11. The results shows that the indicators depend on each other because of the positive correlation between WASH indicators as shown in Table 11. It also indicates an increasing relationship between improved drinking water and sanitation.

Table 11: Theta (θ) results from model fitted with spatial, fixed and non-linear effects

| Theta | Estimated correlation, with CI |
|---|---|
| Improved hygiene and drinking water | 0.21 (0.163, 0.256) |
| Improved hygiene and sanitation | 0.217 (0.173, 0.251) |
| Improved drinking water and sanitation | 0.229 (0.173, 0.286) |

### 5.4 Conclusions

Joint analysis of WASH indicators is of importance because these indicators depend on each other. Unsafe drinking water is additionally confounded by poor sanitation, which can make water become contaminated and incredibly increases the need for hygiene. In addition, improved drinking water has no meaning on health if improved sanitation is not concurrent. This study has shown that improved drinking water, sanitation and hygiene are of importance in improving the health and livelihood of the Namibian population.

# CHAPTER 6: GENERAL DISCUSSION, CONCLUSIONS AND RECOMMENDATIONS

## 6.1 Discussion

This section summarizes the study main findings. Despite continued national and international efforts, access to improved drinking water, sanitation and hygiene remains a concern in many developing countries. Lack of improved sanitation facilities, poor knowledge and low practice of basic hygiene or proper environmental sanitation behaviours combined with poor health are manifested in increasingly high levels of diarrhoea related mortality and morbidity. Diarrhoea is the third most common cause of hospital attendance and the second highest cause of paediatric admissions globally (UNICEF, 2010).

According to the Namibia Sanitation and Hygiene Program, nearly 1.3 million of Namibia's population of just over 2 million do not have access to proper toilet facilities, including 84 percent of all people living in the rural areas. In July 2016, the then Minister of Urban and Rural Development in Namibia, Mrs Sophia Shaningwa declared to eradicate the use of the bucket system in Namibia. This programme, started in the !Karas and Hardap regions, ran from August 2016 to March 2018 by the then Ministry of Urban and Rural Development. Statistically, the quest is to improve sanitation from the current 34 percent to 70 percent of the population. This task is not going to be an easy one for the government. Nevertheless, the government is doing all it can to achieve its objective of improving access to proper sanitation in Namibia.

In guaranteeing access to improved drinking water, the government has embarked on various projects such as the construction of the Neckartal Dam in !Karas region, construction of the Divudu pipeline in the Zambezi region and execution of projects such as the expansion of Ondangwa-Omutele pipeline, Katima Mulilo-Kongola pipeline and water treatment plants. The contamination of water supplies and poor hand-washing practices following toilet usage is estimated to be around 75-80 percent (UNICEF, 2010). Namibia's execution in such a manner is disturbing by any standard.

The study aimed at applying Gaussian copula to model improved drinking water, sanitation and hygiene to identify spatial variations at the regional level in Namibia. Specific objectives of the study included estimating the spatial distribution of WASH indicators; model the joint spatial dependency of WASH indicators and quantifying the determinants of WASH indicators in Namibia.  The subsections below covers the discussion of results found using flexible approaches in the analysis of improved drinking water, sanitation and hygiene.

### 6.1.1 Results of spatial distribution of WASH indicators using copula on geostatistical data

The distribution of percentages of households with improved drinking water, sanitation and hygiene, odds estimates from the GCMR models fitted with fixed effects and spatial effects, as well as predicted and prediction standard errors maps produced were the methods used in the study. The percentages of households with improved drinking water did not really provide an exact coverage of improved drinking water, sanitation and hygiene at the regional level in Namibia, hence predicted smooth map was produced. With regards to households with improved drinking water, the results of this study indicated that Kavango, Kunene, Omusati and Oshana regions had the lowest percentages of households with improved drinking water. Kunene, Omusati, Ohangwena, Kavango and Zambezi regions had households with the low improved sanitation. A similar pattern was observed for improved hygiene, with Kavango, Ohangwena and Omusati regions having a low improved hygiene too.

As for the distribution of WASH indicators, it is important to understand the relationship between the indicators and covariates. This relation is essential to enable government to design effective policies and tools to tackle the problem. As known, WASH indicators are associated with socio-economic, demographic and geographic factors. The model fitted with fixed effects shows that sex, wealth index, primary and secondary education, marital status, household size, residence, literacy, reading newspaper/magazine, listening to the radio once a week, watching television, occupation and working status were significant to improved drinking water, as shown in Table 5. This is quite different from the model fitted with an additional covariate region, as only sex, poor households, no education, household size, age, able to read a whole language, no card with required language, blind/visually impaired and watching television once a week were significant to improved drinking water.

The results further showed that household size and watching television at least once a week were not significant to improved sanitation, in the model fitted with fixed effects

(Table 7), while in the model fitted with an additional covariate region, only wealth index, education level, never married, married, age and blind/visually impaired were significant to improved hygiene (Table 9). Moreover, covariates sex, wealth index, education level, marital status, household size, age, residence, literacy, high skilled and working status were significant to improved hygiene in the model fitted with an additional covariate region (Table 9). With regard to covariates becoming non-significant in the model fitted with an additional covariate region (Table 9), this confirmed the importance of accounting for spatial correlation when analyzing geographic data to avoid over-estimation of the standards errors of model covariates (Raso et al., 2012).

Studies by Adam, Boateng, and Amoyaw (2015), Koskei, Koskei, Koske, and Koech (2013) and Rahut, Behera, and Ali (2015), pointed out that improved drinking water was significant to education level and occupation, which is similar to this study, while gender was not significant to improved drinking water, which is different for this study. Existing literature suggests a positive relationship between the education level of household head and the probability of having improved drinking water (Rahut et al., 2015). In addition, age was not significant in Koskei et al. (2013) study, which is a similar case for this study. Moreover, Koskei et al. (2013) results indicate a decrease in improved drinking water and sanitation in female-headed households, which is the case for this study. It is noted that the literature that has focused on the link between poverty and household headship suggests that female-headed households have limited access to resources (Koskei et al., 2013).

Furthermore, a study by Sarmento (2015) shows that residence and wealth index is significant to improved drinking water, which is the case for this study. The odds of improved drinking water increased in urban areas. This pattern is consistent with the previous study done in Indonesia by Irianti, Prasetyoputra, and Sasimartoyo (2016). However, literacy was not significant in Sarmento (2015) study, but for this study, it was significant.

The findings from this study showed that the odds of improved drinking water decreased if the respondent is working, which is unexpected. In general, it is expected that the households where the respondent is working should have improved drinking water increases.

Koskei et al. (2013) result further show that marital status was significant to improved sanitation, while household size, gender, occupation and education level were not significant to improved sanitation. This is only similar to this study with covariate household size. Koskei et al. (2013) findings are however different from the study by Adam et al. (2015) whose study highlighted the importance of gender and education level to improved sanitation, which is consistent with this study. Adam et al. (2015) pointed out that, households with educated household heads are more likely to have better sanitation facilities compared to households with less educated household heads.

This study indicated that poor and middle income households were less likely to have access to improved sanitation compared to the rich households, thus the odds of having improved sanitation increased from poor to rich households. This could be because rich households are able to connect and pay for basic needs (Adam et al., 2015 and Mulenga, Bwalya, & Kaliba- Chishimba, 2017). WHO & UNICEF (2015) noted that there is a strong relationship between wealth and use of improved drinking water and sanitation. As expected, the odds of improved sanitation increase when the respondent is working. The results also show that literacy plays an important role in having improved drinking water and sanitation. As anticipated the odds of improved drinking water is higher in households where the woman is able to read whole language.

The rural households showed lower odds of having improved drinking water, sanitation and hygiene. In terms of improved drinking water, the results were consistent with Nketiah-Amponsah et al. (2009), Adams et al. (2015) and Mulenga et al. (2017). According to Mulenga et al. (2017), having improved drinking water is an urban phenomenon. In addition, globally it has been observed that rural households are less likely to have improved drinking water, compared to urban areas. Mulenga et al. (2017) pointed out that this could be because rural households are not able to pay for improved drinking water and water supply corporations are mostly found in urban areas.

The results further showed higher odds of improved drinking water and sanitation in male-headed households. These results are not as expected because, in the African culture, females are responsible for fetching water and using clean water for households purposes including sanitation (Adams et al., 2015). Hence, these results were not consistent with studies by Adams et al. (2015), Irianti et al. (2016) and Mulenga et al. (2017). In terms of improved hygiene, the odds of improved hygiene is higher in female-headed households.

Furthermore, the results showed that household heads with higher education were more likely to have improved drinking water, sanitation and hygiene than household heads with primary and secondary education. These results were consistent with the study by Adam et al. (2015), who pointed out that household heads with higher education may have more knowledge about the health risks associated with using unsafe drinking water and poor sanitation facilities. Another explanation for this pattern is that educated household heads might be more disposed towards influencing the fundamental speculations for their families to have better access to household amenities. Moreover, household heads with higher education may be well employed, acquire more salary and are able to afford improved drinking water, sanitation and handwashing detergents.

Mapping the geographic distribution of the observed number of households with improved drinking water, sanitation and hygiene were not sufficient, as the variability in data was not accounted for. Therefore, the model estimates often used in spatial variation was used to overcome this.

Results from this study predicted that Kavango, Ohangwena, Omusati and Kunene regions will have low improved drinking water. In addition, the study predicted low improved sanitation in Zambezi, Kavango, Oshikoto, Ohangwena, Oshana, Omusati and Kunene regions. The Namibia government initiative to eradicate the use of the bucket system has started in !Karas and Hardap regions. However from the study results, these regions were not much of a concern compared to the others. This study would have helped the government to start the program in regions that have low improved sanitation.

Furthermore, the results predicted that Zambezi, Kavango, Oshikoto, Ohangwena, Oshana, Omusati, Kunene, !Karas, Erongo and Hardap regions will have low improved hygiene.

The results from the study on the spatial distribution of households with low improved sanitation were consistent with what has been reported in other reports. For instance, the study predicted low improved sanitation in Kunene, Omusati, Ohangwena, Oshana, Zambezi and Kavango regions that is similar with the results from UNICEF (2010), who found low coverage of improved sanitation in Ohangwena (5%), Zambezi (9%), Kavango (12%), Omusati (15%) and Kunene (19%) (UNICEF, 2010).

Moreover, this study found geographical inequalities between regions whereby households in some regions were more likely to have access to improved drinking water, proper sanitation and good hygiene, which is consistent with preceding studies by Adam et al. (2015), Irianti et al. (2016) and Mulenga et al. (2017).

### 6.1.2 Results of simultaneous modelling of water, sanitation and hygiene using generalized joint regression model

The study used a joint model to account for dependency between WASH indicators. Results from the joint model fitted with spatial effects, fixed effects and non-linear effects showed that residence was significant to improved drinking water and hygiene, while female-headed households show a negative association with improved drinking water. Occupation and working status is significant to improved hygiene, with skilled and low skilled negatively associated. The results also showed that widowed and divorced negatively associated with improved hygiene and widowed was negatively associated with improved drinking water. In addition, middle and rich household showed a positive association with improved drinking water, sanitation and hygiene.

In terms of spatial variation from the joint model with spatial effects, !Karas and Oshana regions showed a low improved hygiene, Erongo region had the lowest improved drinking water and !Karas, Zambezi & Oshana regions had the lowest improved sanitation. The joint model results for improved drinking water were quite different from the GCMR model estimates, as Erongo region was among the regions with high improved drinking water and it had also shown a moderate percentage of households with improved sanitation in Figure 1(b). For improved hygiene, the results were a reflection of some of the regions with low improved hygiene. In addition, for improved sanitation, Zambezi, Ohangwena and Omusati regions were among the regions with low improved sanitation.

The joint model show that residence and wealth index were important predictors of improved drinking water, sanitation and hygiene. Moreover, the results show that WASH indicators depend on each other, with improved drinking water and sanitation highly correlated. This is similar to the findings by Nketiah- Amponsat, Aidam, and Senadza (2009). Furthermore, the AIC results showed that the joint model with non-linear effects is the best model amongst the other joint model fitted.

## 6.2 Limitations of the study

The limitation of this study was using secondary data, as the researcher had no control over the variables in the dataset. Another limitation was that the data might not be of quality representation because the data was collected with a different purpose. Even though the geostatistical data used for the analysis were from 550 locations, they were sparsely distributed in the south, west, northeast and north-west of Namibia. This could be because the regions are vast and this can cause bias in estimates. It was then important to be vigilant when interpreting the predicted estimates, with the map of prediction errors as in Figures 2(b), 3(b) and 4(b). For example, the model predicted the west to have higher improved sanitation with no survey points, which may not be true on the ground. Regions such as Omusati, Ohangwena, Oshana and Zambezi regions were rich in survey points and this could be because the regions are narrow. The study was limited to explore other forms of copula in the programming part, as the semiParTRIV function that was used to join the WASH indicators in R (R Core Team, 2016) does not allow for several choices for copula, apart from the Gaussian copula that was used.

Regardless of these limitations, the predicted maps of improved drinking water, sanitation and hygiene give helpful information that can be utilized to identify areas that need significant intervention.

## 6.3 Conclusions and recommendations

In terms of spatial modelling of WASH indicators in Namibia, the availability of limited literature for the study suggests that this study is among the few that attempted to estimate the spatial distribution of WASH indicators at region level in Namibia.

More generally, this work has highlighted regions that are struggling with improved drinking water, sanitation and hygiene. It is the responsibility of the government and policy makers to provide these regions with assistance in the development of strategies and investment plans to reduce inequality and marginalization. The result of this study shows that government has done a lot to provide improved drinking water to its citizens, but still some areas, especially, in the northern part of Namibia still need intervention. The study has identified that the need for improved sanitation and hygiene is still high in Namibia. It is therefore hoped that this study will assist policy makers and those involved in budget execution to allocate enough resources to affected regions that have low improved drinking water, sanitation and hygiene. However, resources should not only be provided in terms of cost for infrastructure, but also in the aspect of sustainability.

This study further suggests that the NDHS data can be particularly useful in identifying geographic variation in improved drinking water, sanitation and hygiene used at different geographic level. Therefore, there is a need to extend this analysis further to examine trends over time and examine the spatial distribution of different levels, as opposed to just the “haves” and “have-nots”. Based on the findings of this study, it is recommended that the formulation of policies geared towards the provision of clean drinking water, safe sanitation and proper hygiene, which will improve public health and save lives. In addition, the study recommends further model analysis beyond the mean, for example, modelling of the variability of the models such as the variance, skewness and media.

# REFERENCES


Adams, E. A., Boateng, G. O., & Amoyaw, J. A. (2015). Socioeconomic and demographic predictors of potable water and sanitation access in Ghana. *Social Indicators Research*, 126(2), 673-687. doi:10.1007/s11205-015-0912-y.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov, & F. Csaki (Eds), *Second international symposium on information theory* (pp. 267-281). Budapest: Academian Kiado.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19, 716-723.

Ayele, D. G., Zewotir, T. T., & Mwambi, H. G. (2013). Spatial distribution of malaria problem in three region of Ethiopia. *Malarial Journal*, 12(207),1-14.doi:10.186/1475-2875-12-201.

Baek, S. (2010). A Copula-based method for analyzing bivariate binary longitudinal data .A dissertation in Epidemiology and Biostatistics. University of Pennsylvania. Retrieved from http://repository.upenn.edu/edissertations/1564.

Bárdossy, A. (2006). Copula-based geostatistical models for groundwater quality parameters. *Water Resources Research*, 42(11),1-12, doi:10.1029/2005wr004754.

Bartram, J., & Cairncross, S. (2010). Hygiene, Sanitation, and Water: Forgotten foundations of health. *PLoS Med*, 7(11), 1-9. doi: 10.1371/journal.pmed.1000367.

Besag, J. (1974). Spatial interaction and the statistical of Lattice system. *Journal of the Royal Statistical Society*, 36(2), 192-236.

Besag, J. (1975). Statistical analysis of non-lattice data. *Journal of the Royal Statistical Society*, 24(3), 179-195.

Besag, J., York, J., & Mollie, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*, 43(1), 1-20. doi:10.1007/BF00116466.

Bhat, C. R., & Sener, I. N. (2009). A Copula- based close form binary logit choice model for accommodating spatial correlation accross observational unit. *Journal of Geographical System*, 11(3), 243-272.

Bhat, C. R., Sener, I. N., & Eluru, N. (2010). A flexible spatially dependent discrete choice model: Formulation and application to teenagers' weekday recreational activity participation. *Transportation Research*, 44(8-9), 903-921.

Blanchet, J., & Davison, A. C. (2011). Spatial modeling of extreme snow depth. *The Annals of Applied Statistics*, 5(3), 1699-1725.

Bohdalova, M., & Nanasiova, O. (2006). *Note to Copula Fuctions.* Retrieved from www.g-casa.co/download/Bohdalova_Copula_fuctions.pdf.

Cavanaugh, J. E., & Neath, A. A. (1999). Generalizing the derivation of the Schwarz information criterion. *Communication in Statistics Theory and Methods*, 28, 49-66.

Cliff, A. D., & Ord, J. K. (1973). *Spatial Autocorrelation*. London: Pion Limited.

Cohen, B. (2006). Urbanisation in Developing countries: Current trends, future projections and key challenges for sustainability. *Technology in Society*, 28, 63-80.

Costello, D. H. (2013). *An evaluation of a Water, Sanitation, and Hygiene (WASH) program for rural communities in Northern Afghanistan.* Retrieved from https://ir.library.oregonstate.edu/xmlui/bitstream/handle/.../CostelloDeniseH2013 .pdf?. Oregon State University.

Cressie, N. (2015). *Statistics for spatial data* (Rev. ed.). New York: Wiley.

Cui, W., & George, E. I. (2008). Empirical Bayes vs. Fully Bayes variable selection. *Journal of Statistical Planning and Inference*, 138(4), 888-900.

Danaher, P. J., & Smith, M. S. (2011). Modelling multivariate distribution using Copulas. *An Application in Marketing Science*, 30(1), 4-21.

De Oliveira, V. (2012). Bayesian analysis of conditional autoregressive models. *Annals of the Institute of Statistical Mathematics*, 64(1), 107-133. doi: 10.1007/s10463-010-0298-1.

Diggle, P. J., & Ribeiro, P. J. (2006). *Model-Based Geostatistics.* Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1111/j.15410420.2008.01026_3.x&rep=rep1&type=pdf.

DiMaggio, C. (2014). Spatial epidemiology notes: Applications and vignettes in R. Columbia University Press. New York.

Earnest, A., Morgan, G., Mengersen, K., Ryan, L., Summerhayes, R., & Beard, J. (2007). Evaluating the effect of neighbourhood weight matrices on smoothing properties of Conditional Autoregressive (CAR) models. *International Journal of Health Geographics*, 6(65), 1-12. doi:10.1186/1476-072X-6-54.

Eid, U. (2015). *The importance of Water, Sanitation and Hygiene as key to national development.* Retrieved from Water Magazine: http://water.jhu.edu/index.php/magazine/climate-change-and-health-why-the-link-to-water-is-critical/.

Eilers, P. H., & Marx, B. D. (1996). Flexible smoothing with b-splines and penalties . *Statistical Science*, 11(2), 89-102.

Embrechts, P., Lindskog, F., & McNeil, A. (2001). Modelling dependence with Copulas and applications to risk management. Retrieved from https://people.math.ethz.ch/~embrecht/ftp/copchapter.pdf.

Ezeh, O. K., Agho, K. E., Dibley, M. J., Hall, J., & Page, N. P. A. (2014). The impact of water and sanitation on childhood mortality in Nigeria: Evidence from Demographic and Health Surveys, 2003–2013. *International Journal of Environmental Research and Public Health*, 11(9), 9256-9272. doi: 10.3390/ijerph110909256.

Fermanian, J.-D. (2017). Recent development in copula models. *Econometrics*, 5(34), 1-3. doi: 10.3390/econometrics5030034.

Filippou, P., Marra, G., & Radice, R. (2017). Penalized likelihood estimation of a trivariate additive probit model. *Biostatistics*, 18(3), 569-585. doi:10.1093/biostatistis/kxx008.

Fleming, M. M. (2004). Techniques for estimating spatially dependent discrete choice models. In L. Anselin, L. Florax, & S. Rey, *Advances in spatial econometrics. Methodology, tools and applications* (pp. 145-168). New York: Springer.

Garriga, R. G., & Foguet, A. P. (2013). Unravelling the linkages between water, sanitation,hygiene and rural poverty: The WASH poverty index. *Water Resources Management*, 27(5), 1501-1515. doi: 10.org/1007/s11269-012-0251-6.

Genest, C., & Neslehova, J. (2007). A primer on copulas for count data. *Astin Bulletin*, 37(2):475.

Goovaerts, P. (1999). Geostatistical approaches for incorporating elevation into the spatial interpolation of rainfall. *Journal of Hydrology*, 228(2000), 113-129.

Goren, E., & Hughes, J. (2017). *Fitting the copCAR regression model for discrete areal data*. Retrieved from http://cran.r-project.org/web/packages/copCAR.pdf.

Hoff, P. D. (2007). Extending the rank likelihood for semiparametric Copula estiamate. *The Annals for Applied Statistics*, 1, 265-283. doi: 10.1214/07-AOAS107.

Huard, D., Evin, G., & Favre, A. C. (2006). Bayesian Copula selection. *Computational Statistics and Data Analysis*, 51(2), 809- 822.doi: 10. 1016/j.cascla.2005.08.010.

Hughes, J., & Haran, M. (2013). Dimension reduction and alleviation of confounding for spatial generalized linear mixed models. *Journal of the Royal Statistical Society, Series B*, 75(1), 139-159.

Human Rights Council. Resolution Adopted by the Human Rights Council. 18/1. The human right to safe drinking water and sanitation. Retrieved from: http://www.worldwatercouncil.org/fileadmin/wwc/.

Hughes, J. (2015). copCAR: A flexible regression model for areal data. J*ournal of Computational and Graphical Statistics*, 24(3), 733–755. doi:10.1080/10618600.2014.948178.

Hutton, G., & Chase, C. (2016). The knowledge base for achieving the sustainable development goal targets on water supply, sanitation and hygiene. *International Journal of Environmental Research and Public Health*, 353-399. doi:10. 3390/ijerph13060530.

Irianti, S., Prasetyoputra, P., & Sasimartoyo, T. P. (2016). Determinants of household drinking-water source in Indonesia: An analysis of the 2007 Indonesian family life survey. *Cogent Medicine*, 3(1151143), 1-13. doi.org/10.1080/2331205X.2016.1151143.

Jane, R., Valle, D. L., Simmonds, D., & Raby, A. (2016). A copula-based approach for the estimation of wave height records. *Coastal Engineering*, 117, 1-18. doi.org/10.1016/j.coastaleng.2016.06.008.

Jasper, C., Le, T.-T., & Bartram, L. (2012). Water and sanitation in schools: A systematic review of the health and educational outcomes. *International Journal of Environmental Research and Public Health*, 9, 2272-2787. doi.org/10.3390/ijerph9082772.

Joe, H. (2015). *Dependence Modeling with Copulas.* Boca Raton: Taylor & Francis group.

Kazembe, L. N. (2017). Bivariate Copula based regression to model timing and frequency of antenatal care utilization. Department of Statistics and Population studies. University of Namibia.

Kazianka, H., & Pilz, J. (2009). Spatial Interpolation using Copula based geostatistical models. In P. M. Atkinson, & C. Lloyd, *GeoENV VII Geostatistics for Environmental Application* (pp. 307-320). Berlin: Springer.

Kazianka, H., & Pilz, J. (2010). Bayesian spatial modelling and interpolation using Copula. *Computer Science & Geoscience*, 37(3), 310-319.

Kolev, N., & Paiva, D. (2009). Copula- based regression models: A survey. *Journal of Statistical Planning and Inference*, 139, 3847-3856. doi: 10.1016/j.jspi.2009.05.023.

Kort, J. D. (2007). *Modeling tail dependence using copulas — literature review*. Retrieved from http://ta.twi.tudelft.nl/users/vuik/numanal/kort_scriptie.pdf.

Koskei, E. C., Koskei, R. C., Koske, M. C., & Koech, H. K. (2013). Effect of socio-economic factors on access toiImproved water sources and basic sanitation in Bomet municipality, Kenya. *Research Journal of Environmental and Earth Sciences*, 5(12), 714-719.

Kuethe, T. H., Hubbs, T., & Waldof, B. (2009). Copula models for spatial point patterns and processes. Retrived from: http://www.ub.edu/sea2009.com/Paper/162.pdf.

Kulldorff, M., Heffernan, R., Hartman, J., Assuncao, R., & Mostashari, F. (2005). A space-time permutation scan statistic for disease outbreak detection. *PLOS Medicine*, 2(3), 59. doi.org/10.1371/journal.pmed.0020059.

Lee, D., Rushworth, A., & Sahu, S. K. (2014). A Bayesian localized conditional autoregressive model for estimating the health effects of air pollution. *Biometrics*, 70, 419-429. doi:10.1111/biom.12156.

Liverani, S. (2016). Modelling collinear and spatially correlated data. *Spatial and Spatio-temporal Epidemiology*, 8, 63-73.

Madsen, L. J. (2009). Maximum likelihood estimation of regression parameter with spatially depedent discret data. *Journal of Agriculture, Biological and Environmental Statistics*, 14(4), 375-391. doi:10.1198/jabes.2009.07116.

Magalhaes, R., Barnett, A. G., & Clements, A. (2011). Geographical analysis of the role of water supply and sanitation in the risk of helminth infections of children in West Africa. *PNAS*, 108(50), 20084-20089.

Marra, G., & Radice, R. (2017a). Bivariate Copula additive models for location,scale and shape. *Computational Statistics and Data Analysis*, 122, 99-113.

Marra, G., & Radice, R. (2017b). Generalized Joint Regression Modelling (GJRM). R package version 3.4.1. https://CRAN.R-project.org/package=GJRM.

Marra, G., Radice, R., Bärnighausen, T., Wood, S. N., & McGovern, M. E. (2017). A simultaneous equation approach to estimating hiv prevalence with non-ignorable missing responses. *Journal of the American Statistical Association*, 112(518), 484-496. doi: 10.1080/01621459.2016.1224713.

Masarotto, G., & Varin, C. (2012). Gaussian copula marginal regression . *Electronic Journal of Statistics* , 6(2012), 1517-1549. doi: 10.1214/12-EJS721.

Masarotto, G., & Varin, C. (2017). Gaussian Copula Regression in R. *Journal of Statistical Software* , 8(77), 1-26. doi: 10.18637/jss.v077.i08.

Ministry of Health and Social Social Services (MoHSS) [Namibia] and Macro International Inc. (2014). Namibia Demography and Health Survey 2013. Windhoek, Namibia and Claverton, Maryland: MoHSS and Macro.

Moe, C. L., & Rheingans, R. D. (2006). Global challenges in water, sanitation and health. *Journal of Water and Health*, 41-57.

Monogan, J. (2013). *Areal data models.* Retrieved from http://pdfs.semanticscholar.org/fcc3/3cbd5d134fdd7e9fe4bf603df2c2987eade6.pdf.

Mulenga, J. N., Bwalya, B. B., & Kaliba- Chishimna, J. (2017). Determinants and inequalities in access to improved water sources and sanitation among the Zambian households. *International Journal of Development and Sustainability*, 6(8), 746-762.

Musgrove, D. R., Hughes, J., & Eberly, L. E. (2016). Hierarchical Copula regression models for areal data. *Spatial Statistics*, 17, 38-49. doi: 10.1016/j.spasta.2016.04.006.

National Planning Commission (NPC). (2004). Namibia Vision 2030. Windhoek, Namibia: NPC.

Nelsen, R. B. (1999). *An Introduction to Copula.* New York: Springer.

Nelsen, R. B. (2006). *An Introduction to Copula.* New York: Springer.

Nketiah-Amponsah, E., Aidam, P. W., & Senadza, B. (2009). *Socio-economic determinants of sources of drinking. Paper presented at the Conference on International Research on Food Security, Natural Resource Management and Rural Development, University of Hamburg, Germany.* Retrieved from http://www.tropentag.de/2009/abstracts/full/185.pdf.

Ntozini, R., Marks, S. J., Mangwada, G., Mbuya, M. N., Gerema, G., & Mutasa, B. (2015). Using Geographic Information Systems and spatial analysis methods to assess household water access and sanitation coverage in the SHINE Trial. *Clinical Infectious Diseases*, 61(7), 716-725. doi.org/10.1093/cid/civ847.

Paciorek, C. J. (2013). Spatial models for point and areal data using Markov random fields on a fine grid. *Electronic Journal of Statistics*, 7, 946-972. doi:10.1214/13-EJS791.

Pebesma, E., & Skoien, J. O. (2016). *Package'intamap'*. Retrieved from https://cran.r-project.org/web/packages/intamap/intamap.pdf.

Perrone, E., & Müller, W. G. (2016). Optimal design for Copula models. *Statistics*, 50(4), 917-929.

Pitt, M., Chan, D., & Kohn, R. (2006). Efficient Bayesian inference for Gaussian Copula regression models. *Biometrika*, 93, 537-554. doi: 10.1093/biomet/93.3.537.

Pruss-Ustun, A., Bartram, J., Clasen, T., Colfor Jr, J. M., Cumming, O., & Curtis, V. (2014). Burden of disease from inadequate water, sanitation and hygiene in low- and middle-income settings: a retrospective analysis of data from 145 countries. *Tropical Medicine and International Health*, 19(8), 894-905. doi:10.1111/tmi.12329.

Pullan, R. L., Freeman, M. C., Gething, P. W., & Broker, S. M. (2014). Geographical inequalities in use of improved drinking water supply and sanitation across Sub-Saharan Africa: Mapping and spatial analysis of cross-sectional survey data. *PLoS Med*, 11(4). doi:10.1371/Journal.pmed.1001626.

Qu, X., & Lee, L. (2015). Estimating a spatial autoregressive model with an endogenous spatial weight matrix. *Journal of Econometrics*. 184(2), 209-232. doi.org/10.1016/j.jeconom.2014.08.008.

Radice, R., Marra, G., & Wojtyś, M. (2016). Copula regression spline models for binary outcomes. Statistics and Computing, 26(15), 981-995.

Rahut, D. B., Behera, B., & Ali, A. (2015). Household access to water and choice of treatment methods: Empirical evidence from Bhutan. *Water Resources and Rural Development*, 5, 1-16. doi:10.1016/j.wrr.2014.09.003.

Raso, G., Schur, N., Utzinger, J., Koudou, B. G., Tchicaya, E. S., & Rohner, F. (2012). Mapping malaria risk among children in Cote d'Ivoire using Bayesian geo-statistical models. *Malaria Journal*, 11(160), 1-11. doi.org/10.1186/1475-2875-11-160.

R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rosenberg, M.S., Sokal, R.R., Oden, N.L., & DiDiovann, D. (1999). Spatial autocorrelation of cancer in Western Europe. *European Journal of Epidemiology*, 15(1), 15-22. doi:10.1023/A:1007559728848.

Rue, H., & Martino, S. (2007). Approximate Bayesian inference for hierarchical Gaussian Markov random fields models. *Journal of Statistical Planning and Inference*, 137(10), 3177-3192. doi.org/10.1016/j.jspi.2006.07.016.

Sarmento, D. R. (2015). *An analysis of access to improved drinking water and sanitation and distance to the water source in a newly independent country, Timor-leste: assessing geographical and socioeconomic disparities.* Georgia State University. Retrieved from http://scholarworks.gsu.edu/iph_theses/391/.

Schabenberger, O., & Gotway, C.A. (2005). *Statistical methods for spatial data analysis.* Florida: Chapman & Hall/CRC Press.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6, 461-464.

Silva, R. S., & Lopes, H. F. (2008). Copula, marginal distribution and model selection. *Statistics and Computing*, 18, 313-320.

Sklar, A. (1959). Distribution functions of n dimensions and margins. Publications of the Institute of Statistics. University of Paris, Paris, pp. 229-231.

Smith, M. S. (2011). Bayesian approaches to Copula modelling. In P. Damien, P. Dellaportas, N. Polson, & D. Stephens, *Bayesian Theory and Applications*, (pp. 336–358).

Smith, M. S. (2014). Bayesian approaches to Copula modelling. In P. Damien, P. Delaaportas, N. Polson, & D. Stephens, *Hierarchial Models and MCMC: A tribute to Adrian Smith.*

Smith, M., Gan, Q., & Kohn, R. (2010). Modelling depedence using skew t Copulas: Bayesian inference and application. *Journal of Applied Econometrics( Forthcoming)*. doi: 10.1002/jae.1215.

Song, P. X.K., Li, M., & Yuan, Y. (2009). Joint regression analysis of correlated data using Gaussian Copula. *Biometrics*, 65, 60-69. doi: 10.1111/j.l541-0420.2008.01058.

Sudell, M., Kolamunnage-Dona, R., & Tudur-Smith, C. (2016). Joint models for longitudinal and time-to-event data: A review of reporting quality with a view to meta analysis. *BMC Medical Research Methodology*, 16(168), 1-11. doi: 10.1186/s12874-016-0272-6.

Tiwari, R., & Nayak, S. (2013). Drinking water and sanitation in Uttar Pradesh: A regional analysis. *Journal of Rural Development*, 32, 61–74.

Trivedi, P. K., & Zimmer, D. M. (2005). Copula modeling: An introduction for practitioners . *Foundations and Trends in Econometrics*, 1(1), 1-111. doi: 10.1561/0800000005.

UNICEF. (2010). Retrieved from UNICEF Namibia: https://www.unicef.org/namibia/health_nutrition_13817.html.

UNICEF. (2015). UNICEF 2016-2030 strategy for health. New York.

UNICEF. (2016a, April 10). *Water, Sanitation and Hygiene*. Retrieved from UN: https://www.unicef.org/wash/3942_3952.html.

UNICEF. (2016b). Strategy for Water, Sanitation and Hygiene 2016-2030. New York.

Varin, C. (2008). On composite marginal likelihoods. *AStA advance in Statistical Analysis*, 92(1), 1-28. doi: 10.1007/s10182-008-0060-7.

White, H. (1994). Estimation, Inference and Specification Analysis. Cambridge University Press.

WHO. (2017). *Radical increase in water and sanitation investment require to meet development targets*. Retrieved from www.who.int/news-room/detail.13-04-2017-radical-increase-in-water-and-sanitation-investment-required-to-meet-development-targets

WHO & UNICEF. (2006). Meeting the MDG Water and Sanitation Targets: The rural and urban challenges of the decade. WHO, Geneva and UNICEF, New York.

WHO & UNICEF. (2015). Progress on sanitation and drinking water. Geneva and New York.

Winkelmann, R. (2011). Copula bivariate probit models: With an application to medical expenditures. University of Zurich. Pp.29.

Wood, S. N. (2017). *Mixed GAM computation vehicle with automatic smoothness estimation. R package version 1.8-23.* Retrieved from https://cran.r-project.org/web/packages/mgcv/mgcv.pdf.

Yang, H., Bain, R., Bartram, J., Gundry, S., Pedley, S., & Wright, J. (2013). Water safety and inequality in access to drinking-water between rich and poor households. *Environmental Science and Technology*, 47(3), 1222–1230.

Yu, W., Bain, R., Mansour, S., & Wright, J. A. (2014). A cross-sectional ecological study of spatial scale and geographic inequality in access to drinking-water and sanitation. *International of Equity in Health*, 113-127.

Zhu, L., & Carlin, B. P. (2000). Comparing hierarchical models for spatio- temporally misaligned data using the deviance information criterion. *Statistics in Medicine*, 19(17- 18), 2265-2278.

Zimmer, D. M., & Trivedi, P. K. (2006). Using trivariate Copula to model sample seletion and treatment effects: Application to family health care demand. *Journal Business Economic Statistics*, 24, 63-76.

## ANNEXURE A: R codes

R is a free statistical package and is the software used for the analysis of this thesis. The software can be freely downloaded from the site (http://www.r-project.org/.)

Below are the codes used for the analysis of the study:

**(a) Codes used to create point maps**

```
> library(foreign)

> per<-read.csv("C:\\Users\\Anastasia\\Desktop\\gcmr\\per.csv",header=T)

> library(sp)

> library(rgdal)

> library(raster)

> coordinates(per) <- ~x + y

> str(per)

> bou <-readRDS("C:\\Users\\Anastasia\\Desktop\\NAM_adm1.rds")

 > spplot(per, "W", scales = list(draw = T), cuts = 5, col.regions
=rev(bpy.colors(cutoff.tails= 0.1,alpha = 1)), cex = 1, sp.layout =
list(list("sp.polygons",bou)))

> spplot(per, "S", scales = list(draw = T), cuts = 5, col.regions =
rev(bpy.colors(cutoff.tails = +0.1,alpha = 1)), cex = 1, sp.layout = list(list("sp.polygons",
bou)))

> spplot(per, "H", scales = list(draw = T), cuts = 5, col.regions =
```

rev(bpy.colors(cutoff.tails = 0.1,alpha = 1)), cex = 1, sp.layout = list(list("sp.polygons", bou))

**(b) Codes used to fit Gaussian Copula Marginal Regression (GCMR) models**

```
> library(foreign)
> thesis<-read.csv("C:\\Users\\Anastasia\\Desktop\\gcrm\\thesis.csv", header=T)
> library(gcmr)
```

**Codes used to fit Water as a response variable**

```
> D <- spDists(cbind(thesis$x, thesis$y)) / 1000

>modelW<-gcmr(cbind(Wcases,Wsize-Wcases)~I(FHH/100)+I(Poor/100)+I(Middle/100)+I(NoEdu/100)+I(Prim/100)+I(Sec/100)+I(NeverM/100)+I(Married/100)+I(Widowed/10)+Hsize+AgeHH+Res+I(Literacy1/100)+I(Literacy2/100+)+I(Literacy3/100)+I(Literacy4/100)+I(Newspaper1/100)+I(Newspaper2/100)+I(Radio1/+100)+I(Radio2/100)+I(TV1/100)+I(TV2/100)+I(Skilled/100)+I(Low.Skilled/100)+I(High.skilled/100)+I(WorkSt.Y/100),data=data,marginal=binomial.marg,cormat= matern.cormat(D),seed = 12345)

> summary(modelW)

> exp(cbind(odds=coef(modelW),confint(modelW)))

>modelWReg<-gcmr(cbind(Wcases,Wsize-Wcases)~I(FHH/100)+I(Poor/100)+I(Middle/+100)+I(NoEdu/100)+I(Prim/100)+I(Sec/100)+I(NeverM/100)+I(Married/100)+I(Widowed/100)+Hsize+AgeHH+Res+I(Literacy1/100)+I(Literacy2/100)+I(Literacy3/100)+I(Literacy4/100)+I(Newspaper1/100)+I(News
```

paper2/100)+I(Radio1/100)+I(Radio2/100)+I(TV1/100)+I(TV2/100)+I(Skilled/100)+I(Low.Skilled/100)+I(High.skilled/100)+I(WorkSt.Y/100)+as.factor(Region),data=data,marginal = binomial.marg, cormat = matern.cormat(D), seed = 12345)

```
> summary(modelWReg)
> exp(cbind(odds=coef(modelWReg),confint(modelWReg)))
> AIC(modelW,modelWReg)
> par(mfrow = c(2, 2))
> plot(modelW)
> fitted(modelW)
```

**Codes used to fit Sanitation as a response variable**

```
>modelS<-gcmr(cbind(Scases,Ssize-
Scases)~I(FHH/100)+I(Poor/100)+I(Middle/100)+I(NoEdu/100)+I(Prim/100)+I(Sec/10
0)+I(NeverM/100)+I(Married/100)+I(Widowed/100)+Hsize+AgeHH+Res+I(Literacy1/
100)+I(Literacy2/100)+I(Literacy3/100)+I(Literacy4/100)+I(Newspaper1/100)+I(News
paper2/100)+I(Radio1/100)+I(Radio2/100)+I(TV1/100)+I(TV2/100)+I(Skilled/100)+I(
Low.Skilled/100)+I(High.skilled/100)+I(WorkSt.Y/100),data=thesis,marginal=
binomial.marg, cormat = matern.cormat(D),
seed = 12345)
> summary(modelS)
> exp(cbind(odds=coef(modelS),confint(modelS)))
```

```
>modelSReg<-gcmr(cbind(Scases,Ssize-
Scases)~I(FHH/100)+I(Poor/100)+I(Middle/100)+I(NoEdu/100)+I(Prim/100)+I(Sec/10
0)+I(NeverM/100)+I(Married/100)+I(Widowed/100)+Hsize+AgeHH+Res+I(Literacy1/
100)+I(Literacy2/100)+I(Literacy3/100)+I(Literacy4/100)+I(Newspaper1/100)+I(News
paper2/100)+I(Radio1/100)+I(Radio2/100)+I(TV1/100)+I(TV2/100)+I(Skilled/100)+I(
Low.Skilled/100)+I(High.skilled/100)+I(WorkSt.Y/100)+as.factor(Region),data=thesis,
marginal = binomial.marg, cormat = matern.cormat(D),seed = 12345)
> summary(modelSReg)
> exp(cbind(odds=coef(modelSReg),confint(modelSReg)))
> AIC(modelS,modelRReg)
> par(mfrow = c(2, 2))
> plot(modelS)
> fitted(modelS)
```

**Codes used to fit Hygiene as a response variable**

```
>modelH<-gcmr(cbind(Hcases,Hsize.1-
Hcases)~I(FHH/100)+I(Poor/100)+I(Middle/100)++I(NoEdu/100)+I(Prim/100)+I(Sec/1
00)+I(NeverM/100)+I(Married/100)+I(Widowed/100)+Hsize+AgeHH+Res+I(Literacy1
/100)+I(Literacy2/100)+I(Literacy3/100)+I(Literacy4/100)+I(Newspaper1/100)+I(News
paper2/100)+I(Radio1/100)+I(Radio2/100)+I(TV1/100)+I(TV2/100)+I(Skilled/100)+I(
Low.Skilled/100)+I(High.skilled/100)+I(WorkSt.Y/100),data=thesis,marginal =
binomial.marg, cormat = matern.cormat(D),
```

```
seed = 12345)
> summary(modelH)
> exp(cbind(odds=coef(modelH),confint(modelH)))
>modelHReg<gcmr(cbind(Hcases,Hsize.1Hcases)~I(FHH/100)+I(Poor/100)+I(Middle/1
00)+I(NoEdu/100)+I(Prim/100)+I(Sec/100)+I(NeverM/100)+I(Married/100)+I(Widowe
d/100)+Hsize+AgeHH+Res+I(Literacy1/100)+I(Literacy2/100)+I(Literacy3/100)+
I(Literacy4/100)+I(Newspaper1/100)+I(Newspaper2/100)+I(Radio1/100)+I(Radio2/100
)+I(TV1/100)+I(TV2/100)+I(Skilled/100)+I(Low.Skilled/100)+I(High.skilled/100)+I(W
orkSt.Y100)+as.factor(Region),data=thesis,marginal = binomial.marg, cormat =
matern.cormat(D),seed = 12345)
> summary(modelHReg)
> exp(cbind(odds=coef(modelHReg),confint(modelHReg)))
> AIC(modelH,modelHReg)
> par(mfrow = c(2, 2))
> plot(modelHReg)
> fitted(modelHReg)
```

**(c) Codes use for Generalized Joint Regression Modelling (GJRM)**

```
> library(GJRM)

> library(sp)

> library(foreign)

> stacy<-read.spss(("C:\\Users\\Anastasia\\Documents\\THESIS\\Second
Draft\\DATA\\THESIS.sav",to.data.frame=T,use.value.labels=T)

> stacy<-as.data.frame(stacy)

> summary(stacy)

> obj <- readRDS("C:/aspatialR/NAM_adm1.rds")

> pol <- polys.setup(obj)

> nam.polys <- pol$polys

> name <- cbind(names(nam.polys), pol$names1)

>name

>  xt <- list(polys = nam.polys)
```

**Model Fitting**

```
>f.1<- list(Hygiene ~
as.factor(Residence)+as.factor(Sex)+as.factor(Marital)+as.factor(Wealth)+as.factor(Lite
racy)+as.factor(Newspaper)+as.factor(Radio)+as.factor(TV)+as.factor
(RespOcc)+as.factor(WorkStatus)+ s(Members)+s(Age)+s(Region,bs="mrf",xt=xt,
k=7), Water ~ as.factor(Residence)+as.factor(Sex)+as.factor(Marital)+as.factor
```

```
(Wealth)+as.factor(Literacy)+as.factor(Newspaper)+as.factor(Radio)+as.factor(TV)+as.
factor(RespOcc)+as.factor(WorkStatus)+s(Members)+s(Age)+s(Region, bs="mrf",xt=xt,
k=7),Sanitation ~ as.factor(Residence)+as.factor(Sex)+as.factor(Marital)+as.factor
(Wealth)+as.factor(Literacy)+as.factor(Newspaper)+as.factor(Radio)+as.factor(TV)+as.
factor(RespOcc)+as.factor(WorkStatus)+s(Members)+s(Age)+s(Region, bs="mrf",xt=xt,
k=7))
> wash1  <- SemiParTRIV(f.1, data = stacy)
> conv.check(wash1)
> set.seed(1)
> sb<-summary(wash1)
> sb
> plot (wash1,eq=1)
> plot (wash1,eq=2)
> plot (wash1,eq=2)
```

# APPENDIX A: Ethical Clearance Certificate

UNAM
UNIVERSITY OF NAMIBIA

ETHICAL CLEARANCE CERTIFICATE

**Ethical Clearance Reference Number:** FOS /350/2017 **Date:** 20 November, 2017

This Ethical Clearance Certificate is issued by the University of Namibia Research Ethics Committee (UREC) in accordance with the University of Namibia's Research Ethics Policy and Guidelines. Ethical approval is given in respect of undertakings contained in the Research Project outlined below. This Certificate is issued on the recommendations of the ethical evaluation done by the Faculty/Centre/Campus Research & Publications Committee sitting with the Postgraduate Studies Committee.

**Title of Project:** Flexible Spatial Modelling Using Copula With Application To Water, Sanitation & Hygiene (Wash) Indicators In Namibia

**Researcher:** Anastasia Johannes

**Student Number:** 200618130

**Supervisor(s)** Prof Lawrence Kazembe

Take note of the following:

(a) Any significant changes in the conditions or undertakings outlined in the approved Proposal must be communicated to the UREC. An application to make amendments may be necessary.
(b) Any breaches of ethical undertakings or practices that have an impact on ethical conduct of the research must be reported to the UREC.
(c) The Principal Researcher must report issues of ethical compliance to the UREC (through the Chairperson of the Faculty/Centre/Campus Research & Publications Committee) at the end of the Project or as may be requested by UREC.
(d) The UREC retains the right to:
(i) Withdraw or amend this Ethical Clearance if any unethical practices (as outlined in the Research Ethics Policy) have been detected or suspected,
(ii) Request for an ethical compliance report at any point during the course of the research.

UREC wishes you the best in your research.

Prof. P. Odonkor: UREC Chairperson

Ms. P. Claassen: UREC Secretary

## APPENDIX B: Permission letter from MoHSS

**REPUBLIC OF NAMIBIA**

*Ministry of Health and Social Services*

| | | |
|---|---|---|
| Private Bag 13198 | Ministerial Building | Tel: 061 – 2032150 |
| Windhoek | Harvey Street | Fax: 061 – 222558 |
| Namibia | Windhoek | Email: shimenghipangelwa71@gmail.com |

**OFFICE OF THE PERMANENT SECRETARY**

**Ref:** 17/3/3 **AJ**
**Enquiries:** Mr. J. Nghipangelwa

**Date:** 07 September 2017

**Ms. Anastasia Johannes**
**P.O. Box 55008**
**Rocky Crest, Windhoek**
**Namibia**

Dear Ms. Johannes

***Re: Flexible spatial modelling using copula with application to water, sanitation and hygiene (WASH) indicators in Namibia.***

1. Reference is made to your application to conduct the above-mentioned study.

2. The proposal has been evaluated and found to have merit.

3. **Kindly be informed that permission to conduct the study has been granted under the following conditions:**

3.1 The data to be collected must only be used for academic purposes;

3.2 No other data should be collected other than the data stated in the proposal;

3.3 Stipulated ethical considerations in the protocol related to the protection of Human Subjects' should be observed and adhered to, any violation thereof will lead to termination of the study at any stage;

3.4 A quarterly report to be submitted to the Ministry's Research Unit;

3.5 Preliminary findings to be submitted upon completion of the study;

3.6 Final report to be submitted upon completion of the study;

3.7 Separate permission should be sought from the Ministry of Health and Social Services for the publication of the findings.

Yours sincerely,

PERMANENT SECRETARY
MINISTRY OF HEALTH & SOCIAL SERVICES

...........................................................

**Andreas Mwoombola (Dr.)**
**Permanent Secretary**

*"Health for All"*

## APPENDIX C: Regional boundary map of Namibia