Modelling spatial non-stationarity of *Oreochromis karongae* in South-East Arm of Lake Malawi

Patrick Jeremy Likongwe

A thesis submitted to the Department of Horticulture in the Faculty of Agriculture in partial fulfillment of the requirements for the award of degree of Master of Science in Research Methods of the Jomo Kenyatta University of Agriculture and Technology

2013

DECLARATION

This thesis is my original work and has not been presented for a degree in any other University.

Signature: ..... Date: .....

Patrick Jeremy Likongwe

This thesis has been submitted for examination with our approval as University Supervisors.

Signature: ..... Date: .....

J. M. KIHORO (PhD)

Jomo Kenyatta University of Agriculture and Technology (JKUAT), KENYA

Signature: ..... Date: .....

Murimi NGIGI (PhD)

Jomo Kenyatta University of Agriculture and Technology (JKUAT), KENYA

Signature: ..... Date: .....

Daniel JAMU (PhD)

WorldFish, Malawi

DEDICATION

To my parents Dr J.S. Likongwe, my mentor and Mrs J. J. Likongwe, my hero.

To Rethabile Coltilda Likongwe, my daughter – be inspired.

ACKNOWLEDGEMENT

I am greatly indebted to Dr D. Jamu and WorldFish, Zomba, Malawi for the guidance and support during the process of developing the proposal and writing the thesis, to Dr J. Kihoro and Dr M. Ngingi for the untiring supervisory role in the whole process from proposal to thesis writing.

I also thank RUFORUM for providing the scholarship to fund the research.

To Mr G. Kanyerere from Fisheries Research Unit, Department of Fisheries, Government of Malawi, Monkey Bay - Thank you so much for providing the data on fishery surveys from where Chambo data was based and for the guidance on how the surveys were designed.

To my lovely wife Patricia Makuzana Mphundi, who missed me most in the first seven months of Rethabile's development on mother earth while I was studying and for her encouragement and patience and to my brothers and sisters for being there in one way or the other, may God keep blessing you always.

To my classmates and all friends, Chester Kalinda, Gabriel Otieno, I say thank you for your continued support, understanding and encouragement, keep the network alive and may God keep blessing you always.

To God almighty, thank you for the gift of life and endless love.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

ABBREVIATIONS

AIC Alkaike Information Criteria

AICc corrected Akaike Information Criteria

BFC Brudson, Fotheringham and Charlton F test

BIC Bayesian Information Criterion

BLUE Best Linear Unbiased Estimator

CRSP Chambo Restoration Strategic Plan

CV Cross Validation

DoF Department of Fisheries

FRU Fisheries Research Unit

GAMs Generalized Additive Models

GLM Generalized Linear Model

GLR Global Logistic Regression

GNP Gross National Product

GoM Government of Malawi

GRASP Generalized Regression Analysis and Spatial Prediction

GWR Geographically Weighted Regression

LMZ Leung, Mei and Zhang F3 test

LRM Linear Regression Model

OLS Ordinary Least Squares

SAR Simultaneous Autoregressive Regression

SEA South East Arm

SEBLUP Spatial Empirical Best Linear Unbiased Predictor

SWA South West Arm

TL Total Length

ABSTRACT

Fisheries researchers are faced with the challenging task of studying complex patterns and processes in aquatic resources. Analysis of such patterns is mostly performed under the assumption that ecological relationships do not vary within management areas (i.e. assuming spatially stationary processes). This assumption was questioned by studying the distribution of a target fish population, (Oreochromis karongae) commonly known as Chambo from South East Arm (SEA) of Lake Malawi where it is in abundance. Presence/absence of Chambo is not equally distributed as factors like depth or distance from shore line to where fish is caught are not the same in the whole of SEA. Survey data from June 1999 and October 2007 were used firstly in the comparison of fish abundance and only data from 2007 was used in the modeling of spatial distribution. Global logistic regression (GLR), generalized additive logistic model (GAM) (both global) and geographically weighted logistic regression (GWR), a local modeling technique, were run to explore the best model that can explain spatial non-stationarity and how they affect fish distribution. Akaike Information Criterion (AIC) was used for best model selection depending on the lowest possible deviance value by comparison with other models. The best model was used in further analysis in mapping the model coefficients. Results from the global model on abundance indicate that there was less likelihood for finding Chambo in 1999 of 12.6% as compared to 2007. Results from the GWR (AIC = 18.62) model explained significantly more variability than the global models GLR (AIC = 40.84) and GAM (AIC = 40.22). Adjusted R2 explained 62.8% in GWR against 41.4% for GAM model. The significant local parameter estimates and t-values for depth were mapped and they provided a visual of their non-stationarity and reduction in the spatial autocorrelation of its model residuals.

CHAPTER ONE

INTRODUCTION

1.1 Background and Motivation

Lake Malawi is endowed with about 500 - 1,000 fish species of which majority are endemic, making it an important living aquatic resource to the country and the world. The Fisheries resource contributes 4% to the Gross National Product (GNP), 60 - 70% to the consumption of the nation's animal protein, 40% to the total protein consumption of the population, which is believed to have reduced drastically to less than 30%, and employs a considerable proportion of the population that is engaged in fishing (Banda et al., 2005b). Of the various fish species available in the lake, Cichlidae family is the number one followed by the Catfish family. *Oreochromis karongae* (a Cichlid) locally known as Chambo is one of the target fish species in the lake as for decades it has been the mainstay of both the commercial and artisanal fisheries in southern Lake Malawi (Turner, 1996; Palsson et al., 1999). It is a delicacy loved by all and carries the Malawi flag as one of the best fish to be taken once one is in the country.

"Chambo" is a common name that is used in reference to four fish species of the genus *Oreochromis* of which three are endemic to Lake Malawi and the other is not. The endemic species are *O. karongae* (38 cm Total Length (TL)) from where the name Chambo comes from, *O. squamipinnis* (37 cm TL) and *O. lidole* (37 cm TL) while *O. shiranus* (37 cm TL) is the only non-endemic species (Turner, 1996; Bell et al., 2012). These occur mainly in shallow waters and have been recorded at depths ranging from 2 m to 50 m although they are most abundant at depths greater than 20 m (Palsson et al., 1999). Knowledge of the distributions and dynamics of such target fish populations is basic to effective fishery man-

agement. For the rest of the document, all the first three fish species endemic to Lake Malawi from the genus *Oreochromis* will be referred to as Chambo.

Different fish species are normally found in areas that are conducive to their own living requirements and Chambo is no exception. The factors affecting their presence can be either environmental, biotic or abiotic. Environmental factors include specific range of water temperatures, pH of the water and salinity; biotic factors include availability of plankton both zoo and phytoplankton while abiotic factors include depth, distance from shore, locations and seasons (Likongwe, 2005). These vary depending on the type of fish species and some fish species are location specific while others are not.

Fishing in the lake is done using different technologies depending on the scale of operation. There are basically three levels of operation which are artisanal (mostly small-scale commercial "traditional" fisheries, with some subsistence fishing), small-scale mechanized (pair trawlers and small stern trawlers) and large-scale "commercial" mechanized fishing (Seymour, 2001). Nets of different sizes in terms of length and breadth are cast in areas where Fishers believe they will catch fish. Some areas have more fish as compared to others and mostly traditional fishermen know where to cast their nets to catch more fish.

Chambo (*O. karongae*) is an endemic fish species of Lake Malawi belonging to tilapiine cichlids of the genus *Oreochromis*. Chambo is one of the fish species that the lake is best known for. The species has been heavily exploited by fishermen both subsistence and small-scale commercial operators who are scattered all over the lake using a variety of fishing gears. The gears in use range from gill nets, beach seine nets, open water seine nets, hooks and lines and fish traps. Over-exploitation has led to the decline of Chambo catches from earlier records of 1940's where its catches were 75% of the annual native artisanal catches from the lake (Palsson et al., 1999). Around the same period of time, Yiannakis started

operating open water ring-nets in the SEA of the lake, landing in excess of two million (individual pieces) annually by 1946 (Banda et al., 2003).

Chambo catches have declined from a record high of 9,400 metric tons in 1985 to as low as 1,400 metric tons by 1999, down to a contribution of 7% of the total annual catch (Banda et al., 2003). In Salima, its catches declined from a maximum record catch of 23% of total catch in 1982 to less than 1% in 1999. In 1980, the Chambo made up about 19% of the total catch in Karonga, but by 1999 it was down to about 2%. The same scenario is registered in the SWA, Nkhotakota and Nkhata Bay. In the period 1976 - 1990, the mean annual percent contribution of the Chambo to the total catch in SEA was 39%, but in the period 1991 - 1999 it decreased to 14% (Kanyerere and Booth, 1999; Banda et al., 2003). A similar reduction in mean annual percent contribution of the Chambo to the whole fisheries landings in these two periods is prevalent in the SWA, Salima, Nkhata Bay, Karonga, Nkhotakota and Likoma.

This decline was observed in all the fishing zones but was more in the SEA which is also a major fishing ground. This fall in its stocks was mainly a result of several factors like overfishing, use of illegal gears, habitat degradation, non-compliance of regulations (Banda et al., 2003), an influx of people on the resource and an exploding lake shore human population with limited alternative livelihood options (Mvula et al., 2003) and also limited spatial coverage of research programs (Kachinjika, 2001). This decline prompted the Government of Malawi (GoM) through the Department of Fisheries (DoF) to develop the Chambo Restoration Strategic Plan (CRSP) which was aimed to restore the stocks by way of reduction of fishing effort, enforcement of laws and regulations and restoration of degraded habitats (Banda et al., 2005a). These management plans have been enforced since 2003 after the Chambo Strategic Restoration Plan. Informatively, the CRSP has not been fully implemented.

1.2 Statement of the Problem

There has been a decline in Chambo fishery resource over the years throughout the lake. Related to that, more research has been done on Biomass and stock assessment (Bell et al., 2012), biology (Banda et al., 2003) and general factors on the decline of Chambo (Bulirani, 2005). However, there has been limited spatial coverage of research programs which can help to provide spatial reference to factors that contribute to the decline. Since early 2000, spatial research has been one of the top research targets according to the Malawi's Fisheries Resource Management, Sustainability and Conservation Act 1997 (Kachinjika, 2001). Ricklefs (1990) in (Windle et al., 2010) observed that fisheries researchers are faced with the challenging task of studying complex patterns and processes in marine resources mostly occurring over large spatial scales. The processes are often examined using population and environmental variables averaged over management units, resulting in a single "global" model applied to an entire study region (Ciannelli et al., 2008). These models function under the assumption of spatial stationarity in the processes under study, whereby the parameters of a process (e.g. mean, variance) are independent of location and direction (Fortin and Dale, 2005).

In aquatic systems, there is dynamic spatial interactions between biological and environmental variables and the fisheries are highly mobile (Rose, 2005; Ciannelli et al., 2008) and one cannot assume stationarity of the processes under study as a rule. Therefore it is important to explore the spatial effects on fisheries resource using other models like global logistic regression and geographically weighted regression against the background that the FRU uses generalized additive models (GAMs) in related analysis of fish distribution and stock assessments.

1.3 Research Questions

Is there any difference in presence of Chambo between the years 1999 and 2007?

Is the probability of finding Chambo at a location the same in the SEA?

Is the distribution of Chambo clustered or spread throughout the SEA?

Which of the models available can better explain the non-stationarity nature of the Chambo in the SEA if present?

1.4 Objective of the Study

The overall objective of the study was to assess the distribution and model population dynamics of Chambo in SEA of lake Malawi at a point in both 1999 and 2007. Specifically, the study wanted to:

1. Compare the distribution of Chambo in the SEA in the years 1999 and 2007.

2. Model the spatial distribution of Chambo in SEA of lake Malawi and identify best best model between global logistic regression (GLR), binomial generalized additive model (GAM) and logistic geographically weighted regression (GWR).

3. Map the parameter coefficients from the best model for spatial interpretation of the observed dynamics.

1.5 Justification

There are a number of undesirable statistical features associated with spatial fisheries data that challenge their analysis through currently available global statistical techniques. The features include patchiness, scale dependency, excess

of zeros or low counts (zero-inflated counts) and spatial correlation (Ciannelli et al., 2008). Furthermore, the many variables that locally affect fish abundance can interact among each other to affect the outcome - a feature known as non-additivity. Both generalized additive models and geostatistical tools offer opportunities in addressing these challenges (Hastie and Tibshirani, 1990; Bez, 2002). Non-additivity implies that interaction is always eminent and in aquatic systems it comes in so many ways for example depth of water affects temperature and availability of dissolved oxygen as well as abundance of phytoplankton and therefore a model that can address the non-additivity nature of the data should be more preferable. A geostatistical approach was used by Kanyerere and Booth (1999) when studying the spatial and temporal distribution of some commercially important fish species in the southeast arm of Lake Malawi. Other models like generalized additive models are available for spatial analysis and FRU currently uses it. Regression techniques like the global logistic regression were also used by FRU before GAMs.

Geographically weighted regression is another recent technique which provides a method for exploring how regression model parameters vary across space (i.e. spatial non-stationarity in the process under study and spatial dependence) (Brunsdon et al., 1996) and represents spatial modification to normal techniques, such as ordinary least-squares regression. The study therefore aimed at comparing the efficiency of these models - global logistic regression, generalized additive model and geographically weighted regression in addressing the statistical challenges faced by fisheries researchers with a case study on Chambo, one of the target fish species in Malawi. The results from the study will help the fisheries research unit through adopting a model that can not only model the distribution better but also map the distribution as GAMs and GLR have limited capacity to map the results.

1.6 Limitations

The findings are mostly in relation to point data from fisheries surveys done in 1999 and 2007. The data that was collected when sampling lacks other ecological parameters that were not captured like temperature, dissolved oxygen and chlorophyll a amount which help to best model the presence/absence of Chambo at a particular location or zone. Fluctuations in the lake levels could also have an effect on the presence or absence of the fish species under study as depth will also vary with varying lake levels. Water current direction and wind direction can also have an effect in the availability of Chambo as these can influence the turbidity of the water. The same also applies to the particular conditions on the day the sampling was done, which can not be the same when done in other different days. Seasonality is also another factor that was not captured as distribution and abundance of fish also depends on it, for breeding and time for recruitment are different seasons.

CHAPTER TWO

LITERATURE REVIEW

2.1 Ecology of Chambo

Chambo is found throughout lake Malawi in shallow areas of 50m in depth and some beyond this depth though not more than the other (Turner, 1996). They are found in shallow vegetated bays, over sand, in purely rocky biotopes and other kinds of habitats. They feed on phytoplankton and on diatom sediment on the sand in both the benthic and pelagic environments by scraping the surfaces of rocks and weeds. They breed from July to March, cresting both in September and in February (Turner et al., 1991; Banda et al., 2005b). It is mostly present in shallower waters from December to March after breeding to nurse the fry and is evenly distributed from June to September. It is generally localized during the breeding season from October when the temperatures are rising, at the end of the rainy season and at the start of the strong south-easterly winds (Van Zalinge et al., 1991; Banda et al., 2005b).

2.2 Modelling Ecological Relationships

Different models for studying ecological relationships in aquatic resources have been used and include geostatistics, (GRASP), generalized additive models, global logistic regression, geographically weighted regression just to mention a few. Geostatistical analysis have often been applied to marine and fisheries ecology data, focusing more on stock abundance and variance estimates e.g. (Kanyerere and Booth, 1999; Wieland and Rivoirard, 2001; Bez, 2002). A geostatistical analysis is defined as referring to: " . . . models and methods for data observed at a discrete set of locations, such that the observed value, z_i, is either a

direct measurement of, or is statistically related to, the value of an underlying continuous spatial phenomenon, $F(x,y)$, at the corresponding sampled location (x_i, y_i) within some spatial region A" (Diggle, 1990). Ecological applications of geostatistical analysis have been very useful for a number of purposes including (a) spatial characterization of fish distribution in relation to biomass (Petitgas, 2001) or season and geographic areas (Mello and Rose, 2005), (b) patterns of spatial correlation over progressively increasing scales (Fauchald et al., 2000), (c) nested spatial structures (Maravelias et al., 2000; Fauchald et al., 2000), (d) spatial distribution of fish in relation to physical habitat (Sanchez and Gil, 2000).

Generalized regression analysis and spatial prediction (GRASP) method basically consists of GAMs used to generate predictions in geographic-grid format and it is known to have solved a significant problem in spatial modeling because it has introduced a way of exporting statistical models to GIS software (GIS, ArcView v.9.2, ESRI, CA, USA). GRASP involves use of two types of distribution, Poisson with a log link function and a binomial with the link function logit for the proportion of response variable in question, while using smoothing spline functions to adjust for the non-linear effects of the model (Cleveland and Delvin, 1988). With it, one can model statistical relationships between a variable of interest and environmental, spatial and temporal variables, then make spatial predictions based on the predictor variables (Lehmann et al., 2002). This model was developed by Lehmann et al. (2002) to partially overcome the problem of using spatial prediction techniques based on interpolation algorithms that are generally data-intensive and requires large quantities of well-distributed data. These are rarely attainable with respect to fisheries, especially when the species of interest is not a target one. GRASP also deals with spatial auto correlation at the data stage, with correlations between the chosen predictors examined to allow the removal of correlated predictors (Carvalho et al., 2011). with the nature

of the binary data in the study, neither geostatistical nor GRASP method will be used but the other regression models, a comparison between the global models (GLR and GAM) to the local model (GWR).

Regression encompasses a wide range of methods for modeling the relationship between a dependent variable and a set of 1 or more independent variables (Charlton and Fotheringham, 2009). There are several ways of explaining the association between variables such as linear, cubic, quadratic, exponential. A Linear Regression Model (LRM) assumes that the association between the independent variables (sometimes known as X-variables, predictor variables or regressors) and the dependent variables (also known as Y-variable, response variables or the regressand) is linear and that the residual error is constant (homoscedastic) (Gujarati, 2004).

A regression model is expressed as an equation and it states that a variable Y changes in a certain manner as another variable, X, changes. In statistical analysis, we are interested in finding out how variables affect each other, for instance how the change in depth in a lake affects the availability of Chambo. In its simplest form, a linear regression model can be expressed as

$$y_1 = \alpha + \beta_1 x_1 + \varepsilon \tag{2.1}$$

where y_1 is the response variable representing the presence of Chambo at a specific depth, x_1 is an independent variable (the depth), α is the constant while β_1 measures the marginal effect (slope parameter) i.e. how y_1 changes when x_1 changes (Charlton and Fotheringham, 2009). The error term is represented by ε and includes all other factors that affect y except from x. According to Charlton and Fotheringham (2009), equation (2.8) holds under the following assumptions: linear in parameters; no perfect collinearity between independent variables; zero conditional mean; homoscedasticity of error terms and normality of error terms. Such a model is usually fitted using Ordinary Least Squares (OLS) procedure.

Mathematically, it can be shown that, given the above assumptions, the estimator from equation (2.8) is Best Linear Unbiased Estimator (BLUE). However, are the assumptions realistic, do they hold in this study's data set and if not, what are the consequences and further what are the solutions? The major challenge regarding the data set is that the percentage of Chambo in a catch from which the dependent variable (presence or not) emanates has a binomial distribution, i.e. the dependent variable is a dummy, which takes a value of zero or one depending on whether or not Chambo is present in a catch (1) or absent (0). However, the independent variables are continuous.

Using LRM in this type of data has a serious defect in that the estimated probability values can lie outside the 0 - 1 range. There are several methods that can be used in the analysis of data involving binary outcomes such as Logit models and generalized additive models that can best fit the non linear relationship between the dependent and explanatory variables. The justification for using logit is its simplicity of calculations and that its probability lies between 0 and 1. Moreover, its probability approaches zero at a slower rate as the value of explanatory variable gets smaller and smaller and the probability approaches 1 at a slower and slower rate as the values of explanatory variable gets larger and larger (Gujarati, 2004).

2.3 The Logistic Regression

The logistic regression is a specialized form of regression that is formulated to predict and explain a binary (2–group) categorical variable rather than a metric dependent measure (McCullagh and Nelder, 1989). The application of logistic regression can be viewed as from six–stage model building perspective. As with all multivariate applications, setting the objectives is the first step in the analysis. Then the researcher must address specific design issues and makes sure

that the underlying assumptions are met. The binary measure is translated into odds of occurrence and then a logit value that acts as a dependent measure. The model formed in terms of independent variables is almost identical to multiple regression.

Model fit is first assessed by looking for statistical significance of the overall model and then determining the predictive accuracy by developing a classification matrix. Then, given the unique nature of the transformed dependent variable, logistic coefficients are given their "original" scale, which is interpreted more like regression coefficients. Each form of the coefficient details a certain characteristic of the independent variables impact. Finally, the logistic regression model should be validated with hold out sample.

The logistic regression model is based on a logistic function, which describes the mathematical form on which the logistic model is based. If we let P_i to be a probability that a catch of fish at a certain depth in a lake will contain Chambo, then the cumulative logistic function is specified as

$$\begin{aligned} P_i &= f(u_i) \\ &= \frac{1}{1 + exp(-u_i)} \end{aligned} \tag{2.2}$$

where $exp(u_i)$ is the exponential function, equivalent to e^u. In turn, e is the exponential constant which is approximately equal to 2.71828. Its defining property is that $log(e^u) = u$. The value of u_i varies from -∞ to +∞ and the range of $f(u_i)$ is between 0 and 1, irrespective of the value of u_i.

Hosmer and Lemeshow (1989) pointed out that the logit model could be written in terms of the odds and log of odds, which enables one to understand the interpretation of coefficients. The odds ratio implies the ratio of the probability (P_i), that there is Chambo presence in a catch to the probability $(1 - P_i)$ that there wont be Chambo where

$$(1 - P_i) = \frac{exp(-u_i)}{1 + exp(-u_i)} \tag{2.3}$$

For notational convenience, we will denote the probability statement as simply $p(x)$, where x is a notation for the collection of variables x_1, x_2 through x_n.

The response variable in logistic regression is usually dichotomous, that is, the response variable takes the value of 1 which is probability of success P, or the value 0 which is the probability of failure $(1 - P_i)$. This type of variable is called a Bernoulli (or Binary) variable. As mentioned previously, the independent or predictor variable in logistic regression can take any form. That is, the logistic regression makes no assumptions about the distribution of the independent variables. They do not have to be normally distributed, linearly related or of equal variance within each group. The relationship between the predictor and the response variables is not a linear function in logistic regression, instead, the logistic function is used which is the logit transformation of P.

To obtain the logistic model from the logistic function, the u_i is written as a linear Σ as follows

$$u_i = \alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n \tag{2.4}$$

where $x's$, α and β are as defined before. Substituting equation 2.4 into 2.2 we obtain

$$f(u_i) = \frac{1}{1 + exp(-\alpha - \Sigma\beta_i x_i)} \tag{2.5}$$

Thus, the logistic model may be written as

$$p(x) = \frac{1}{1 + exp(-\alpha - \Sigma\beta_i x_i)} \tag{2.6}$$

However, the above logistic function is non linear, the logit transformation would be used to make it linear, and this is given as

$$logit\ p(x) = ln_e \left[\frac{p(x)}{1 - p(x)} \right] \tag{2.7}$$

This transformation allows us to compute a number, called logit $p(x)$, for a location/area with independent variables given by x. By substituting equation 2.3 and 2.6 into equation 2.7, we obtain

$$\begin{aligned} ln_e\left[\frac{p(x)}{1-p(x)}\right] &= ln_e\left[\frac{\frac{1}{1+e^{-u_i}}}{\frac{e^{-u_i}}{1+e^{-u_i}}}\right] \\ &= ln_e\left[e^{u_i}\right] \\ &= \alpha+\Sigma_{i=1}^{n}\beta_i x_i \end{aligned}$$

and therefore

$$\begin{aligned} x_i\ logit\ p(x) &= \alpha+\Sigma\beta_i x_i \\ &= \alpha+\beta_1 x_1+\beta_2 x_2+...+\beta_n x_n \end{aligned} \tag{2.8}$$

Thus, the logit $p(x)$ simplifies to the linear Σ. The quantity $p(x)|1-p(x)$, whose log value gives the logit, describes the odds for a location or an area to have no fish in the presence of independent variable. $p(x)|1-p(x)$ = Odds for location i in the presence of independent variables. The goal of logistic regression is to correctly predict the category of outcome for individual cases using the most parsimonious model. To this end, a model is created that includes all predictors that are useful in predicting the response variable.

Modelling of the factors that may affect the ability of fish to grow, migrate, survive and reproduce using logistic regression model assumes a linear relationship between fish performance and environmental variables, when actually the relationships are very likely to be non linear (Bigelow et al., 1999). Despite the advantages of linear regression techniques in determining model parameters and

their interpretation, the method has little flexibility because of its relatively restricted range of application (Chong and Wang, 1997). To overcome such difficulties, generalized additive models have been used to identify, characterize and estimate the relationships between extrinsic factors and catch rates of certain fish species (Walsh et al., 2002; Zagaglia et al., 2004; Damalas et al., 2007).

Logistic regression models have been widely applied in marine and fisheries management. Narumalani et al. (1997) applied logistic regression in aquatic macrophyte modelling; Grift et al. (2003) used it in fisheries-induced trends in reaction norms for maturation in North Sea plaice where logistic regression was applied in analyzing the age and length at maturity; Ellis et al. (2006) used it in predicting macrofaunal species distributions in estuarine gradients; Bi et al. (2007) applied it in modelling the pelagic habitat of salmon off the Pacific Northwest (USA) coast.

2.4 Generalized Additive Model

GAMs were first proposed by Hastie and Tibshiran (1986), as a non-parametric regression techniques that are not restricted by linear relationships and it is flexible regarding the statistical distribution of the data (Hastie and Tibshirani, 1990; Swartzman, 1997; Valavanis et al., 2008). The only underlying assumption made is that the functions are additive and that the components are smooth. A GAM, like a generalized linear model (GLM), uses a link function to establish a relationship between the mean of the response variable and a "smoothed" function of the explanatory variable(s). The strength of GAMs is their ability to deal with highly non-linear and non-monotonic relationships between the response and the set of explanatory variables.

GAMs are sometimes referred to as data- rather than model-driven. This is because the data determine the nature of the relationship between the response and

the set of explanatory variables rather than assuming some form of parametric relationship (Yee and Mitchell, 1991). Like GLMs, the ability of this tool to handle non-linear data structures can aid in the development of ecological models that better represent the underlying data, and hence increase our understanding of ecological systems, now widely applied in fisheries science.

In GAM, the coefficient β_i in equation 2.8 is replaced by a smooth function such that

$$logit\ \ (p(x)) = \alpha + \Sigma f_i x_i \tag{2.9}$$

where $f(x)$ is a non-parametric function describing the effect of x_i on $p(x)$ (McCullagh and Nelder, 1989). The model remains additive with respect to the covariates but is no longer linear in them. The logit link function will be used when fitting the binomial GAM with the shape of the f_i function determined by penalized regression splines with automatic smoothness selection (Wood, 2006).

There are numerous applications of GAM to marine and fisheries spatial data. They have been applied to acoustic data sets to investigate the relationship between environmental factors and horizontal distributions of other fish species like walleye pollock in the Bering Sea (Swartzman et al., 1994, 1995); herring in the Northeastern Atlantic (Bailey et al., 1998; Maravelias et al., 2000,a); Japanese anchovy (*Engraulis japonicus*), sand lance (*Ammodytes personatus*), and krill (*Euphausia pacifica*) in Sendai Bay, Japan (Murase et al., 2009); spatial prediction of demersal fish distribution (Moore et al., 2009); spatial distribution of blue shark (*Prionace glauca*) catch rate and catch probability of juveniles in Southwest Atlantic (Carvalho et al., 2011).

Other GAM applications include, but are not limited to distribution of: flatfish and fish stomach contents (Swartzman et al., 1992), eggs (Fox et al., 2000; Wood and Augustin, 2002; Ciannelli et al., 2007), squids (Denis et al., 2002), sea trout (Kupschus, 2003), tuna (Zagaglia et al., 2004), and crabs (Jensen et al., 2005).

It was also used by Francis et al. (2005) in predicting small fish presence and abundance in northern New Zealand harbours and Giannoulaki et al. (2006) to identify the relationship between anchovy presence and environmental variables. However, there have been few applications of GAMs to create maps of the distribution of fish (Beare et al., 2002).

Both global logistic regression and GAM regression models are unit level random effects models that are widely applied in small area estimation. Typically, such models assume independence of random area effects and individual effects (Chandra et al., 2010). In environmental, ecological and geographical applications, however, observations that are spatially close may be more related than observations that are further apart. This spatial correlation can be modeled by extending random effects models to allow for spatially correlated area effects, for example via a Simultaneous Autoregressive Regression (SAR) random effects model (Anselin, 1990; Cressie, 1993). Pratesi and Salvati (2008) investigated the use of Spatial Empirical Best Linear Unbiased Predictor (SEBLUP) for small area estimation in this situation. SAR models allow for spatial correlation in the error structure. An alternative approach to incorporating the spatial information in the regression model is by assuming that the regression coefficients vary spatially across the geography of interest.

2.5 Geographically Weighted Regression

These are types of models that were introduced by McMillen (1996) and McMillen and McDonald (1997) as non–parametrically locally linear regression models where the cases are geographical locations and Brunsdon et al. (1996) labeled them as GWR. Brunsdon et al. (1996); Fotheringham (1997); Fotheringham et al. (2002) extended the traditional regression model by allowing local rather than global parameters to be estimated. That is, GWR directly models spatial non-

stationarity in the mean structure of the outcome variable. In GWR, the observations taken in its consideration for the formation of the model are weighted with regard to their location.

The underlying idea behind GWR is that parameters may be estimated anywhere in the study area given a dependent variable and a set of one or more independent variables which have been measured at places whose location is known. Considering Toblers observation of nearness and similarity, estimated parameters for a model at some location i will result in observations nearer the location having a greater weight in the estimation than observations further away (Tobler, 1970).

GWR therefore provides a method of exploring how regression model parameters vary across space, i.e. spatial non-stationarity in the process under study. It provides spatial modifications to normal techniques such as OLS, GLMs, GAMs and linear mixed models, more especially when modelling vegetation and animal distributions in their terrestrial ecology (Kupfer and Farris, 2007; Osborne et al., 2007; Kimsey et al., 2008). This model shows to be a promising exploratory tool for investigating spatial non-stationarity in other ecological niches. In a binomial or logistic GWR, the variable being modeled takes the binary condition of either being present or absent as outlined in equation 2.10. This regression model is essentially a modified global regression which incorporates a set of geographic coordinates for each location i, taking the form

$$logit\ p(x) = \beta_{0i} + \sum_{i} \beta_{ij} x_{ij} + \varepsilon_{ij} \tag{2.10}$$

where β_{0i} is the intercept parameter specific to location i; β_{ij} is the parameter coefficient of independent variable x_j at location i ; ε_{ij} is the Gaussian error at location j and j is defined by the x-y coordinate of the i^{th} location while $\beta_{ij}x_{ij}$ are coefficients varying conditional on the location.

Local variable coefficients in this model are determined by a weighting matrix that uses a distance-decay function, resulting in local regression points being

more influenced by observations closer in space. The spatial kernel controlling the distance-decay function can take either a fixed (distance) or adaptive (number of samples) approach to establishing the radius of the local GWR model, in effect creating a moving window regression for each observation point in the study area (Fotheringham et al., 2002). The size of the kernel bandwidth has a large impact on the outcome of the GWR analysis and should be selected carefully. Increasingly smaller bandwidths result in parameter estimates that are highly localized and have a large degree of variance, whereas increasingly larger bandwidths tend towards the normal global regression estimates.

Suppose d_{ij} denotes the distances between region i and j, and $w_j(i)$ denotes the element of weighting matrix at region i assigned to region j. A possible choice of $w_j(i)$ is expressed as function of

$$w_j(i) = exp\left[-\frac{1}{2}\left(\frac{d_{ij}}{b}\right)^2\right] \tag{2.11}$$

where $j = 1,2,L,n$ and b is the bandwidth. If i is exactly equal to j, the weights at the point will be unity and the weight of the others will decrease and follow a Gaussian curve as d_{ij} increases. An alternative to Gaussian function to determine elements of weighting matrix is bi square function which sets the weights of data point inside radius of bandwidth to decrease to zero as d_{ij} increases and discards the others. Bi square function is expressed as

$$w_j(i) = \left[1 - \left(\frac{d_{ij}}{b}\right)^2\right]^2 \tag{2.12}$$

if $d_{ij} < b$ otherwise $w_j(i) = 0$. Both Gaussian and Bi square functions allow the weights to vary continuously over region (Fotheringham, 1997; Fotheringham et al., 2002). In choosing the weighing matrix, it is very important to predetermine an optimum bandwidth b which can be done by minimizing either the cross-validation (CV) score or the Akaike Information Criterion (AIC). In this

study, the Gaussian weight was used based on the Euclidean distances between the two observation sites/locations and the bandwidth in map units. The kernel bandwidth was determined by minimizing the corrected Akaike Information Criteria (AICc) for the fitted regression model.

The GWR model significance or goodness–of–fit is estimated statistically by using either of these methods: (a) Brunsdon, Fortheringham and Chartlon F Test, (b) Leung, Mei and Zhang F1 Test, (c) Leung, Mei and Zhang F2 Test. Since GWR produces parameter estimates for each location, a test of their variability is needed to establish the significance of the parameter estimates. Leung et al. (2000) proposed *F3* test which takes place in diagnosing the null hypothesis that the set of parameters tend to be constant over region. A large value of *F3* supports alternative hypothesis that a given parameter tends to vary over region. Besides formal procedures described above for significance tests, performance of GWR model could be established according to determination coefficient (R^2) which was used in the study and AIC statistic, the latter when compared to other models.

GWR models have been applied in fisheries and marine studies not as much as GLRs and GAMs have been applied. Some examples of studies done using GWR model were by Jones et al. (2009) who applied it to quantify relationships between corals occurrence/abundance and six environmental parameters to determine how these parameters may control the distribution of cold-water coral species in the Northwest Atlantic region. Windle et al. (2010) used it in exploring spatial non–stationarity of fisheries survey data in Northwest Atlantic.

2.6 Best Model Selection

GLR, GAM and GWR are all models which can either be closer to reality or far from it in estimating the parameters of interest (McCullagh and Nelder, 1989;

Rodríguez, 2005). The model whose explanatory variables are explaining more of the variation in the data under study is much better preferred as compared to those which are explaining very little. This process is referred to as best model selection. There are however different methods of selecting the best model and different tests are employed in model selection. The methods for analyzing multivariate binary responses can be classified into two broad classes of methods: likelihood based and estimating equation based methods (Moore et al., 2009).

The likelihood based methods require complete specification of the joint distribution of the multivariate responses, whereas the estimating equation based methods can be employed when joint distribution is not fully specified. The most common likelihood based methods for multivariate binary data are multivariate probit and multivariate logit models which consider univariate normal and logistic distributions as univariate margins, respectively (Latif et al., 2008). Model selection is an important part of data analysis which leads to a search of "best" model. By "best" model, it means selecting the best subset of the covariates from the available covariates in the data. Usually model selection is done by using a specific criterion. For likelihood-based methods, (AIC) (Akaike, 1973) is widely used as a model selection criterion as compared to Bayesian Information Criterion (BIC) (Schwarz, 1978).

AIC takes the form of

$$AIC = -NL_N\left(\hat{\theta}\right) + d \tag{2.13}$$

while BIC takes the form of

$$BIC = -NL_N\left(\hat{\theta}\right) + \frac{d}{2}logN \tag{2.14}$$

AIC and BIC are derived from distinct perspectives: AIC intends to minimize the Kullback-Leibler divergence between the true distribution and the estimate from a candidate model while BIC tries to select a model that maximizes the posterior

model probability. Due to the rather different motivations, it is not surprising that they have different properties. However, the most well known properties of AIC and BIC are asymptotic (loss) optimality and consistency (in selection) respectively (Yang, 2003). These two properties of AIC and BIC are respectively called consistency and asymptotic (nonparametric) optimality (under the average squared error loss). Generally, AIC is not consistent and BIC is not asymptotically (loss) optimal in the nonparametric case, hence the need for using AIC due to the asymptotic nature of the models used in the study.

CHAPTER THREE

RESEARCH METHODOLOGY

3.1 Study Area

Fisheries research in Malawi is conducted by the Fisheries Research Unit (FRU) which is based in Monkey Bay, Mangochi district. This institution has the mandate to provide the Department of Fisheries with such information based on understanding of the biology, life history and distribution of the target species (Kachinjika, 2001). Depending on the depth and abundance of fish species, the lake was divided into fishing grounds from the south going up to1 the northern region. The sections are Lake Malombe, Upper Shire River, South East Arm (SEA) in Mangochi, South West Arm (SWA) in Mangochi and Salima, Domira Bay in Salima, Nkhotakota, Likoma and Chizumulu Islands, Nkhata Bay and Karonga (Figure 3.1). However, Chambo is abundant in the SEA of the lake followed by SWA, making these areas to be major fishing grounds. The SEA, like other regions, is divided into three sections: A, B and C for management purposes. In a study by Kanyerere and Booth (1999), it was established that more of the Chambo is landed in the SEA.

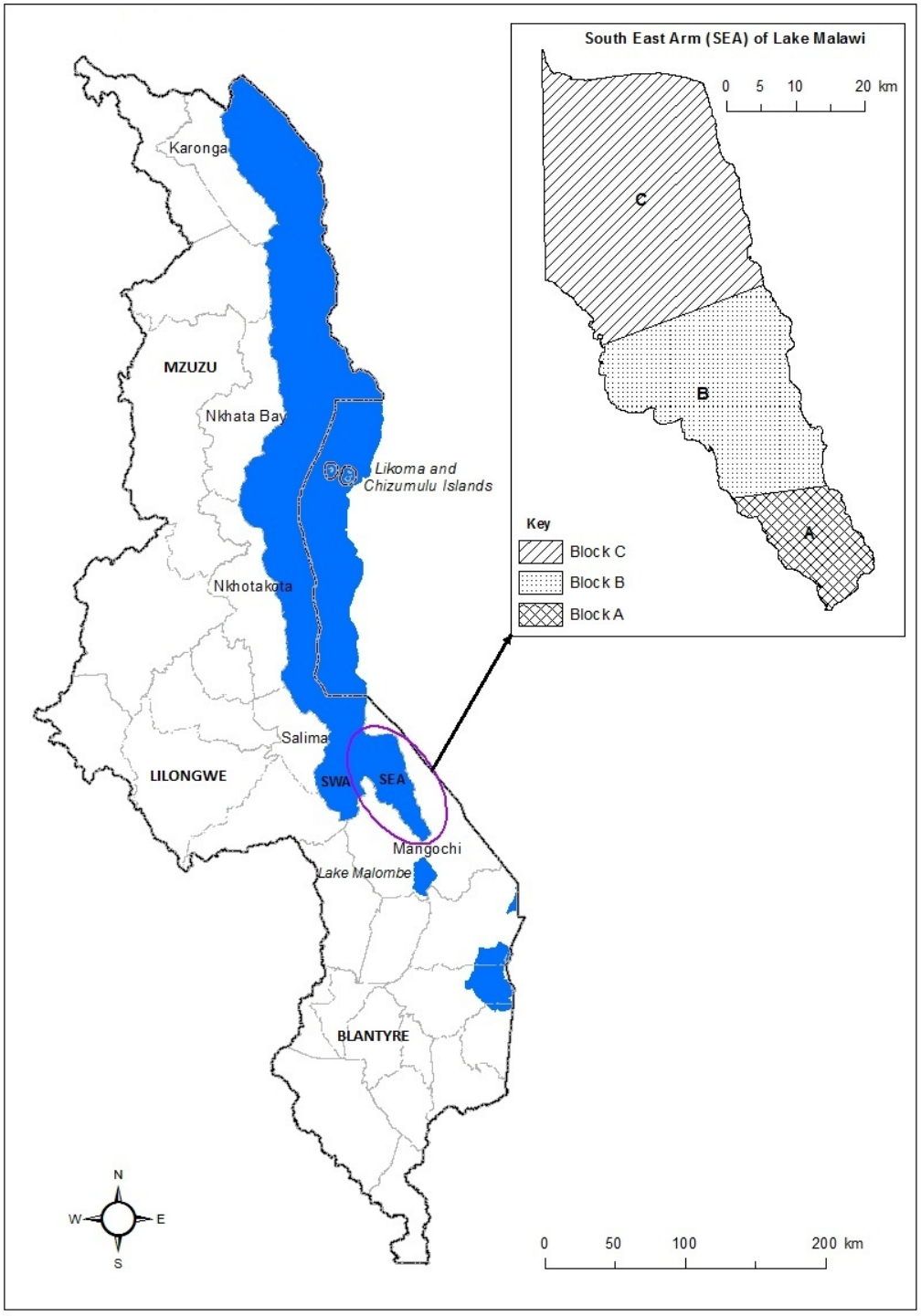

Figure 3.1: Map of Malawi showing the Lake Malawi and the sampling site
Source: WorldFish, 2010

The study covers the South East Arm (SEA) of Lake Malawi which shares its boundary with Mangochi district. There is a fisheries research station in Monkey Bay which is responsible for fisheries research on fishery resource and associated limnological aspects of Lake Malawi with other similar small stations in Salima in the central region and Karonga in the northern region of Malawi. The SEA under study has three fishing zones. The zones were demarcated for management purposes and done according to depth and fish abundance, designated as A, B and C which are shallow, medium and deep waters respectively. Figure 3.2 is the map showing the SEA and the fishing zones with areas for fishery data collection. There are a total of 50 small polygons or areas from where fisheries surveys on the SEA of the lake take place, here referred to as locations. These represent all the 3 zones, with 7 data collection stations in zone A, 18 in zone B and 25 in zone C, each proportional to size of area.

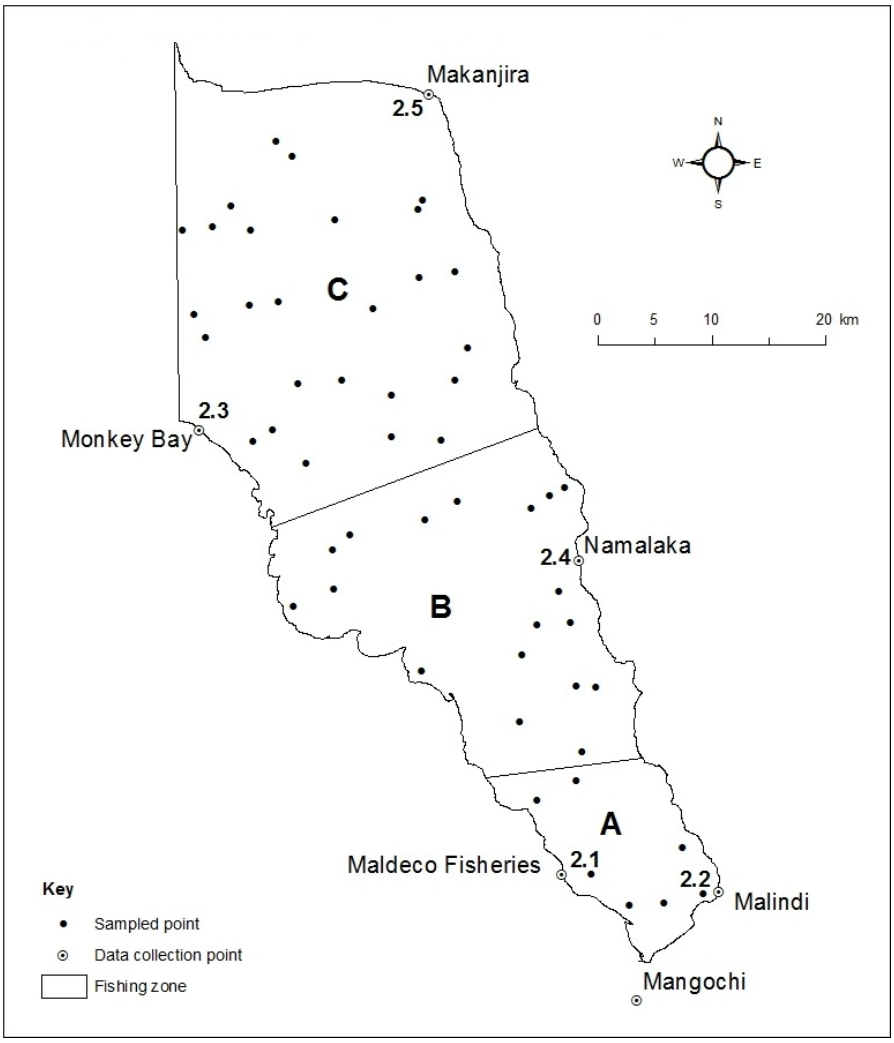

Figure 3.2: Map showing SEA and its fishing zones, fishery data collection points and sampled points
Source: Monkey Bay Fisheries Research Unit

3.2 Data Collection

3.2.1 Data Sources

Data used in this study is retrospective, collected from two of the multispecies fisheries surveys conducted on the lake in June 1999 and October 2007 by the

Monkey Bay Fisheries Research Unit. These surveys are normally done twice a year but due to unavailability of funds, there has been only a single survey done since late 1990's up to 2007. At each trawl location, data were collected on the shooting latitude and longitude and after 30mins of trawl, hauling latitude and longitude was also recorded plus fish biomass and abundance of the various fish species available and depth of the catch. The catch is done once per location and then move on to another location. Once the fish are caught, they are sorted according to size and species, labeled and stored for further action like weighing and length measurements at the research station. Big fish are weighed individually and the total length is measured. For small fish, samples are taken to represent the catch. All this is recorded in data sheets which are now given to data clerks to enter in a database. Data was therefore retrieved from the data base and screened to only have Chambo data which was used. We give credit to Monkey Bay Fisheries Research Unit for providing the data.

3.2.2 Sampling Method

MV Ndunduma is the vessel used in the surveys and has these hauling specifications: velocity (V) of the trawl over the ground when trawling taken to be 3.5nm; head rope (h, nm) length which is 0.01242; trawling time (t) which is 30mins per trawl; x2 is that fraction (0.639) of the head-rope which is equal to the width of the path swept by the trawl, and the 'wing spread' is h*x2 (courtesy of Monkey-Bay Fisheries Research Unit). These are the values used in the variogram when geostatistics is used to analyze the data. Attached to the vessel is a 38mm mesh size trawl net with a 38mm cod-end mesh size which is the minimum mesh size restriction for the trawl cod-end (Kachinjika, 2001). Sampling during the fishery survey is done in specific locations indicated above, guided by the coordinates that demarcate them. These locations are randomly spread across the lake from

where trawling takes place. As they trawl, an ecological area or polygon is covered from where different fish species are caught. The whole SEA is represented by 50 different and independent samples of which each is represented as a location with mean coordinates X and Y.

3.2.3 Variables Definition

The study had two types of dependent variables, one for comparing the presence of Chambo between two years (odds ratio) while the other was for modelling Chambo presence in 2007. For comparison of Chambo presence, density of Chambo (amount of Chambo in a catch) was examined based on the weight contribution of Chambo in the total catch. Chambo was classified as present or absent at each sampling station and any weight above 0kgs was taken as presence of Chambo, otherwise 0 if there was no Chambo in the catch. This was the trend in the data from 1999 and 2007. According to a 1999 study on the fishery resource in the SEA, 14% was established as the percent contribution of Chambo in the SEA according to (Kanyerere and Booth, 1999). This density was to be cross-checked with 2007 data and appreciate any changes within the Chambo density. The presence of Chambo bears the code names CP99 for 1999 and CP07 for 2007, both as derived dependent variables as outlined in the data set and models.

On modelling the presence and spatial distribution of Chambo in 2007, the dependent variable used is either the presence or absence of Chambo (y_i) at a location based on data generated from the catch. If there was more than 0 kg on a catch, the generated dependent variable (CP07 in the data set) was coded as 1, 0 otherwise as outlined below

$$y(i) = \begin{cases} > 0Kg & \text{Chambo Present (1)} \\ \leq 0Kg & \text{Otherwise (0)} \end{cases}$$

Both primary and secondary data were used to analyze the probability of finding Chambo on a station. Depth (m), Distance (m) and Area covariates were used to explain the distribution and abundance of Chambo in the study area. Depth is the length into the water column from the surface to where the fish were caught within the water column, measured in meters. Distance is the shortest length measured from the nearest shoreline to the location where fish catches were executed. Area is the zone that is composed of several locations. Table 3.1 shows the variables used in the study.

Table 3.1: Description of variables used in the study

Type of Variable	Description
Binary	
CP99	1 = Presence of Chambo at a station in 1999
	0 = Absence of Chambo at a station in 1999
CP07	1 = Presence of Chambo at a station in 2007
	0 = Absence of Chambo at a station in 2007
Metrical	
Depth (m)	Distance from surface to where fish is caught
Mdepth (m)	Mean depth from the 1999 and 2007 depth data
Distance (m)	Shortest length from either shore to data collection point
Spatial	
Location	50 stations where fish were sampled, coordinate x,y
Area	3 structured areas demarcated for fishery management

Depth, Location and Area are primary covariates while Chambo and distance are secondary data variables, derived from the data. Distance was estimated using ArcView GIS by digitizing the available 50 coordinate positions to the nearest shore line. For the coordinates, shooting (casting net) longitude and latitude plus hauling (dragging the net out) longitude and latitude meant that the coordinates present an area (a polygon). These coordinates were then averaged to have a single coordinate point (X,Y) to represent the area and this point is the one being

used in the analysis. Distance was measured from a fixed map as the satellite pictures of the lake in the different years 1999 and 2007 were not present and this does not mean that the shoreline is fixed. Other environmental factors which could also help explain the presence of Chambo in an area could be water current flow, water temperature at depth where fish was caught and chlorophyll a amount in the area the fish were caught. However, these were not part of the parameters the survey records during the exercise.

3.3 Data Analysis

3.3.1 Modelling Approaches

The statistical approaches mostly used to analyze Chambo spatial distribution and abundance data can be loosely grouped into two categories, according to whether the emphasis is on the relationships among neighboring observations or on the relationship among the observations and the collected environmental variables (Ciannelli et al., 2008). The first group is based on techniques developed for geographical analysis and mining resources (Matern, 1986) also known as geostatistical analysis while the second group is an extension of common regression techniques applied to spatial data (e.g. Guisan et al., 2002). Separately, the second technique captures important ecological processes of fish distribution. It is important to note that both analytical techniques attempt to model the local species abundance (y_i) based on a similar underlying statistical model of the type:

$$y_i = u_i + \varepsilon_i \tag{3.1}$$

where u is a mean effect (i.e. the known and explainable proportion of the model) and ε is the error (the unknown proportion of the model).

The dependent variable in this study as explained in the previous sections, is binary in nature (present =1 or absent = 0), hence a logistic regression approach (Cox, 1970; Hosmer and Lemeshow, 1989) is adapted in the relationship between the dependent variable and other environmental variables (depth, distance). For uniformity of the covariates in determining the odds ratio between 1999 and 2007 Chambo presence; distance, depth and area covariates are the same, only the presence of the Chambo is different as it is from two different points in time. For distance, it is assumed to be the same in both years as the location or area sampled is maintained, disregarding the differences in lake levels which are not known. On depth from the different years, a t-test was run to check the null hypothesis that the true difference between the means is equal to zero (Equation 3.2). A mean depth ($MDepth$) was used for running the logistic regression for comparing presence of Chambo between the two years. A t-test and the comparison model (M_{COMP}) were run as outlined below:

$$T-test: \quad Depth\,99 = Depth\,07 \tag{3.2}$$

$$M_{COMP}: \quad ln\left[\frac{CP99}{CP07}\right] = \alpha+\beta_1(MDepth)+\beta_2(Distance)+Area \tag{3.3}$$

where Depth 99 and Depth 07 is the depth at which fish were caught in 1999 and 2007 respectively, resulting in mean depth variable used in the comparative model M_{COMP}; $In[CP99/CP07]$ is the odds ratio for finding Chambo in 1999 against finding it in 2007, α is the intercept value, β_1 is the parameter coefficient for mean depth (MDepth), β_2 is the parameter coefficient for distance and area is a factor

The following models were fit and compared to find the best model that can explain the non-stationarity of Chambo much better using 2007 data; GLR, GAM

and GWR as follows:

$$M_{GLR}: \ ln\left[\frac{P(x)}{1-P(x)}\right] = \alpha + \beta_1(Depth) + \beta_2(Distance) \quad (3.4)$$

$$M_{GAM}: \ ln\left[\frac{P(x)}{1-P(x)}\right] = \alpha + f_1(Depth) + f_2(Distance) \quad (3.5)$$

$$M_{GWR}: \ ln\left[\frac{P(x)}{1-P(x)}\right] = \alpha + \beta_{1j}(Depth) + \beta_{2j}(Distance) \quad (3.6)$$

where $P(x)$ in this regard is the probability of finding Chambo; $1-P(x)$ is the probability of not finding Chambo; $ln[P(x)|1-P(x)]$ is the odds of finding Chambo given the covariates of depth and distance as the determinants of Chambo presence or absence. α is the intercept value; β_1 and β_2 are parameter coefficients for depth and distance variables used in GLR model (M_{GLR}); f_1 and f_2 are smooth parameter coefficients for depth and distance used in GAM model (M_{GAM}); β_{1j} and β_{2j} are parameter coefficients with j^{th} location for depth and distance as used in GWR model (M_{GWR}).

GLR and GAM were applied in an attempt to explain global relation between the dependent and independent variables. Note that in logistic models, there is no error term as in linear models. In the GWR model, a t-statistic for each coefficient and local regression was run to inform us if there are places in our study area where the coefficient for a given variable is significantly different from the expected value. A Bonferoni test was run to check on the significance of the variation. The model run was tested using the (BFC) and (LMZ) which diagnoses the null hypothesis that the set of parameters tend to be constant over a region. BFC compares OLS model fit to GWR model fut using ANOVA test and the LMZ test tests for spatial variability of the parameter estimates.

3.3.2 Analysis Package

Microsoft Excel was used to organize the data and derive the other secondary variable of Chambo presence or absence determined by weight and mean depth for 1999 and 2007 saved in comma delimited file format (.csv) for easy of use in R package. Arc View GIS software was used in establishing the attributes table for the SEA of the lake from where the boundary shape files for SEA needed in R package were used. Estimation of the nearest distances from the station to the shore line was also done in ArcView GIS by using the straight line distance measure. The statistical package used for the analysis was R version 2.15.0, using the "glm", "mgcv", "spgwr", "sp" and "Maptools" packages downloaded from R cran website. All programs/codes run during the analysis are attached in appendix A.

CHAPTER FOUR

RESULTS AND DISCUSSION

4.1 Exploratory Data Analysis

Main variables used in the models were Depth, Distance and Area. Depth (m) was the equivalent length in meters from where the catch of the fish was effected. Distance (m) on the other hand was the shortest distance from either shore to where the fish were sampled or caught. Since the samples are done on the same area each time, distance maybe the same in both years if there is no significant variations in lake levels. A distribution of the values in depth from both years are given in the box plots (Figure 4.1). Depth in 1999 has no outlier, however there are two outliers in 2007 in area A, both above and below the median depth. Median depth from respective areas compared in both years has little variation. It is clearly shown that area A is shallow with a median depth of about 20m as compared to area C whose median value is slightly above 80m in both years.

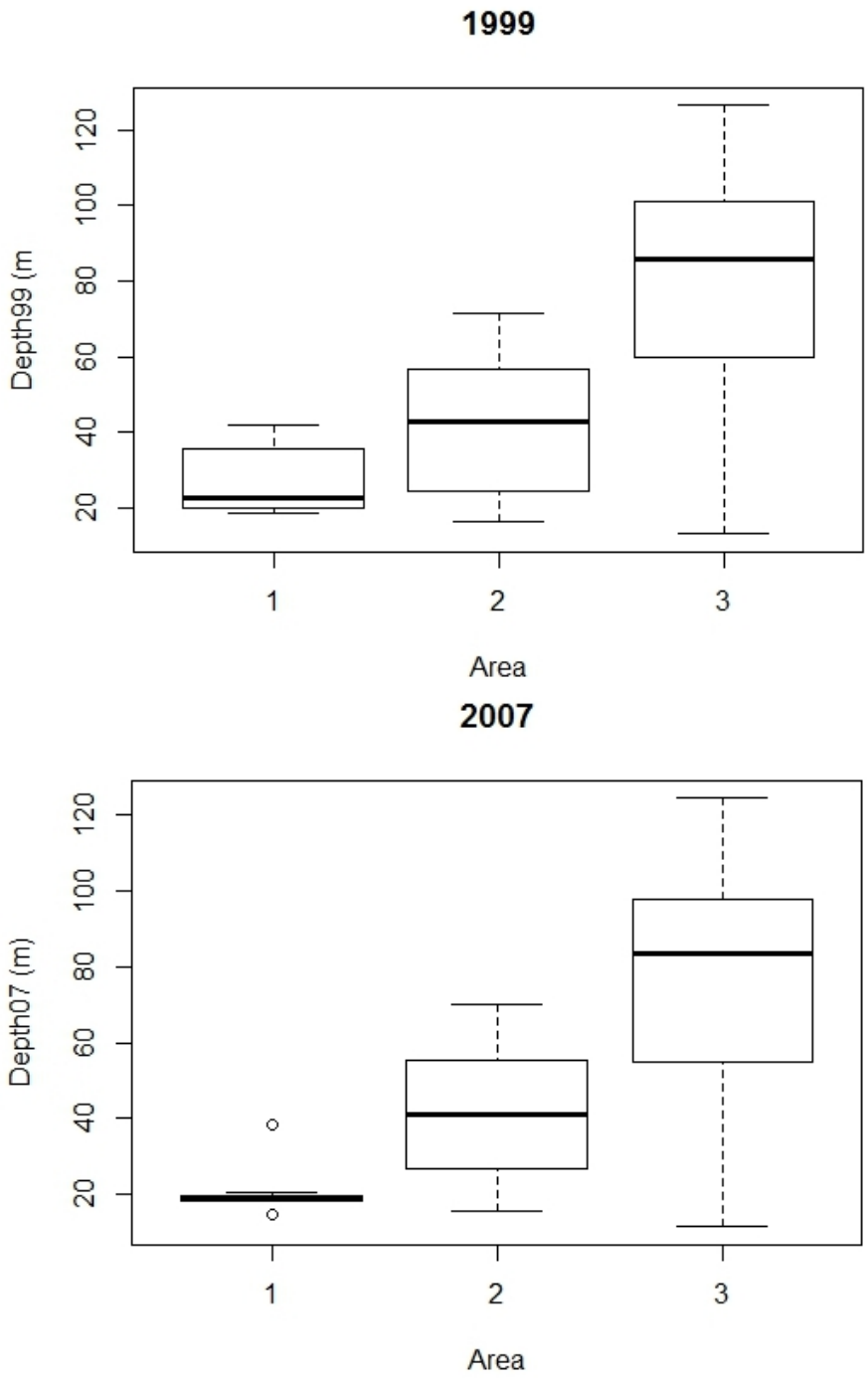

Figure 4.1: Box plots for depth in 1999 and 2007 respectively per area in SEA

Figure 4.2 below signifies that the trend is similar to that of depth as distance from shoreline to fishing location/site also increases from area A to C in both

years. One value for distance is an outlier in area A and another in area B and these can have an effect in the results of the models run. The same trend is depicted with distance such that the distance from either shore to where fish were sampled in area A is slightly less than 2000m as opposed to area C whose median distance is slightly above 5000m. Area A has a median distance of less than 2000m and the distribution is almost normally distributed unlike in areas B and C which have values of around 5000m. There is more variation in areas B and C, with a positive and negative skewness respectively.

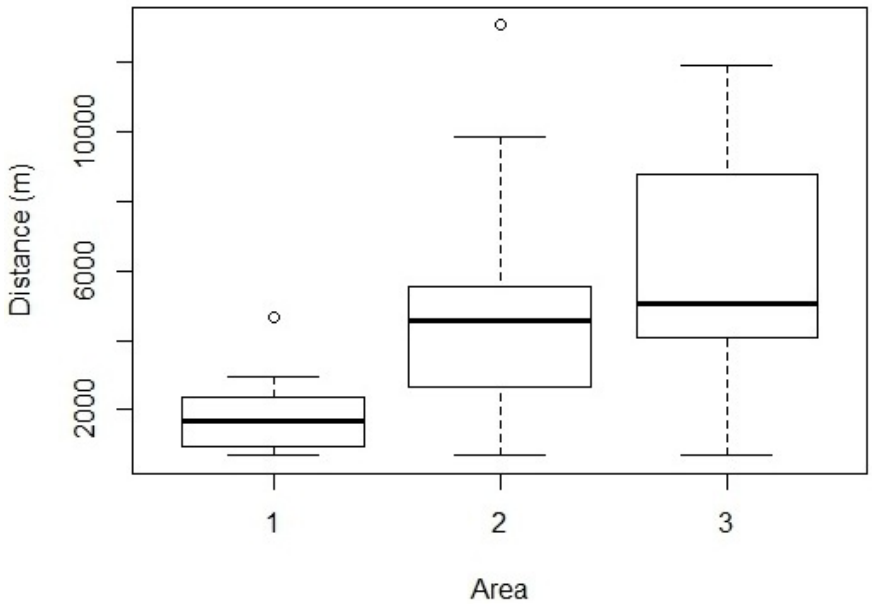

Figure 4.2: Box plot showing the distribution of distance from shoreline by area for both 1999 and 2007

Figure 4.3 reveals that Chambo is more available in area A in both years with a median value of above 50kgs in 2007 as compared to about 10kgs in the same area in 1999. Area B is scanty in both years as the distribution of the Chambo catch is evidenced by several outliers. Area C in 1999 had a very small catch of 0.5kg which is not present in 2007. Note should also be taken of the highest

catches in the two years where 1999 had 150kgs and 2007 had catches closer to 250kgs.

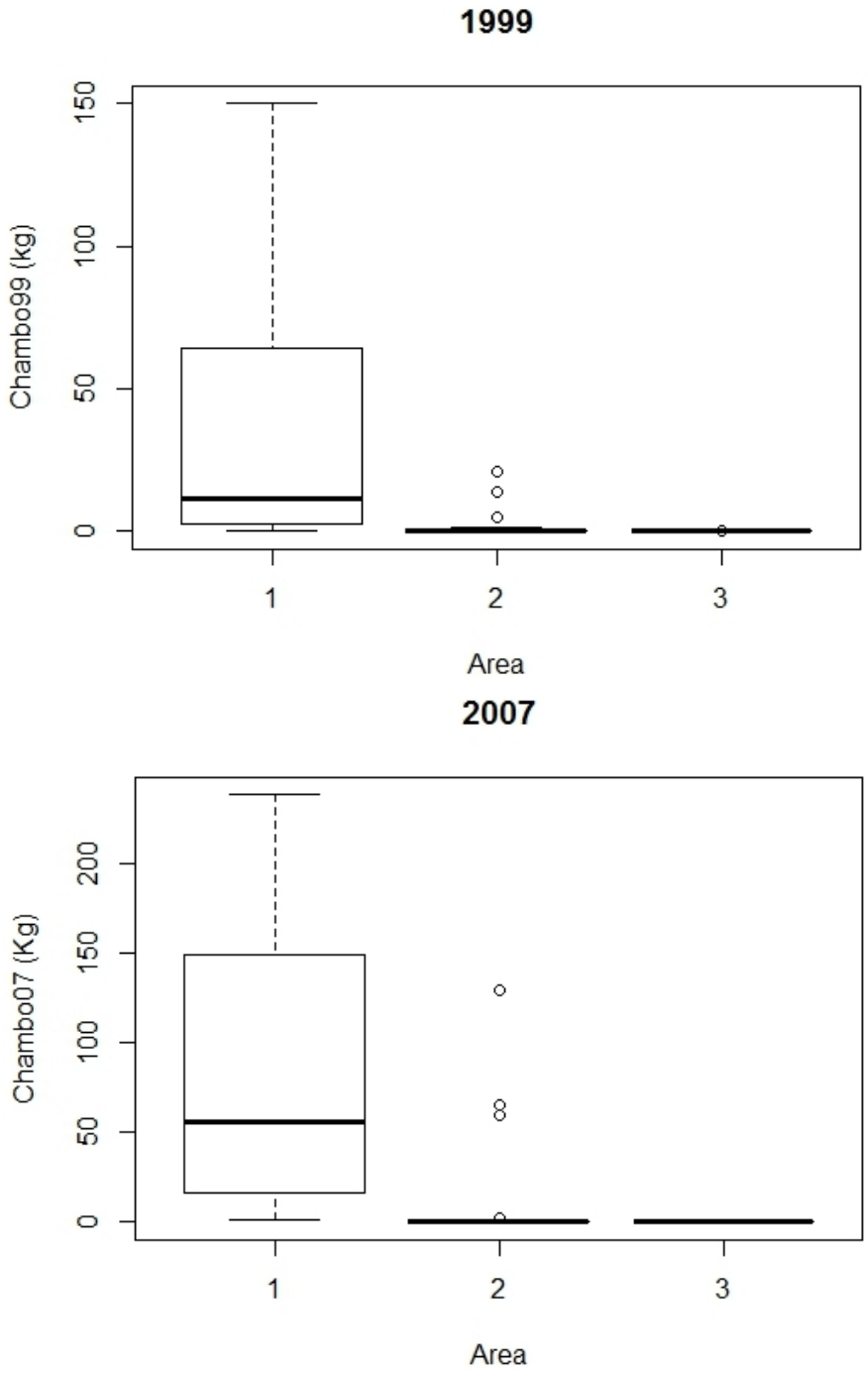

Figure 4.3: Box plots for Chambo distribution and abundance in 1999 and 2007 respectively per area in SEA

Pairs of variables to be explored in the models were also plotted just to check if there is any relationship and dependency in the variables. Figure 4.4 outlines the relationships between the variables. It is clearly seen that as the depth increases, Chambo availability reduces in both years (graph number 1 and 2 after Chambo, second row). Increase in distance from shoreline results in decrease in Chambo availability and abundance (graph number 3 after Chambo, second row). With scatter plot panel smoother, there is a relationship between depth and distance from the pair plots below showing some curve which is an indication of a relationship of an increasing depth as distance increases from shoreline into the water body (graph numbers 3 and 4, last row before Distance).

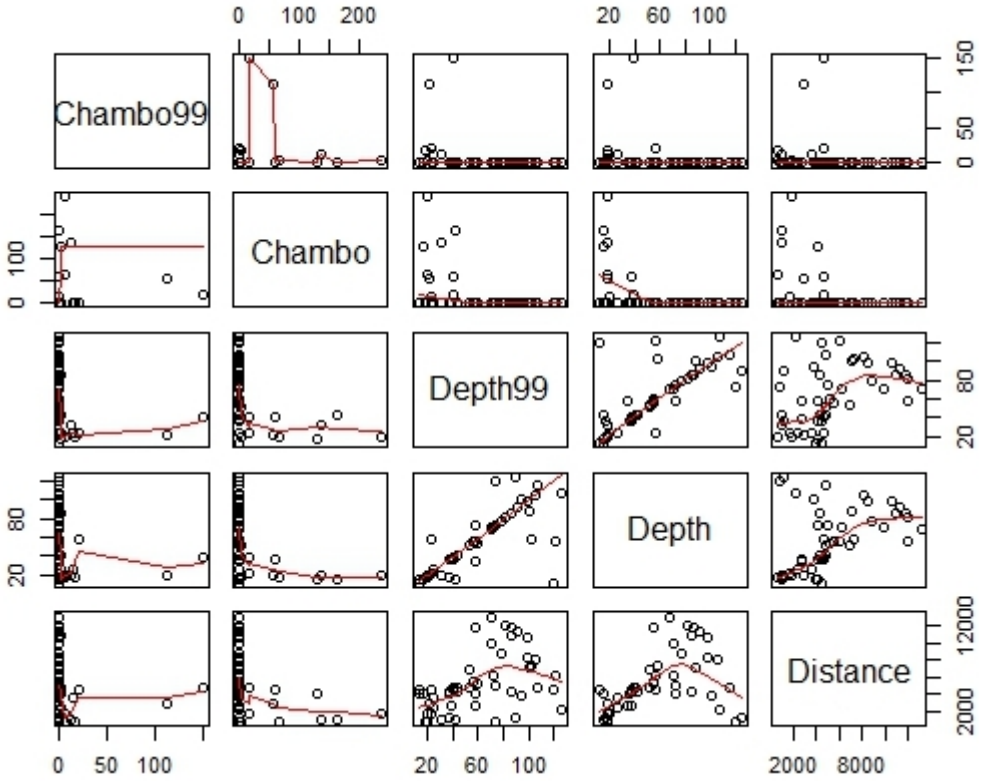

Figure 4.4: Pair plots for variables used in the models to highlight any relationships between the vertical and horizontal variables

However, since distance is from either shoreline, depth is even higher at a short distance in other areas of the lake like in area C and along the center of the lake as represented by the curve which is an indication that it is deeper at the center than close to either shore lines. A close view on the distance and depth plot also shows that as the depth increases, so does the distance though not to be generalized to the whole area. More Chambo in a catch above 200kgs was caught at distances below 5,000m while for depth, most of the Chambo was caught at depth less than 60m in both. The pattern of depth in both years is the same.

4.2 Distribution of Chambo in SEA between 1999 and 2007 compared

Distribution of Chambo was compared using the weight of Chambo realized from the point catches in the years 1999 and 2007. A logistic regression regression was used where the log odds of finding more Chambo in one year was compared with the other year against variables of depth, distance and area as a factor. The t-test (Equation 3.2) showed that there were no significant differences between the depth in the two years ($p-value$ 0.482), an indication that the two means (60.71m and 56.14m from 1999 and 2007 respectively) are not significantly different. Table 4.1 presents the logistic regression model results (Equation 3.3) which were run to assess the presence-absence difference in availability and abundance of Chambo between 1999 and 2007.

Table 4.1: Logistic regression coefficients and odds ratio

Coefficients	*Estimate*	*Std. Error*	*z value*	*Pr(>|z|)*	*Odds Ratio*
(Intercept)	2.338	2.351	0.995	0.320	10.361
Mdepth	-0.130	0.111	-1.164	0.244	0.8785
Distance	0.001	0.0006	1.474	0.140	1.0010
Area B	-2.588	1.769	-1.463	0.143	0.075

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

From the results above, no variable significantly explained the spatial behavior of Chambo at 0.05 level of significant. Despite that, the likelihood of finding Chambo in 1999 for every unit change in depth while holding distance and area fixed is 0.8785 times that of 2007. In other words, the likelihood of finding Chambo in the SEA in 1999 was less by 12.15% as compared to 2007 for every unit change in depth while holding distance and area fixed. This means that 2007 had more Chambo in the SEA than in 1999 and the different management measures instituted on the lake to reverse the problem of overfishing which saw Chambo catches reducing drastically over the years more especially in the SEA (Mvula et al., 2003) can have contributed to this high probability. Otherwise, the increase in chances of finding Chambo in 2007 could be as a result of natural regeneration considering the time that has elapsed in between the years. For area, holding depth and distant constant, there is a less likelihood of finding Chambo in area B of 92.5% as compared to area A, all in favour of 1999 as compared to 2007. On the other hand, there is no likelihood of finding Chambo in area C.

4.3 Modelling Spatial Distribution of Chambo in SEA of Lake Malawi

The study also aimed at modelling the spatial distribution of Chambo using three different models and at the end to find the best model which can better explain the distribution of Chambo. The first statistical model (Equation 3.4) modeled the log odds of finding Chambo given depth and distance as the independent variables. The model produced these results (Table 4.2).

Table 4.2: Logistic regression coefficients and log odds

Coefficients	*Estimate*	*Std. Error*	*z value*	*Pr(>|z|)*	*Log Odds*
(Intercept)	3.009	1.148	2.622	0.009**	20.276
Depth	-0.040	0.017	-2.299	0.022*	0.961
Distance	-0.0005	0.0003	-1.966	0.049*	0.999

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

In assessing the model parameters for significance of the explanatory variables and the whole model, deviance explained was 59.30%. All parameters are significant $(p-values < 0.05)$ for depth and distance while the intercept is highly significant $(p-value < 0.009)$. This then signifies that the presence or absence of Chambo in the areas A, B and C are affected mostly by depth followed by distance. There could also be some other factors which can explain the presence/absence of Chambo like availability of plankton which was not captured. Holding depth and distance constant, the log odds of finding Chambo increases by 202% and decreases with increase in depth and distance.

When the GAM model was run (Equation 3.5) using the ***mgcv*** package, on the same depth and distance variables, now smoothed with the additive nature of a GAM model, the logit operation on the same random sample data set showed the intercept to be significant and of the two smoothed variables, only distance is significant with $(p-value < 0.038)$ (Table 4.3). This model explained 44.8% of the deviance in the presence or absence of Chambo as compared to Model 4.2 whose deviance is higher at 59.3%. Since GAMs are non-parametric extensions of linear model regressions that apply non-parametric smoothers to each predictor variable and additively calculate the response, an anova test for model evaluation gave a statistically significant difference in the models, the additive model describing the relationship between presence/absence of Chambo and depth and

distance much better, by using the Chi-square test for linearity which yielded a value of 0.28.

Table 4.3: GAM model results of Chambo presence/absence and log odds

Parametric coefficients:				
Coefficients	*Estimate*	*Std. Error*	*z value*	*Pr(>|z|)*
(Intercept)	-2.353	0.773	-3.043	0.002**
Approximate significance of smooth terms:				
	edf	*Ref.df*	*Chi.sq*	*p-value*
s(Depth)	1.749	2.194	3.879	0.165
s(Distance)	1.000	1.000	4.305	0.038*

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

The GAM model plots for depth and distance are outlined in Figure 4.5. We see from these figures that odds of the presence of Chambo is highest at depth below 55m and distance below 5,000m and reduces as these parameters increase in magnitude. According to Hastie and Tibshirani (1990), the largest partial residual in the figures as seen is a potentially valuable observation since it corresponds to a presence (1) in a region with very low predicted probability.

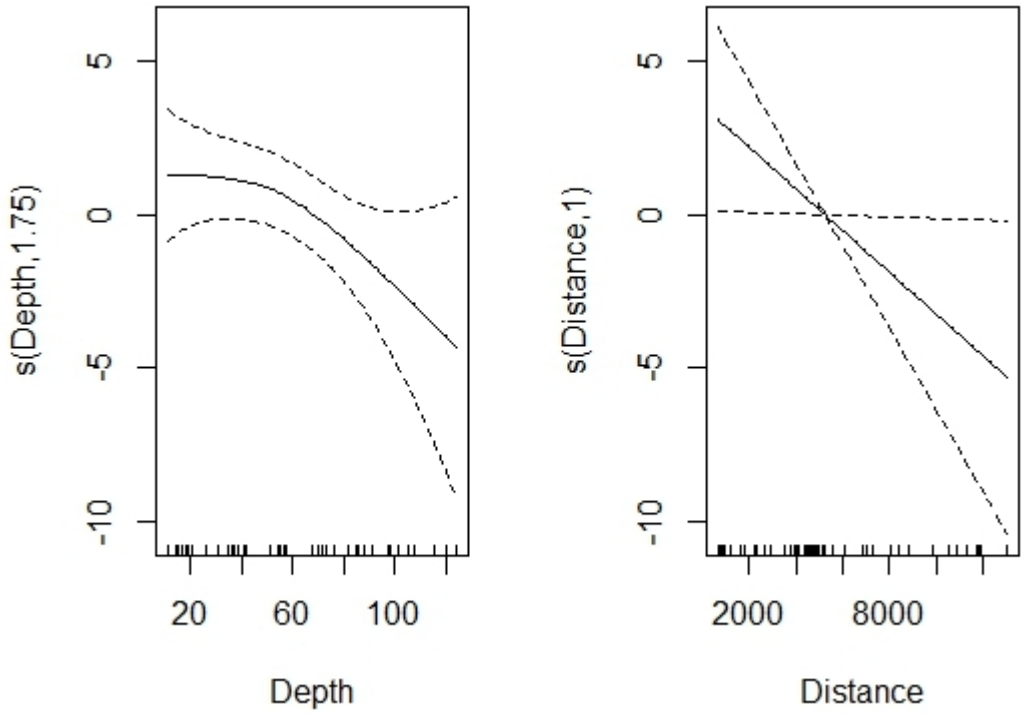

Figure 4.5: Fitted GAM models and their respective variables in explaining the presence / absence of Chambo

Lastly, geographically weighted regression model (Equation 3.6) run on the same variables of depth and distance but with an inclusion of coordinates for location explained more of the deviance as compared to the two models of GLR and GAM, with an R^2 of 62.8, more than the GAM model which had an adjusted R^2 value of 41.4, Table 4.4. The intercept and depth are significant where as distance is not (range of values from minimum to maximum includes a zero). The odds of the presence of Chambo in SEA increased by 45.2% and reduced with increase in depth and distance.

Table 4.4: Summary statistics of the logistic GWR parameter estimates

Variable	*Min*	*1st Quart.*	*Median*	*3rd Quart.*	*Max*	*Global*
(Intercept)	0.081	0.279	0.570	0.940	1.340	0.822
Depth	-0.0248	-0.0086	-0.0041	-0.0018	-0.0005	-0.0059
Distance	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000	0.0000

From the results of the three models on logistic regression, GAM and GWR, GWR had a better AIC and AICc values as compared to GLR and GAM (Table 4.5) as applied to the distribution of Chambo in the SEA of Lake Malawi.

Table 4.5: Comparison of fit for the GLR, GAM and GWR models

Model	n	K_e	*AIC*	AIC_c
GLR	50	3	40.84	41.4
GAM	50	3	40.22	41.0
GWR	50	6.3	18.62	29.84

where: n = number of observations, k_e = effective number of parameters, AIC_c = corrected Akaike Information Criterion

From the results in Table 4.5, results of GLR and GAM are not better as compared to GWR model. The difference in AIC values between GAM and GLR to GWR is 22, which signifies the significant difference in the models. As a rule of thumb, a difference of >3 between AIC values from two competing models are assumed to represent significant differences between them. The AIC is a relative goodness-of-fit statistic for comparing competing models, where the model with the smallest AIC provides the closest approximation to reality.

The value of AIC for GWR (AIC = 18.62) is lower than those of GAM (40.22) and GLR (40.84), indicating that GWR resulted in a significantly better fit for

both variables. Following this therefore, GWR will be used to explain the results of the coefficients from this model, followed by mapping of the coefficients. Effective number of parameters for GWR based models is a more general measure of model complexity unlike in OLS regression which is simply the number of linear coefficients in the model (Fotheringham et al., 2002).

4.4 Geographically Weighted Regression Results

Descriptive statistics for the local parameter coefficients produced by GWR revealed much variance in the parameter value (Table 4.4) confirming the presence of spatial non-stationarity in the relationships between Chambo distribution and the explanatory variables. The depth and distance variables have negative parameter values and the intercept is positive. As the depth and distance increases, chances of finding Chambo decreases.

The mean intercept value (0.569) gave an odds value of 1.77 while for depth it is 0.99. The odds of the presence of Chambo is high in SEA with 77% but decreases with increase in depth. Related probabilistic studies on presence of Chambo for the SEA have not been done, only descriptive studies on its presence presented as percentages. Despite that GWR is a better model for exploring Chambo presence as compared to others, it still can do much better if other factors that influence presence of Chambo are factored in. As Tweddle and Magasa (1989); Bell et al. (2012) reiterated that Chambo catch was related to lake height three years prior and that the productivity of all fish in the lake was related to primary productivity which is a function of the wind velocity and thermal structure of the water column, factors which were not included in the analysis. They further suggested that periods of declining lake heights occurred under conditions which enhanced nutrient upwelling and provided more food for the juvenile Chambo.

However, Bell et al. (2012) noted that there is a change in lake heights from

three years to two years due to a decline of larger fish such that the bulk of the catch were age-1 individuals. The total biomass however is a function of the environmental conditions as well as the anthropogenic factors around the lake. These anthropogenic factors in turn are tied to changes in the climate and the economics of the country. The lake height in turn affects the distance as well as depth at which Chambo is caught. On top of effect of lake heights, Bell et al. (2012) also realized that the main driver of Chambo biomass, however, was fishing pressure which was above the level that would achieve maximum sustainable yield during the entire time series studied (1976 to 2003). Despite lack of other data parameters to explain the relationship, GWR model has proved to be far much better model than GAM and GLR. These results concurs with Windle et al. (2010) who also found that geographically weighted regression was superior in performance as compared to GAMs and GLR.

4.5 Mapping Local GWR Parameters

According to Fotheringham et al. (2002), in GWR, the regression is re–centered many times–on each observation–to produce locally specific GWR parameter results. These local GWR results combined generate a complete map of the spatial variation of the parameter estimates. That is, GWR results, unlike global model results, are mappable and 'given that a very large number of potential parameter estimates can be produced, it is almost essential to map them in order to make some sense of the patterns they display'. Mapping GWR results facilitates interpretation based on spatial context and known characteristics of the study area. The focus is on maps of parameter estimates and t-values as these are the most commonly reported maps in research using GWR. Data classification for t-values should account for certain exogenous criteria that are of importance to the variable being mapped, especially the threshold values that distinguish parameter

estimates that are significant from those that are not (Fotheringham et al., 2002; Mennis, 2006).

The statistical output of GWR software typically includes a baseline global model result (parameter estimates), GWR diagnostic information, a convenient parameter 5- number summary of parameter estimates that defines the extent of the variability in the parameter estimates (the 5-number summary is based on the minimum, lower quartile, median, upper quartile, and maximum local parameter estimates reported in the GWR model – Table 4.4) and Monte Carlo test result for non-stationarity in each parameter. Even with the 5-number summary of parameter estimates and the formal Monte Carlo test, to better understand and interpret non-stationarity in individual parameters it is necessary to visualize the local parameter estimates and their associated diagnostics. GWR models estimate local standard errors, derive local t-statistics, calculate local goodness-of-fit measures (e.g., R^2), and calculate local leverage measures. The output from GWR includes data that can be used to generate surfaces for each model parameter that can be mapped and measured, where each surface depicts the spatial variation of a relationship with the outcome variable (Fotheringham et al., 2002; Mennis, 2006; Matthews and Yang, 2012).

Since depth is significant in the model, parameter estimates were plotted (Figure 4.6) followed by t-values (Figure 4.7) to visualize the variations in depth and their effect in determining the availability of Chambo.

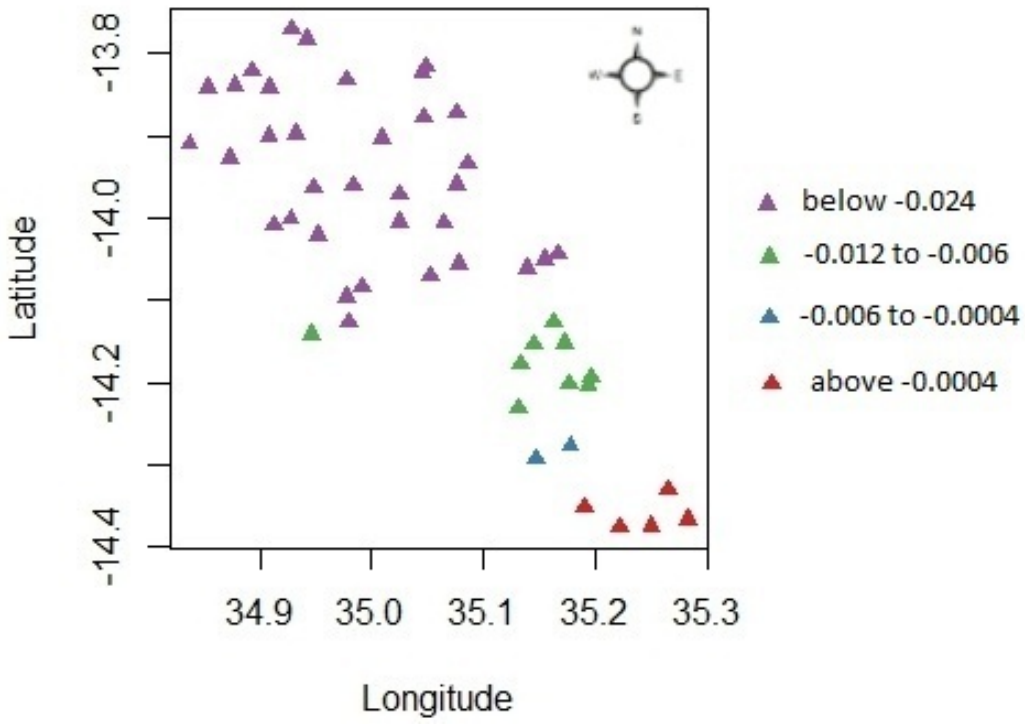

Figure 4.6: Map showing the geographical patterning of the depth parameter estimates

Depth as the only variable significant in the study, is being clearly mapped and from the residuals, the first quartile represents area A and the difference is minimal (0.0026) progressing to the second set which is also getting into area B as a transition between the two areas. The last quartile has residual values for the upper and right side of area C, each quartile sharing the same characteristics in the area of the same dotted colour. There is more variation in the upper quartile (area C) where there is no fish available.

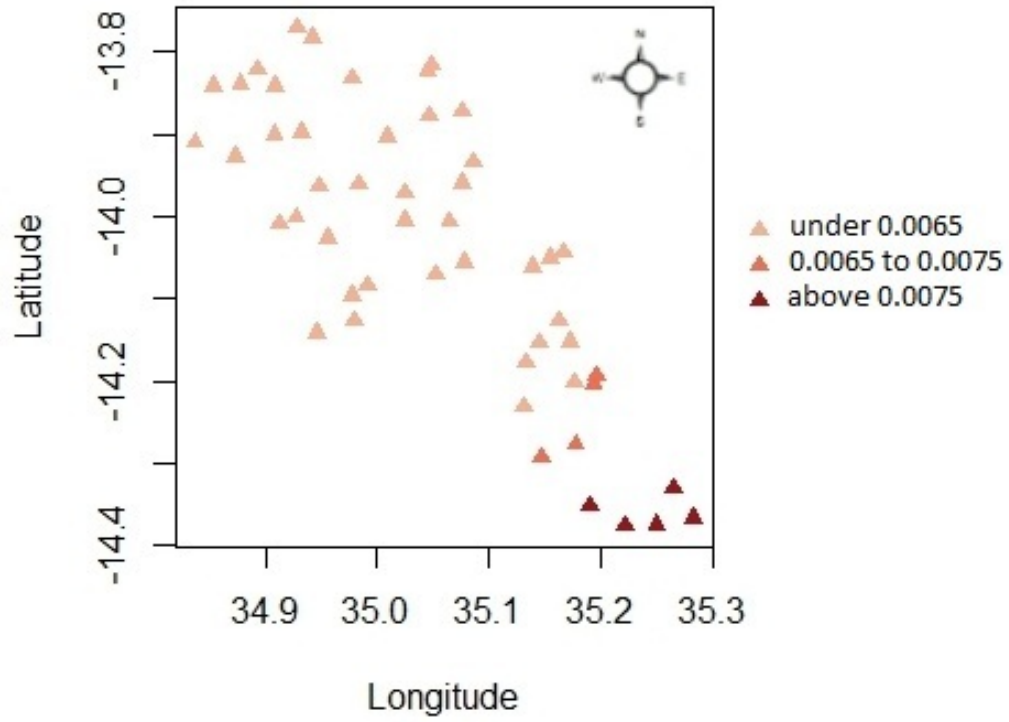

Figure 4.7: Estimated t-values for depth from the GWR model.
Note: Points shown in blue indicate a relationship that is not significant

Figure 4.7 above shows the estimated t values that are not significant, in this case shown in blue while those in orange and green are significant. By referring to the area demarcation, it means areas A and part of B have their t values from the GWR model significant. This means the relationship between the variables in the model and the areas are significant, Chambo available in the far south and not available in the upper most. Though the depth is transforming, with other areas having Chambo and others not, the same shade signifies the non significance of the model coefficient residuals where Chambo is not found. If is was found, the modeling would also show the different shades of colours to signify the spatial non-stationarity of the variable explaining the behavior of the dependent variable.

CHAPTER FIVE

CONCLUSIONS AND RECOMMENDATIONS

5.1 Conclusions

The assumption of spatially stationary processes of ecological relationships within management areas was questioned by studying the distribution of a target fish population, Chambo from SEA of Lake Malawi. The competing models of GLR, GAM and GWR were run to explore the best model that can best model spatial non-stationarity better based on the model with the lowest AIC value. The same Chambo fishery data for 2007 was used for all models plus a comparison of the differences in the probability of finding Chambo in 1999 and 2007, using the odds ratio. There was no significant difference in the probability of finding Chambo either in 1999 or 2007, however, there was a high probability of finding Chambo in 2007 by 12.15%. Results showed that GWR outperformed the other models with a lower AIC value of 18.62 against 40.84 for GLR and 40.22 for GAM. Again, the goodness-of-fit measure from adjusted R^2 explained 62.8% in GWR model against 41.4% from GAM. Depth in this case was significant and was further analyzed by mapping the parameter coefficients and related t-values for visualization of the non-stationarity aspect of the significant variable.

There was more Chambo in 2007 as compared to the year 1999 by percent contribution, as evidenced by the higher probability of 12.15% for finding Chambo in the same year as compared to 1999. Among the competing models used in the analysis of Chambo distribution and its non-stationarity in the variables affecting its presence or absence, geographically weighted regression provided better results as compared to generalized additive models and global logistic regression. Of the parameters from the models used in the study, depth explained better con-

firming the availability and spatial distribution of Chambo which is mostly found at low depth and not in deep waters.

5.2 Recommendations and Areas for Future Research

Apart from the usual depth and coordinates taken on each sampling point, other environmental variables like temperature, primary productivity of the area, wind direction, water current direction and other variables that can best describe the area being sampled can also be captured to help in providing an informed decision once analyzed. These variables can be useful and help improve performance in modelling the abundance and distribution of the different fish species in the lake. Funds permitting, these kind of surveys can be done several times a year (like every 3 or 4 months) so as to capture the seasonality and have a clear picture of fish distribution within a year.

With the revelation done by geographically weighted regression in exposing the local variation in the Chambo–environment relationships under study, it can be used in analyzing fisheries data, more especially in understanding and predicting the spatial dynamics of aquatic ecosystems. An expanded study can therefore be instituted to explore the full time-series of Chambo data, now for the whole of Lake Malawi using GWR and additional abiotic and biotic variables. The same can also be done on different fish species of interest, either at a point in time or in time-series aspect to unveil the significant factors and their spatial distribution so as to help in management decisions of the fishery resource under study.

REFERENCES

Akaike, H., 1973. Information theory and an extension of the maximum likelihood principle. Budapest: Akademiai Kiado.

Anselin, L., 1990. Spatial dependence and spatial structural instability in applied regression analysis. Journal of Regional Science 30, 185–207.

Bailey, M. C., Maravelias, C. D., Simmonds, E. J., 1998. Changes in the spatial distribution of autumn spawning herring (*Clupea harengus L.*) derived from annual acoustic surveys during the period 1984-1996. ICES Journal of Marine Science 55, 545–555.

Banda, M., Jamu, D., Njaya, F., Makuwila, M., Maluwa, A., 2005a. The chambo restoration strategic plan. Tech. rep., Department of Fisheries Proceedings of a national workshop, 13–16 May, 2003, Mangochi.

Banda, M., Kanyerere, G., Rusuwa, B., 2003. The status of the chambo in malawi: Fisheries and biology. In: The Chambo Restoration Strategic Plan.

Banda, M., Kanyerere, G., Rusuwa, B., 2005b. The status of the chambo in malawi: Fisheries and biology. In: The Chambo Restoration Strategic Plan.

Beare, D. J., Reid, D. G., Petitgas, P., 2002. Spatio-temporal patterns in herring (*Clupea harengus L.*) school abundance and size in the northwest north sea: modelling space-time dependencies to allow examination of the impact of local school abundance on school size. ICES Journal of Marine Science 59, 469–479.

Bell, R., Collie, J., Jamu, D., Banda, M., 2012. Changes in the biomass of chambo in the southeast arm of lake malawi: A stock assessment of *Oreochromis spp.* Journal of Great Lakes Research.

Bez, N., 2002. Global fish abundance estimation from regular sampling: a geostatistical transitive method. Can. J. Fish. Aqua. Sci. 59, 1921–1931.

Bi, H., Ruppel, R., Peterson, W., 2007. Modeling the pelagic habitat of salmon off the pacific northwest (usa) coast using logistic regression. Mar Ecol Prog Ser 336, 249–265.

Bigelow, A. K., Boggs, C. H., He, X., 1999. Environmental effects on swordfish and blue sharks catch rates in the us north pacific longline fishery. Fisheries Oceanography 8, 178–198.

Brunsdon, C., Fotheringham, A., Charlton, M., 1996. Geographically weighted regression: a method for exploring spatial nonstationarity. Geographical Analysis 28, 281–298.

Bulirani, A., 2005. Observation on the factors behind the decline of the chambo in lake malawi and lake malombe. In: The Chambo Restoration Strategic Plan. WorldFish Center Conference Proceedings 71, 112 p.

Carvalho, F., Murie, D., Hazin, F., Hazin, H., Leite-Mourato, B., Burgess, G., 2011. Spatial predictions of blue shark (*Prionace glauca*) catch rate and catch probability of juveniles in the southwest atlantic. ICES Journal of Marine Science 68(5), 890 – 900.

Chandra, H., Salvati, N., Chambers, R. a Tzavidis, N., 2010. Small area estimation under spatial nonstationarity. Centre for Statistical and Survey Methodology, University of Wollongong, Working Paper 21 10, 33p, http://ro.uow.edu.au/cssmwp/71.

Charlton, M., Fotheringham, S., 2009. Geographically weighted regression, white paper. Tech. rep., National Centre for Geocomputation, National University of Ireland, Maynooth.

Chong, Y. S., Wang, J. L., 1997. Statistical modeling via dimension reduction methods. Nonlinear Analysis, Theory, Methods and Applications 30, 3561–3568.

Ciannelli, L., Bailey, K., Chan, K., Stenseth, N., 2007. Phenological and geographical patterns of walleye pollock spawning in the western gulf of alaska. Can. J. Fish. Aqua. Sci. 64, 713–722.

Ciannelli, L., Fauchald, P., Chan, K.S. Agostini, V., Dingsor, G., 2008. Spatial fisheries ecology: recent progress and future prospects. Journal of Marine Systems 71, 223 – 236.

Cleveland, W. S., Delvin, S. J., 1988. Locally weighted regression: an approach to regression analysis by local fitting. Journal of the American Statistical Association 83, 596 – 610.

Cox, . R., 1970. Analysis of Binary Data. Meihuen.

Cressie, N., 1993. Statistics for Spatial Data. New York: John Wiley and Sons.

Damalas, D., Megalofonou, P., Apostolopoulou, M., 2007. Environmental, spatial, temporal and operational effects on swordfish (*Xiphias gladius*) catch rates of eastern mediterranean sea longline fisheries. Fisheries Research 84, 233–246.

Denis, V., Lejeune, J., Robin, J., 2002. Spatial–temporal analysis of commercial trawlers data using general additive models: patterns of loliginid squid abundance in the north–east atlantic. ICES Journal of Marine Science 59, 633–648.

Diggle, P., 1990. A point process modeling approach to raised incidence of a rare phenomenon in the vicinity of a prespecified point. J. Royal Stat. Soc. A, 153, 349– 62.

Ellis, J., Ysebaert, T., Hume, T., Norkko, A., Bult, T., Herman, P., Thrush, S., Oldman, J., 2006. Predicting macrofaunal species distributions in estuarine gradients using logistic regression and classification systems. Marine Ecology Progress Series 316, 69–83.

Fauchald, P., Erikstad, K., Skarsfjord, H., 2000. Scale–dependent predator–prey interactions: the hierarchical spatial distribution of seabirds and pray. Ecology 81, 773–783.

Fortin, M.-J., Dale, M., 2005. Spatial Analysis: a Guide for Ecologists. Cambridge university Press, Cambridge, UK.

Fotheringham, A., 1997. Trends in quantitative methods i: stressing the local. Progress in Human Geography 21, 88–96.

Fotheringham, A., Brunsdon, C., Charlton, M., 2002. Geographically Weighted Regression: The analysis of spatially varying relationships. John Wiley and Sons Ltd, England.

Fox, C., O'Brien, C., Dickey-Collas, M., Nash, R., 2000. Patterns in the spawning of cod (*Gadus morhua* l.), sole (*Solea solea* l.) and plaice (*Pleuronectes platessa* l.) in the irish sea as determined by generalized additive modeling. Fish Oceanogrgaphy 9, 33–49.

Francis, M. P., Morrison, M., Leathwick, J., Walsh, C., Middleton, C., 2005. Predictive models of small fish presence and abundance in northern new zealand harbours. Estuarine, Coastal and Shelf Science 64, 419–435.

Giannoulaki, M., Machias, A., Valavanis, V., Somarakis, S., Palialexis, A., Papaconstantinou, C., 2006. Spatial modelling of the european anchovy habitat in the eastern mediterranean basin using gams and gis technology. In: General Fisheries Commission for the Mediterranean Scientific Advisory Commit-

tee, Sub-Committee for Stock Assessment Working Group on Small Pelagic Species, FAO, Rome, 11–14 September 2006.

Grift, R., Rijnsdorp, A., Barot, S., Heino, M., Dieckmann, U., 2003. Fisheries-induced trends in reaction norms for maturation in north sea plaice. Mar 257, 247–257.

Gujarati, D., 2004. Basic Econometrics. The MacGraw-Hill Companies.

Hastie, T., Tibshirani, R., 1990. Generalized Additive Models. Chapman & Hall.

Hosmer, D., Lemeshow, S., 1989. Applied Logistic Regression. John Wiley and Sons, New York.

Jensen, O., Seppelt, R., Miller, T., Bauer, L., 2005. Winter distribution of blue crab (*Callinectes sapidus* in chesapeake bay: application and croos–validation of a two–stage generalized additive model. Marine Ecology Progress Series 299, 239–255.

Jones, K., Devillers, R., Edinger, E., 2009. Relationships between cold-water corals off newfoundland and labrador and their environment, all authors from Memorial University of Newfoundland, Canada.

Kachinjika, O., 2001. A general overview of fisheries research and development in malawi. Tech. rep., Department of Fisheries, Lake Malawi Fisheries Management Symposium, 4–9 June 2001, Lilongwe.

Kanyerere, G., Booth, A., 1999. Spatial and temporal distribution of some commercially important fish species in the southeast arm of lake malawi: A geostatistical analysis. Tech. rep., Department of Fisheries, Lake Malawi Fisheries Management Symposium, 4–9 June 2001, Lilongwe.

Kimsey, M. J., Moore, J., McDaniel, P., 2008. A geographically weighted regression analysis of douglas-fir site index in north central idaho. Forest Science 54, 356–366.

Kupfer, J. A., Farris, C. A., 2007. Incorporating spatial nonstationarity of regression coefficients into predictive vegetation models. Landscape Ecology 22, 837–852.

Kupschus, S., 2003. Development and evaluation of statistical habitat suitability models: an example based on juvenile spotted seatrout (*Cynoscion nebulosus*. Marine Ecology Progress Series 265, 197–212.

Latif, A., Hossain, M., Islam, M., 2008. Model selection using modified akaike information criterion: An application to maternal morbidity data. Austrian Journal of Statistics 37(2), 175–184.

Lehmann, A., Overton, J. M., Leathwick, J. R., 2002. Grasp: generalized regression analysis and spatial prediction. Ecological Modelling 157, 189 – 207.

Leung, Y., Mei, C.-L., Zhang, W.-X., 2000. Statistical tests for spatial nonstationarity based on the geographically weighted regression model. Environment and Planning A 32(1), 9–32.

Likongwe, J., 2005. A preliminary study on biodiversity of riverine fishes in malawi and their aquaculture potential. In: African Crop Science Conference Proceedings. Vol. 7. pp. 1293–1296.

Maravelias, C. D., Reid, D. G., Swartzman, G., 2000. Seabed substrate, water depth and zooplankton as determinants of the prespawning spatial aggregation of north atlantic herring. Marine Ecology Progress Series 195, 249–259.

Maravelias, C. D., Reid, D. G., Swartzman, G., 2000a. Modelling spatio-temporal effects of environment on atlantic herring, (*Clupea harengus L.*). Environmental Biology of Fishes 58, 157–172.

Matern, B., 1986. Spatial Variation [Early book providing mathematical detail for correlation functions]. New York: Springer.

Matthews, S., Yang, T.-C., 2012. Mapping the results of local statistics: Using geographically weighted regression. Demographic Research 26 (6), 151–166.

McCullagh, P., Nelder, J., 1989. Generalized Linear Models (2nd Edition). Chapman and Hall, New York, USA.

McMillen, D., 1996. One hundred fifty years of land values in chicago: a non-parametric approach. Journal of Urban Economics 40, 100–124.

McMillen, D., McDonald, J., 1997. A nonparametric analysis of employment density in a polycentric city. Journal of Regional Science 37, 591–612.

Mello, L., Rose, 2005. Usingg geostatistics to quantify seasonal distribution and aggregation patterns of fishes: an example of atlantic cod (*Gadus morhua*). Can. J. Fish. Aqua. Sci. 62, 659–670.

Mennis, J., 2006. Mapping the results of geographically weighted regression. The Cartographic Journal 43 (2), 171–179.

Moore, C., Harvey, E., Van Niel, K., 2009. Spatial prediction of demersal fish distributions: enhancing our understanding of species–environment relationships. ICES Journal of Marine Science 66, 2068 – 2075.

Murase, H., Nagashima, H., Yonezaki, S., Matsukura, R., Kitakado, T., 2009. Application of a generalized additive model (GAM) to reveal relationships between environmental factors and distributions of pelagic fish and krill: a

case study in sendai bay, japan. ICES Journal of Marine Science 66, 1417–1424.

Mvula, P., Njaya, F., Nkoko, B., 2003. Socio-economic factors influencing the decline of the chambo. In: The Chambo Restoration Strategic Plan.

Narumalani, S., Jensen, J. R., Althausen, J. D., Burkhalter, S., Mackey Jr., H. E., 1997. Aquatic macrophyte modeling using gis and logistic multiple regression. Photogrammetric Engineering & Remote Sensing 63, 41–49.

Osborne, P. E., Foody, G. M., Suaé rez Seoane, S. ., 2007. Non-stationarity and local approaches to modelling the distributions of wildlife. Diversity and Distributions 13, 313–323.

Palsson, O., Bulirani, A., Banda, M., 1999. A review of biology, fisheries and population dynamics of chambo (*Oreochromis spp.*, cichlidae) in lakes malawi and malombe. Tech. rep., Department of Fisheries, Bulletin No. 38, Lilongwe.

Petitgas, P., 2001. Geostatistics in fisheries survey design and stock assessment: models, methods and applications. Fish Fish. 2, 231–249.

Pratesi, M., Salvati, N., 2008. Small area estimation: the eblup estimator based on spatially correlated random area effects. Statistical Methods and Applications 17, 114–131.

Ricklefs, R., 1990. In-exploring spatial non–stationarity of fisheries survey data using geographically weighted regression (gwr): an example from northwest atlantic. International Council for the Exploration of the Sea, Oxford Journals 167, 145–154.

Rodríguez, C., 2005. The abc of model selection: Aic, bic and the new cic. In: AIP Conference Proceedings. Vol. 803. p. 80.

Rose, G. A., 2005. On distributional responses of north atlantic fish to climate change. ICES Journal of Marine Science 62, 1360–1374.

Sanchez, F., Gil, J., 2000. Hydrographic mesoscale structures and poleward current as a determinant of hake (*Merluccius merluccius*) recruitment in southern bay of biscay. ICES Journal of Marine Science 57, 152–170.

Schwarz, G., 1978. Estimating the dimension of a model. Ann. Statistics 6, 461–464.

Seymour, T., 2001. Fisheries development, management, and the role of government. Tech. rep., Department of Fisheries, Lake Malawi Fisheries Management Symposium, 4–9 June 2001, Lilongwe.

Swartzman, G., 1997. Analysis of the summer distribution of fish schools in the pacific eastern boundary current. ICES Journal of Marine Science 54, 105–116.

Swartzman, G., Huang, C., Kalzuny, S., 1992. spatial analysis of the bering sea groundfish survey data using generalilzed additive models. Canadian Journal of Fisheries and Aquatic Science 49, 1366–1378.

Swartzman, G., Silverman, E., Williamson, N., 1995. Canadian journal of fisheries and aquatic sciences. Relating trends in walleye pollock (Theragra chalcogramma) abundance in the Bering Sea to environmental factors. 52, 369–380.

Swartzman, G., Stuetzle, W., Kulman, K., Powojowski, M., 1994. Relating the distribution of pollock schools in the bering sea to environmental factors. ICES Journal of Marine Science 51, 481–492.

Tobler, W., 1970. A computer movie simulating urban growth in the detroit region. Economic Geography 46(2), 234–240.

Turner, G., 1996. Offshore Cichlids of Lake Malawi. Cichlid Press, Lauenau.

Turner, G., Witimani, J., Robinson, R., Grimm, A., Pitcher, T., 1991. Reproductive isolation and nest sites of lake malawi chambo, oreochromis (nyasalapia) spp. Fish Biology 39, 775–782.

Tweddle, D., Magasa, J., 1989. Assessment of multispecies cichlid fisheries of the southeast arm of lake malawi, africa. International Council for the Exploration of the Sea 45(2), 209–222.

Valavanis, V., Pierce, G., Zuur, A., Palialexis, A., Saveliev, A., Katara, I., Wang, J., 2008. Modelling of essential fish habitat based on remote sensing, spatial analysis and gis. Hydrobiologia 612, 5– 20.

Van Zalinge, N., Alimoso, S., Donda, S., Mdaihli, M., Seisay, M., Turner, G., 1991. Preliminary note on the decline of the chambo catches in lake malombe. Tech. rep., Department of Fisheries, Government of Malawi, UNDP and FAO. FI:DP/MLW/86/013, Field Document 9.

Walsh, W. A., Kleiber, P., McCracken, M., 2002. Comparison of logbook reports of incidental blue shark catch rates by hawaii-based longline vessels to fishery observer data by application of a generalized additive model. Fisheries Research 58, 79–94.

Wieland, K., Rivoirard, J., 2001. A geostatistical analysis of ibts data for age 2 north sea haddock (*Melanogrammus aeglefinus*) considering daylight effects. Sarsia 86, 503–516.

Windle, M. J. S., Rose, G. A., Devillers, R., Fortin, M.-J., 2010. Exploring spatial non-stationarity of fisheries survey data using geographically weighted regression (GWR): an example from the northwest atlantic. ICES Journal of Marine Science 67, 145–154.

Wood, S. N., 2006. Generalized Additive Models: an Introduction with R. Chapman and Hall, CRC Press, Boca Raton.

Wood, S. N., Augustin, N., 2002. Gams with integrated model selection using penalized regression splines and integrated model selection using penalized regression splines. Ecological Modelling 157, 157–177.

Yang, 2003. Can the strengths of aic and bic be shared? Tech. rep., Iowa State University, Department of Statistics, downloaded on 25th January 2012 from: www.stat.iastate.edu/preprint/articles/2003-10.pdf.

Yee, T., Mitchell, N., 1991. Generalized additive models in plant ecology. j. veg. sci.,. J. Veg. Sci. 2, 587–602.

Zagaglia, C., Lorenzetti, J., Stech, L., 2004. Remote sensing data and longline catches of yellowfin tuna (*Thunnus albacore* in the equatorial atlantic. Remote Sensing and Environmental 93, 267–281.

APPENDICES

Appendix A: R Codes Run in the Study

```
## The data for the study
>f< read.csv("Fish.csv",header=T)
>attach(f)
>names(f)
## T test for 1999 and 2007 depth comparison if
significantly different
>t.test(Depth,Depth99)
## Descriptive for the variables
>par(mfrow=c(1,1))
>boxplot(Depth~Area,main="2007",xlab="Area",
                    ylab="Depth07 (m)")
>boxplot(Depth99~Area,main="1999",xlab="Area",
                    ylab="Depth99 (m")
>boxplot(Distance~Area,main="",
                    xlab="Area",ylab="Distance (m)")
>boxplot(Chambo~Area,main="2007",xlab="Area",
                    ylab="Chambo07 (Kg)")
>boxplot(Chambo99~Area,main="1999",xlab="Area",
                    ylab="Chambo99 (kg)")
>pairs(Chambo99~Chambo+Depth99+Depth+Distance)

## Running models for analysis
# 1a. Odds Ratio Logistic Regression for Comparison
# between 1999 and 2007
>glrc < glm((CP99/CP07)~Mdepth+Distance+Area,data=f,
```

```
                    family=binomial)
>summary(glrc)
>coefs < round(exp(glrc$coef),digits=6)
>coefs AIC(glrc)

# 1b. Logistic Regression with factor (Area)
>glr < glm(CP07~Depth+Distance+Area,
            data=f,family=binomial)
>summary(glr)
>coefs < round(exp(glr$coef),digits=6)
>coefs full.sum< summary(glr)$coef full.sum
>plot(glr)
>plot(glr,res=T)
>AIC(glr)

# 2. Binomial GAM
>library(mgcv)
>bgam< gam(CP07~s(Depth)+s(Distance)+Area,
          data=f,binomial)
>summary(bgam)
>coefgam < round(exp(bgam$coef),digits=6)
>coefgam
>list(coefgam)
>data.frame(coefgam)
>plot(bgam,pages=1)
>par(mfrow=c(1,2))
>gam.check(bgam)
```

```
>plot.gam(bgam,residuals=T,pch=16,all.terms=T)
>AIC(bgam)

# 3. Binomial GWR
### Map for plotting
>library(MASS)
>sea< read.csv("SEAMapPoints.csv")
>attach(sea)
>head(sea)
>smap< eqscplot(Longitude,Latitude,type="l",xlab=
        "Longitude",ylab="Latitude",main="")

### Plotting residuals in a map
>library(spgwr)
>library(RColorBrewer)
>library(shapefiles)
>library(maptools)
>f< read.csv("Fish.csv",header=T)
>attach(f)
>cord1< cbind(f$X,f$Y)
>head(cord1)
>fdata< cbind(f,cord1)
>head(fdata) farea< factor(f$Area)

## using fixed weight supported by Cross validation
>best.bw1< gwr.sel(CP07~Depth+Distance,
                    data=fdata,coords=cord1)
```

```
>best.bw1
>gwr.model1 < gwr(CP07~Depth+Distance , data=fdata ,
                        coords=cord1 , bandwidth=best.bw1 ,
              se.fit=TRUE, hatmatrix=TRUE)
>gwr.model1  summary(gwr.model1)  head(gwr.model1$SDF)

## Mapping the results
## Depth coeficients following
>gwr.model1
>Depth.coefs1 < gwr.model1$SDF$Depth
>round(4)
>fivenum(Depth.coefs1 ,4)
>min(Depth.coefs1)
>max(Depth.coefs1)
>colour.catg < findInterval(Depth.coefs1 ,
           c( 0.0240, 0.0181, 0.0122, 0.0063, 0.0004),
                      all.inside=T)
>pallete < brewer.pal(5,'Set1')
>depth.colour < pallete[colour.catg]
>par(mfrow=c(1,1),pty="s")
>plot(cord1 ,col=depth.colour ,smap,pch=17,main="",
                    xlab="Longitude",ylab="Latitude")

## for t values
>Depth.se < gwr.model1$SDF$Depth_se
>round(4)
>fivenum(Depth.se ,4)
```

```
>min(Depth.se)
>max(Depth.se)
>colour.catg< findInterval(Depth.se,c(0.00,0.0015,
            0.0030,0.0045,0.0060,0.0075),all.inside=T)
>pallete< brewer.pal(5,'Reds')
>depth.colour< pallete[colour.catg]
>par(mfrow=c(1,1),pty="s")
>plot(cord1,col=depth.colour,smap,pch=17,
                    main="",xlab="Longitude",ylab="Latitude")
```