Remote sensing of reef fish communities

by

Anders Jensen Knudby

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Geography

Waterloo, Ontario, Canada, 2009

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

ABSTRACT

During the last three decades of coral reefs studies, the large areal coverage of data derived from satellite images has increasingly been used to complement the more detailed but spatially limited data produced by conventional fieldwork. Continuous improvement in sensor capabilities, along with the development of increasingly refined methods for image processing, has lead to ever more accurate maps of physical and biological variables of importance to reef ecology.

During the same period, an abundance of field studies have documented statistical relationships between aspects of the reef habitat and its fish community. Despite numerous stochastic influences, such as spatially concentrated and temporally variable fish recruitment pulses or the selective and patchy mortality caused by fishing, several aspects of habitat have been shown to significantly influence the fish community. Fortunately the most important of these, water depth, the structural complexity of the reef, and the cover of live coral, are possible to estimate from currently available satellite imagery.

The research presented in the following pages has combined the statistical relationships between the fish community and its habitat with the capability of satellite imagery to map that habitat, thereby answering the research question:

How can remote sensing be used to map coral reef fish communities?

In the process, a set of new techniques for predictive modeling of complex relationships have been compared, the influence of a range of habitat variables on the fish community quantified, the spatial scales at which the fish-habitat relationships are strongest have been explored, and new methods for deriving estimates of some aspects of the coral reef habitat from satellite imagery have been developed. The results presented in this thesis thus contribute to the further understanding of fish-habitat relationships, while providing a template for producing spatially explicit predictive models of fish community variables. This is not only of scientific interest, but also of substantial value to the conservation community that tries to protect the world’s remaining healthy coral reef ecosystems, and their fish communities, from an array of man-made influences.

ACKNOWLEDGEMENTS

A PhD cannot be done without a little help from your friends, and I would therefore like to express my sincere thanks to a number of people who have provided assistance of one kind or another along the way.

Before I even began the work, Jeremy Kemp and Steffan Howe infected me with their passion for coral reefs and the animals that live on them, as I learned from both of them while working in Eritrea. Michael Schultz Rasmussen, my M.Sc. thesis supervisor, similarly made me appreciate the capabilities of remote sensing and spatial analysis of natural resources. The combination of these skills and interests set me on course for the PhD work.

Once started at the University of Waterloo (UW), my supervisor Ellsworth LeDrew and his grad students Alan Lim and Candace Newman provided valuable strategic advice along with numerous opportunities for fieldwork and access to funding (CIDA, NSERC, UW). Useful advice was also provided along the way by my committee members Richard Kelly, David Barton, and last but not least Alexander Brenning who provided invaluable statistics and programming support. A surprising number of external academics were also willing to share their time and thought with me along the way, including Sam Purkis, Simon Pittman, Peter Mumby, Peter Sale and Michael Risk.

I also received support from several people during my fieldwork in Zanzibar, including Frida Lanshammar, Omari Nyange and the rest of the fabulous staff at Chumbe Island, Christopher Muhando at the Institute of Marine Science, and Lina Nordlund from the Stockholm University.

Furthermore, I would like to thank my family for providing moral support and perspective during the 4 years this PhD has been underway. A special thank you to my older brother Christen who blazed the PhD trail and shared his experiences with me, and to my younger brother Søren for not completing his PhD before me. Finally, many thanks are due to Candace Newman for providing lots of solid advice and for standing by me even when I went on fieldwork for several months during difficult times.

TABLE OF CONTENTS

List of Tables ix

List of Figures xi

List of Acronyms xiv

A note on Terminology xv

CHAPTER 1 : INTRODUCTION 1

1.1 Coral reefs in decline 1

1.2 MPAs and the use of remote sensing 2

1.3 Sub-questions 3

1.4 Research question 4

1.5 Thesis structure 4

CHAPTER 2 : RESEARCH CONTEXT 5

2.1 Global distribution of coral reefs and associated fishes 5

2.2 Reef-scale distribution of biodiversity 6

2.2.1 Fish-habitat relationships on coral reefs 7

2.2.2 The statistical nature of fish-habitat relationships 8

2.3 Remote sensing of coral reefs 10

2.4 Mapping habitat variables 11

2.4.1 Remote sensing of depth 11

2.4.2 Remote sensing of live coral cover 12

2.4.3 Remote sensing of structural complexity 13

2.4.4 Remote sensing of habitat variety 13

2.5 Mapping fish biodiversity 14

2.6 Summary 15

CHAPTER 3 : STUDY AREA 16

3.1 East Africa and the Western Indian Ocean 16

3.2 Zanzibar 18

3.2.1 Fishing 18

3.2.2 Conservation 18

3.3 Two reefs, one protected and one unprotected 19

3.3.1 Chumbe 19

3.3.2 Bawe 20

3.4 Summary 21

CHAPTER 4 : DATA AND RESEARCH METHODS 22

4.1 Research approach 22

4.2 *In situ* data 22
4.2.1 Fish data 23
4.2.2 Habitat data 27
4.2.3 Note on geolocation 28
4.3 Remote sensing data 29
4.3.1 Satellite imagery 29
4.3.2 Processing 30
4.3.2.1 Datum, projection, and geometric correction 31
4.3.2.2 Atmospheric correction 31
4.3.2.3 Deglinting 32
4.3.2.4 Water column correction 32
4.3.3 Substrate classifications 33
4.3.4 Determination of geomorphologic zones 36
4.3.5 Estimation of live coral cover 36
4.3.6 Estimation of depth and rugosity 38
4.3.7 Estimation of habitat variety 39
4.4 Predictive models 42
4.4.1 Linear Model (LM) 42
4.4.2 General Additive Model (GAM) 42
4.4.3 Tree-based models 43
4.4.3.1 Bagging 44
4.4.3.2 Random forest 44
4.4.3.3 Boosted trees 44
4.4.4 Support Vector Machine (SVM) 45
4.5 Model comparison 45
4.5.1 Resampling methods 45
4.5.1.1 Bootstrapping 46
4.5.1.2 Cross-validation 46
4.5.2 Spatial autocorrelation 47
4.6 Identifying important habitat variables 47
4.7 Determining the characteristic scale of response 48
4.8 Summary 50
CHAPTER 5 : FISH-HABITAT RELATIONSHIPS 51
5.1 Fish communities and habitats on Chumbe and Bawe 51
5.1.1 Fish communities 51
5.1.2 Habitats 52
5.2 Prediction of the fish community from habitat variables 53

5.2.1 Model accuracies 53
5.2.1.1 Description of model differences 53
5.2.1.2 Tests of difference between model prediction accuracies 57
5.2.2 Model precision 57
5.2.3 Finding the important habitat variables 58
5.3 Discussion 61
5.3.1 Model types 61
5.3.2 Important variables and model types 62
5.3.2.1 The influence of conservation status 64
5.3.2.2 Comparison with findings from other studies 65
5.4 Summary 67
CHAPTER 6 : REMOTE SENSING OF HABITAT VARIABLES 68
6.1 Remote sensing of live coral cover 71
6.1.1 Accuracy of IKONOS-based live coral cover estimates 74
6.1.1.1 Assessing the classification accuracy 74
6.1.1.2 Assessing the accuracy of live coral cover estimations 75
6.1.2 Live coral cover estimates and their spatial variation 76
6.2 Remote sensing of depth and rugosity 78
6.2.1 Rugosity estimates and their spatial variation 78
6.3 Remote sensing of habitat variety 81
6.3.1 Habitat variety estimates and their spatial variation 82
6.3.1.1 Comparison of the variety measures 82
6.4 Discussion 86
6.4.1 Remote sensing of live coral cover 86
6.4.2 Remote sensing of depth and rugosity 87
6.4.3 Remote sensing of habitat diversity 87
6.4.4 Characteristic scale of the relationship between rugosity and fish species richness 88
6.4.5 Characteristic scale of the relationship between habitat variety and fish species richness 90
6.4.6 The targeted landscape approach 91
6.5 Summary 91
CHAPTER 7 : REMOTE SENSING OF REEF FISHES 93
7.1 Comparison of predictive model accuracy by data set 94
7.1.1 Models predicting species richness 96
7.1.2 Models predicting biomass 96
7.1.3 Models predicting diversity 98

7.2 Variable importance results 99
7.2.1 Important variables from the IKONOS data set 99
7.2.2 Remotely sensed variables with complementary information from the "All" data set 101
7.3 Map of predicted species richness, Chumbe Island 105
7.4 Discussion 105
7.4.1 Predictive performance of IKONOS-based models 107
7.4.2 Variable importance in IKONOS-based models 108
7.4.2.1 Rugosity 108
7.4.2.2 Live coral cover 109
7.4.2.3 Habitat diversity 109
7.4.3 Complementarity of remotely sensed information 110
7.5 Summary 110
CHAPTER 8 : CONCLUSION 112
Bibliography 117

LIST OF TABLES

Table 4.1: List of variables derived from field data. Log-transformations indicated by *. 41
Table 4.2: List of *in situ* and remotely sensed habitat variables excluded from GAM models. 43
Table 4.3: List of variables used for investigations of the characteristic scale of response. 49
Table 4.4: List of variables used for comparison of remotely sensed and *in situ* habitat variables. 49
Table 5.1: A comparison of average biomass values (g/100 m^2) of sites on the two reefs, for commonly fished species. The top six families are all commonly fished. Serranidae and Balistidae are families with desirable target species, though their current rarity means that they constitute a small percentage of the local fishery. 51
Table 5.2: Summary of minimum, mean and maximum values for each fish and habitat variable on the two reefs. 52
Table 5.3: Average RMSE values for an 'average predictor' model. 53
Table 5.4: Correlation coefficients obtained between individual habitat and fish variables. 58
Table 6.1: Confusion matrix for classification of shallow-water habitats around Bawe Island. Field data in columns, classification results in rows. 75
Table 6.2: Confusion matrix for classification of shallow-water habitats around Chumbe Island. Field data in columns, classification results in rows. 75
Table 6.3: Characteristic scales for fish-habitat relationships. 86
Table 6.4: Characteristic scales for relationships between remotely sensed rugosity and fish species richness from recent studies. Note that for several studies, other scales than the one noted obtained very similar correlations between the two variables. 89
Table 6.5: Characteristic scales for relationships between remotely sensed habitat diversity and fish species richness from recent studies. Note: Habitat variety has been quantified using different approaches in all four studies. N.S. = no significant relationships found. 90
Table 7.1: Median RMSE values for predictive models by input data, model type, and response variable. 95

Table 7.2: Comparison of predictive performance achieved by recent studies predicting fish community variables from remotely sensed data. Coefficients of determination for this study are provided along with values (in parentheses) obtained from testing accuracy on the training set. * Wedding et al. (2008) report their results with Spearman rank correlations (R) – these values have been squared here for a rough comparison with the coefficients of determination from other studies. 107

Table 7.3: Coefficients of determination for IKONOS-based models of all three fish community variables. Calculations based on test data outside parentheses, calculations based on training data in parentheses. 108

LIST OF FIGURES

Figure 2.1: Global distribution of scleractinian corals (blue areas) and coral reefs (red areas) (Veron 2000). 5

Figure 2.2: Global distribution of generic richness in scleractinian corals (Veron 1995). Because of the difficulty of field identification of corals to the species level, generic richness is typically used as a surrogate for species richness. 6

Figure 3.1: Overview of major ocean currents and monsoon winds in the Western Indian Ocean. Modified from Ngusaru (2002). 17

Figure 3.2: The location of Chumbe and Bawe islands in relation to Unguja and the Tanzanian mainland. 20

Figure 4.1: Field sites distributed in clusters on the reef area around Bawe Island. 24

Figure 4.2: Field sites distributed on the reef around Chumbe Island. 25

Figure 4.3: Processing flow for fish data. Data are in ellipses, processing steps in boxes. Some simple processing, e.g. counting the number of species in a list, has been omitted. Variables used in the subsequent analysis have bold borders. 26

Figure 4.4: Processing flow for habitat data. Graphic conventions are similar to those used in Figure 4.3. 28

Figure 4.5: Field sites used for classification of the Bawe image. 34

Figure 4.6: Field sites used for classification of the Chumbe image. 35

Figure 4.7: A graphical illustration of the principle behind the rugosity calculations of the Benthic Terrain Modeler (Jenness 2002). 39

Figure 4.8: The geometry behind habitat variety calculations. Each colour represents a separate substrate type. With an IKONOS pixel size of 4 m, this example illustrates the calculation with a radius of 10 m. 40

Figure 5.1: Results from 100 runs of the six models for each dependent variable. Results are shown using both cross-validation and bootstrapping resampling methods. 55

Figure 5.2: Examples of non-linear relationships between fish species richness and three habitat variables. Similar relationships exist for the other two fish variables. Lines are inserted to aid interpretation. 55

Figure 5.3: First example of an interaction effect. At shallow depths (<4 m) conservation status has no significant influence on fish species richness, whereas the effect is significant at greater depths (>4 m). 56

Figure 5.4: Second example of an interaction effect. The variable 'coarse rugosity' has significant positive influence on fish species richness at shallow depths (<4 m), but not at greater depths (>4 m). 56
Figure 5.5: Summary of variable importance for all combinations of response variable, for LM, GAM and Bagging models. The boxes and whiskers in this and all similar figures are drawn to according to Tukey's definitions. The centre-line marks the median value, the two ends of a box mark the first and third quartiles, whiskers mark end of 3/2 of the interquartile range, circles mark outliers (outside whiskers). 59
Figure 6.1: Result of Maximum Likelihood Classification, Bawe Island. Note scale difference between Figure 6.1 and Figure 6.2. 69
Figure 6.2: Result of Maximum Likelihood Classification, Chumbe Island. Note scale difference between Figure 6.1 and Figure 6.2. 70
Figure 6.3: Linear relationships between the depth-invariant index and live coral cover on the two reefs. 72
Figure 6.4: Spatially distributed prediction of live coral cover for Bawe Island. Coral areas are shown in red, overlaid on a true-color composite of the original data. 72
Figure 6.5: Spatially distributed prediction of live coral cover for Chumbe Island. Coral areas are shown in red, overlaid on a true-color composite of the original data. 73
Figure 6.6: Correlations between IKONOS-based live coral cover estimates and *in situ* estimates of live coral cover, fish species richness, fish biomass and fish diversity, respectively. Spearman rank correlation coefficients shown on y-axis. 77
Figure 6.7: Correlations between IKONOS-based rugosity estimates and *in situ* estimates of depth range, coarse rugosity, fine rugosity, fish species richness, fish biomass and fish diversity, respectively. Spearman rank correlation coefficients shown on y-axis. 80
Figure 6.8: IKONOS-based rugosity estimates of areas around Chumbe Island, at pixel sizes of 8 and 150 m, respectively. The two maps cover the exact same area. Note the details on the reef flat west and south of the island in the left image (8 metre pixel size), and the complete dominance of the reef edge on the right image (150 metre pixel size). 81
Figure 6.9: Correlations between IKONOS-based habitat variety estimates and *in situ* estimates of substrate diversity, fish species richness, fish biomass and fish diversity, respectively. The habitat richness measure is shown in red; the

habitat diversity measure in blue. Spearman rank correlation coefficients shown on y-axis. 83
Figure 6.10: Map of habitat richness around Bawe Island. For each pixel, the map shows the number of different substrate types present within a 30 m radius. 84
Figure 6.11: Map of habitat richness around Chumbe Island. For each pixel, the map shows the number of different substrate types present within a 30 m radius. 85
Figure 7.1: Prediction accuracies for all models types, input data, and response variable. Note the variation in y-axis ranges between plots.......... 94
Figure 7.2: Comparison of species richness prediction accuracy from Bagging models using the four different data sets. 97
Figure 7.3: Comparison of biomass prediction accuracy from Bagging models using the four different data sets.......... 97
Figure 7.4: Comparison of diversity prediction accuracy from Bagging models using the four different data sets.......... 98
Figure 7.5: Variable importance from Bagging models based on the IKONOS data set.......... 100
Figure 7.6: Variable importance for species richness prediction using the LM and GAM models, based on "All" data. The many variables with zero influence in the GAM models are caused by the maximum of 17 variables that could be included in this model.......... 102
Figure 7.7: Variable importance for species richness prediction using the Bagging and Random forest models, based on "All" data. 103
Figure 7.8: Variable importance for species richness prediction using the Boosted trees and SVM models, based on "All" data.......... 104
Figure 7.9: Map showing predicted fish species richness around Chumbe Island. Calculations done with Bagging model and IKONOS data set.......... 106

LIST OF ACRONYMS

ANN	-	Artificial Neural Network
COTS	-	Crown-Of-Thorns Starfish
DEM	-	Digital Elevation Model
FDR	-	False Discovery Rate
FWHM	-	Full Width at Half Maximum
GAM	-	General Additive Model
GCP	-	Ground Control Points
GPS	-	Global Positioning System
LM	-	Linear Model
MPA	-	Marine Protected Area
NIR	-	Near-Infrared
PECCA	-	Pemba Channel Conservation Area
RMSE	-	Root Mean Square Error
RT	-	Regression Tree
SVM	-	Support Vector Machine
TIN	-	Triangulated Irregular Network

A NOTE ON TERMINOLOGY

In this thesis, several terms that do not have a single agreed upon definition are used. This section briefly describes the most common of these terms, in order to define their use in this thesis.

Band: The terms '*band*' and '*channel*' are often used interchangeably to denote "**the measurements of detectors on a remote sensor that fall within one particular wavelength interval**". The IKONOS sensor thus has 5 '*bands*', four of them with wavelength intervals in the blue, green, red, near-infra-red spectra, respectively, and the last one covering one large interval from the blue to the near-infrared.

Biodiversity: The term '*biodiversity*' is often defined very broadly as "hereditarily based variation at all levels of organization" (Wilson 1997). As a more practical definition for this study, it is defined as "**the diversity of the fish assemblage in the area of interest**", as calculated using one of the mathematical diversity indices in common use (Dickman 1968). The index used here is the Shannon-Weaver index (Shannon 1948), which was chosen mainly due to its widespread use in the literature.

Characteristic scale: The spatial scale at which two variables are most correlated. It is assumed to be the scale at which one or more important ecological processes linking the two variables.

Fish community: Multiple definitions and measures of a '*fish community*' are used in existing studies, for a comprehensive example see Friedlander and Parrish (1998). In this thesis, the '*fish community*' is defined as "**the totality of fish found within a given area at the time of observation**", and is used when describing more than the measure of biodiversity defined above. The '*fish community*' is quantified using three measures: Species richness, species diversity, and total biomass.

Habitat / Substrate: These two terms are often used interchangeably in the coral reef literature, as habitats are largely defined in terms of substrate type. I have used them interchangeably in discussion paragraphs, but distinguished between them when naming variables. The *in situ* measures of variety have thus been named substrate diversity and substrate evenness, as they are based on direct observations of substrate covers. On the other hand, the IKONOS-based measures of variety at various spatial scales have been named *'habitat variety'* as they are based on but not identical to the substrate classes, since most of the habitat classes contain a range of substrate types within them.

Habitat diversity: This term is used to denote the remotely sensed measures of *'habitat variety'*, when this variety is calculated as the **Shannon Diversity Index on the basis of classified substrate types**.

Habitat richness: This term is used to denote the remotely sensed measures of *'habitat variety'*, when this variety is calculated as the **number of substrate types present within a given radius**.

Habitat variety: This term is used as a **general reference to the concept of diversity of habitat types**. As such, it covers both the terms *'habitat richness'* and *'habitat diversity'*, which both relate to specific quantified measures of *'habitat variety'*.

IKONOS: The IKONOS satellite is a commercial earth observation satellite, launched in 1999, which provides high-resolution satellite imagery from almost anywhere on Earth. The satellite can be programmed for specific acquisitions, or images can be acquired from a large library of previously recorded data. The term '*IKONOS*' is used to denote **both the IKONOS satellite, and the sensor**.

Importance: The word '*important*' is used in two ways in this thesis, one of which is in the meaning of crucial or significant (in the non-scientific sense). When referring specifically to the function of a habitat variable as an explanatory variable in a predictive model, the word is used to mean **the influence that permutation of this variable has on the predictive performance of the model, quantified as the increased in RMSE the permutation causes**.

***In situ* / Remotely sensed**: These terms are used repeatedly throughout the thesis to describe parts of the dataset and variables used in the analyses. "*In situ*" refers to data collected during fieldwork and variables derived from these data.

“Remotely sensed”, on the other hand, refers to data produced by the IKONOS satellite, and variables derived therefrom.

R: Unfortunately the capital letter ‘*R*’ can be used with a range of different meanings in the scientific literature, often without sufficient explanation. Unless otherwise specified, in this thesis it is limited to mean **correlation coefficient**. In each case, it will be specified whether Pearson’s correlation coefficient or Spearman’s rank correlation coefficient has been used for the particular calculation. Pearson’s coefficient has been used whenever a linear relationship between the two variables could be assumed, whereas Spearman’s coefficient has been used when this assumption did not seem justifiable. Because many non-linear relationships between variables were identified in this study, Spearman’s coefficient has been used for the majority of calculations. Whenever the square – ‘R^2’ – has been used, calculations are based on Pearson’s coefficient. The lowercase ‘*r*’, unless otherwise specified, is limited to mean the **radius** of a circle. Some calculations in the thesis are based on squares rather than circles, in these cases the ‘*r*’ is the **half side length** of the square. This is used as a rough parallel to the radius of a circle. Other (always specified) used of the letter ‘*R*’ include reflectance, and the statistical computing software package. Rugosity, a measure of the structural complexity of a surface, is not abbreviated.

Remote sensing: ‘*Remote sensing’* is often defined as “the science and art of obtaining information about an object, area, or phenomenon through the analysis of data acquired by a device that is not in contact with the object, area, or phenomenon under investigation” (Lillesand and Kiefer 1994), which then hinges on an unclear definition of “contact”. In this thesis, the term ‘*remote sensing*’ is confined to “**the use of airborne or spaceborne instruments, detecting the emission or reflection of electromagnetic radiation from an object on Earth**”. This is an operational definition for this thesis only, which specifically excludes acoustic instruments, or ground-based sensors.

Scale: The term scale is used according to different conventions by people in the two fields of research this thesis bridges – remote sensing geographers, and coral reef biologists. In this thesis, it is used in two different ways. In the introduction, global, regional, and “reef” scales are referred to. These should be self-explanatory. In the remainder of the thesis, “small scale” or “fine scale” refers to measurements made either within a relatively small area, or with relatively high spatial frequency, depending on context. The scale in these cases should be seen in opposition to “large scale” or “coarse scale” measurements made either within a relatively large area or with low spatial frequency, again depending on context. This use most closely resembles the way the term is used by coral reef biologists, and care must therefore be taken by readers used to the geographic/cartographic convention.

Substantial / Significant: These two terms are often used interchangeably. In this thesis, the term *‘substantial’* has been used to describe **“a subjective, qualitative, assessment of importance”**, whereas *‘significant’* has been consistently used to describe **“statistical significance”**.

CHAPTER 1: INTRODUCTION

1.1 Coral reefs in decline

Coral reefs exist in the warm and shallow waters off tropical coastlines. Built from the calcium carbonate skeletons of myriads of individual coral animals, the reefs create the largest biogenic structures on Earth, and form the home of close to one million marine species (Reaka-Kudla 1997). Biogenic shallow-water reefs have existed in various forms since the Carboniferous, and the current dominance of scleractinian corals began in the Triassic, more than 200 million years ago (Wood 1998). However, despite their longevity through geologic time, coral reefs are currently in rapid decline across the globe. One fifth of all current coral reefs are considered degraded beyond their ability to recover, and more than half of the rest are headed in the same direction (Wilkinson 2004). The degradation can take different forms, but typically includes a loss of coral cover (Bruno and Selig 2007), an increased dominance of algae (Hughes 1994), and a flattening of the three-dimensional structure of the reef (Graham et al. 2006b).

The current situation has numerous causes, all arguably anthropogenic (Jackson 2008). Local issues such as overfishing and use of destructive fishing methods, coral mining, sedimentation and nutrient enrichment all impact the reef-building corals negatively, while some (e.g. nutrient enrichment) at the same time improve conditions of life for their competitors (Lapointe 1997). These issues can be managed at the local scale through coastal management efforts, including designation of marine protected areas (Salm and Clark 2000), but other issues, global in nature, add to the list of threats: warming waters in the upper layers of the ocean cause mass coral bleaching events (Goreau and Hayes 1994; Hoegh-Guldberg 1999), and ocean acidification reduces the ability of corals and other animals to precipitate their calcium carbonate skeletons and shells (Hughes et al. 2003; Kleypas et al. 1999; Veron 2008). All in all, the future of coral reefs is bleak (Knowlton 2001).

As coral reefs degrade, the animals that depend on them suffer the effects. More than 10,000 species of fish are dependent on coral reefs for their existence (Paulay 1997), and their decline is not only an ethical problem, but of immediate importance to human society. Fish constitute an important source of income and protein to coastal communities (Brainerd 1994), they are a source of attraction for dive and snorkel tourism, and they serve a range of ecological functions in reef ecosystems (Hughes 1994; Jackson et al. 2001).

The effects of degrading reefs on their resident fish communities have been studied in both manipulated (Syms and Jones 2000) and natural settings (Friedlander and Parrish 1998), with a range of results sometimes including local extinction of species (Jones et al. 2004b). Collectively, the body of studies of fish-habitat relationships have shown that the specific relationships vary between individual reefs, between protected and unprotected areas, along spatial scales, and with absolute values of habitat variables (Syms and Jones 2000). This variation has so far precluded anything but broad conclusions, such as the benefit of live coral and structural complexity for the fish community (Knudby et al. 2007). However, one 'natural' experiment - the widespread mass coral bleaching events following the 1998 El Niño - has shed light on the likely future for many reefs. The bleaching-induced loss of live coral is of immediate consequence to the corallivorous part of the fish fauna, and the subsequent breakdown of the reefs' structural complexity, when coral skeletons erode and collapse, impact the rest of the resident fishes (Garpe et al. 2006; Graham et al. 2006b).

1.2 MPAs and the use of remote sensing

In the face of coral reef decline, the most widely adopted management response has been development of Marine Protected Areas (MPAs), protecting against local threats to both corals and fish. MPAs are designed to incorporate a range of habitats, species and areas of high biodiversity (Roberts et al. 2002), and are ideally incorporated into networks whose design is based on typical larval dispersal of critical species (Carr and Reed 1993). MPAs are therefore in need of accurate spatial information on both fish and coral distributions, and the influence of distribution and changes in habitat on the fish fauna. Spatial information can also be used by MPAs to determine the boundaries of a minimum effective area of protection, and to design zonation plans. Due to the relatively inaccessible nature of reef environments for fieldwork, remote sensing is the only tool that realistically can provide the needed spatial information. Passive optical remote sensing has been used since the launch of Landsat 1 (Smith et al. 1975) to outline the spatial distribution of geomorphologic zones (Andréfouët et al. 2001), dominant substrate types (Mumby et al. 1997b), reef community classes (Turner and Klaus 2005), and bathymetry (Lyzenga 1978; Stumpf et al. 2003). However, the link between the remotely mapped benthic structures and the fish community that relies upon them has only rarely been made (Pittman et al. 2007; Purkis et al. 2008; Wedding et al. 2008), and the actual use of remotely sensed habitat maps for MPA management has been limited to a few examples in highly developed countries (Newman and LeDrew 2008). Some questions remain, the answers to which will facilitate mapping of fish communities by remote sensing, and thereby increase the utility of this tool for coral reef MPA management. Three of these (henceforth: sub-questions) will be addressed in chapters of this thesis.

1.3 Sub-questions

A) What is the statistical nature of fish-habitat relationships? Many studies have implicitly assumed linearity in fish-habitat relationships, and developed predictive models on that basis. Others have allowed for continuous but non-linear relationships, or used classification-based approaches with breakpoints for individual variables. In order to increase the practical value of such predictive models, this research aims to test various approaches to modeling fish-habitat relationships, both in terms of their relative predictive capability, and in terms of the habitat variables they identify is important for predicting the fish community.

B) How accurately can habitat variables be estimated remotely, and at what spatial scales are these variables most predictive for the fish community? Because of the limitations of fieldwork on a coral reef, typically using SCUBA or snorkel gear, most studies have been limited to measuring habitat variables at small spatial scales. Both fish and habitat variables are typically sampled either through point counts with a typical radius of 5 m, or in transects with lengths of 20-50 m. However, both benthic and fish communities are influenced by physical processes at larger scales, and many fish species migrate well beyond such distances. Using remote sensing, this research adds to a small but growing body of literature that aims to find the spatial scales at which specific measures of habitat exert the greatest influence on the fish community, and interpret these scales in terms of their significance for reef ecology and conservation.

C) How does remote sensing compare to traditional fieldwork for mapping a coral reef fish community? Remote sensing can provide a cost-effective alternative to traditional fieldwork for mapping and monitoring habitat variables. However, remote sensing is unable to map habitat as accurately and at as fine a spatial scale as *in situ* surveys. If remote sensing is to be applied to map and monitor reef fish communities through its ability to map habitat, the accuracy it can provide, compared to traditional fieldwork, must be established. This research compares the predictive abilities of models based on remotely sensed data with similar models based on *in situ* data.

1.4 Research question

In combination, answers to the three sub-questions combined will answer the research question of this thesis:

How can remote sensing be used to map coral reef fish communities?

The wording of the question is meant to allow for investigation of all three sub-questions, while arriving at a general conclusion about the possibility of mapping reef fish communities through their habitat.

1.5 Thesis structure

In chapters 2 and 3 we expand on the background for this research through a review of the existing literature and a description of Zanzibar, particularly the two reefs where the research took place, respectively. In chapter 4, we outline the methodology and the specific methods used for data collection, processing, and statistical analysis, along with a justification for their use. Chapters 5 through 7 then deal with each of the sub-questions: comparing predictive models based on *in situ* data, deriving the most relevant spatial scales for remote sensing of fish habitat, and investigating the use of remote sensing for reef fish mapping. Discussions of methods and results are included in each of these chapters. In chapter 8 we provide a conclusion for the thesis by giving a synthetic answer to the research question, putting the research into context, and outlining promising avenues for future research.

CHAPTER 2: RESEARCH CONTEXT

In this chapter we provide a background for the research through a review of the relevant existing literature with particular emphasis on studies of the relationships between reef fishes and their habitat, and the potentials and limitations of remote sensing of coral reefs.

2.1 Global distribution of coral reefs and associated fishes

At the global scale, coral reefs are limited to tropical nearshore areas with clear and shallow water and mean annual water temperature of 18 ºC or higher, roughly corresponding to the area between latitudes 30º north and south (Yonge 1940). Within this broad tropical belt, the occurrence of coral reefs is moderated by ocean currents, which govern the movement of nutrients, oxygen and coral larvae, in addition to the cold and warm water masses themselves. The current distribution of continents produces upwelling at the Eastern margins of the two major oceans, the Atlantic and the Indo-Pacific, reducing reef growth, while reefs flourish on the Western margins of the oceans (Hubbard 1997; Veron 1995). Factors such as tidal ranges, nutrient levels, and river outflows can restrict reef growth more locally, e.g. at the mouths of the Amazon and Orinoco rivers. The reef-building corals themselves have a wider distribution, extending into areas where their survival, though not reef-building, is possible (Wood 1983) (Figure 2.1).

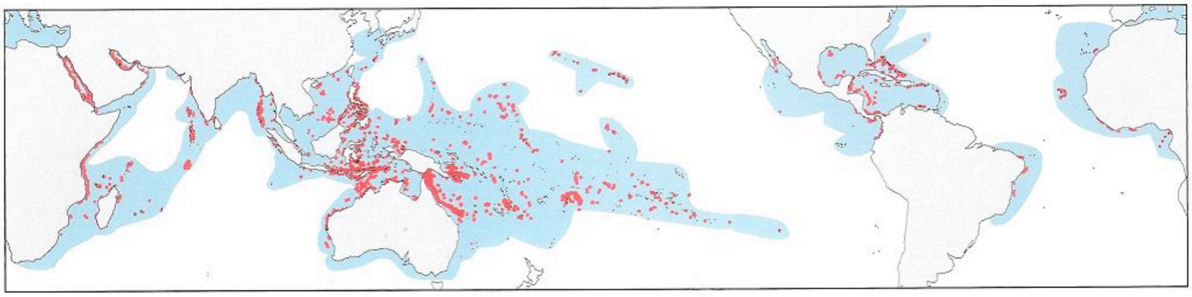

Figure 2.1: Global distribution of scleractinian corals (blue areas) and coral reefs (red areas) (Veron 2000).

Within the areas conducive to reef building, spatial patterns of biodiversity, relatively constant within a human time-frame, have been established by numerous field expeditions (Veron 2000). Both the two major oceans have well-known centres of biodiversity, from which the richness of coral species gradually diminishes with the distance from the centre. In the Indo-Pacific, the large area of shallow seas formed by the archipelagos of Indonesia, the Philippines and New Guinea, the 'coral triangle', forms the centre of coral biodiversity (Bellwood and Hughes 2001; Briggs 1999), as the south western Caribbean does in the Atlantic. From the coral triangle, coral species richness gradually declines in any direction - towards the Western Indian Ocean, the Eastern Pacific, Japan or Southern Australia (Veron 1995) (Figure 2.2). A similar biogeographic pattern is found for reef fishes (Bellwood and Hughes 2001; Bellwood 2002), and for all other reef-associated taxa for which data are available (Paulay 1997).

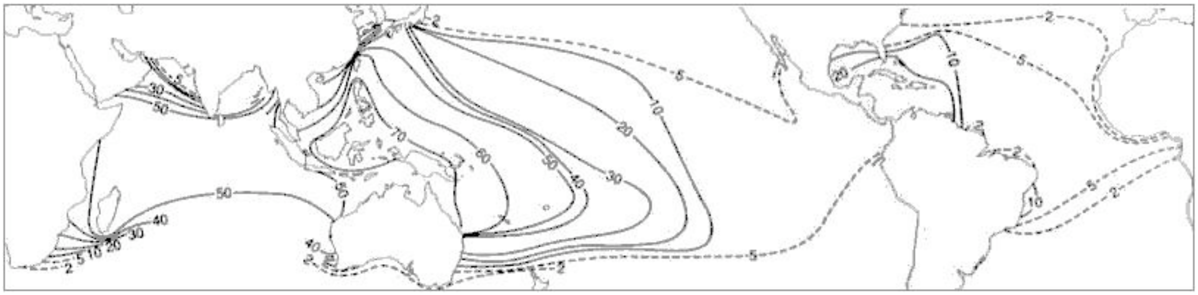

Figure 2.2: Global distribution of generic richness in scleractinian corals (Veron 1995). Because of the difficulty of field identification of corals to the species level, generic richness is typically used as a surrogate for species richness.

At the scale of individual coral reefs or reef complexes, biodiversity is composed of a subset of the regionally available species, but these local spatial patterns of biodiversity are less well known, and subject to more rapid change. They are determined by a combination of the physical environment (Friedlander and Parrish 1998), stochastic processes (Sale and Dybdahl 1975), and human intervention (Chapman and Kramer 1999). It is at this scale that most management interventions exist, and where conservation is easiest to implement and enforce.

2.2 Reef-scale distribution of biodiversity

At the scale of individual reefs, the spatial taxonomic covariance seen globally is less pronounced. Nevertheless, functional indicators have been proposed as surrogate measures of local spatial variations in biodiversity, typically for conservation planning purposes. These include fish species richness as an indicator for invertebrate and plant biodiversity (Ward et al. 1999) or for coral biodiversity (Beger et al. 2003), or molluscs as an indicator of the biodiversity of macroalgae

(Gladstone 2002). More importantly, some attributes of coral reef habitats themselves also co-vary with the biodiversity of fish and mobile invertebrate taxa, and can function equally well or better than species as biodiversity indicators (Mumby et al. 2008; Ward et al. 1999). This suggests that taxonomic covariance at the local scale can be due to co-varying habitat influences, and that the relevant habitat variables can influence local biodiversity strongly. However, the utility of habitat measures as surrogates for biodiversity are not ubiquitous (Stevens and Connolly 2004), and need to be established locally.

2.2.1 Fish-habitat relationships on coral reefs

Numerous studies have demonstrated statistical relationships between habitat variables and measures of the fish community on coral reefs (Knudby et al. 2007). Sea urchins and *Conus* sp. snail abundances have also been shown to co-vary with structural complexity (Kohn and Leviten 1976; McClanahan 1988). However, even within the scale of individual reefs, the strength of these relationships depends on spatial scale. So far a limited number of local scale effects have been shown on reefs, but the issue is of importance for practical conservation reasons, particularly in MPA design where the spatial extent of MPAs needs to match the scales of critical habitats and territory size of species within them (Kendall et al. 2004). Problems arise because the relevant fish-habitat relationships, as well as the relevant spatial scale, vary depending on species and life stage (Grober-Dunsmore et al. 2008; Sale 2002). Some fish migrate daily as adults, others migrate annually to spawn, and some migrate to find new habitat types during ontogenetic shifts. Nevertheless, aggregate fish community variables can show general trends that aggregate the information from individual species (Purkis et al. 2008).

Causal mechanisms have been proposed for these relationships. The influence of depth has been related to disturbance, where wave action at shallow depths strongly favours wave-resistant and fast-colonizing coral species, whereas less disturbance and diminished light at greater depths favour slow-growing and metabolically more efficient species. The intermediate depth provides a habitat that houses representatives of both extremes, and thus maximizes species richness (Huston 1994). The positive influence of live coral cover on fish species richness has been related to larval settlement success, and to the survival of coral-dwelling and corallivorous species (Jones et al. 2004b). In addition, the live coral creates a structurally complex habitat that provides shelter for prey species and a range of structural niches for species of varying body size (Friedlander and Parrish 1998). Increasing species richness with proximity to the reef edge may be related to the increasing availability of food items for planktivores at the reef edge.

Relations between habitat and fish variables have also been shown to exist temporally. Experimentally, Syms and Jones (2000) noted a decline in fish abundance with loss of reef structure in a two-year experiment, results confirmed by an observational (non-experimental) 30-year study by Connell et al. (1997). Jones et al. (2004a) noted a decline in fish species richness with loss of live coral cover over 8 years in Papua New Guinea, with local extinction of species particularly dependent on coral. Impacts of the 1998 El Niño were studied in the Seychelles by Graham et al. (2006a), who found that bleaching-induced loss of live coral cover did not significantly impact species richness, though it did lead to possible local extinction of species highly dependent on coral. However, they also found that the subsequent loss of structural complexity, following erosion of bleached coral skeletons, did affect species richness significantly, changes in coral cover and structural complexity together explaining 57% of the decrease in species richness. They also found that small species were lost first, larger species only as more structure degraded. Garpe et al. (2006) confirmed these findings in Tanzania, and documented different responses from different functional groups, coral-dependent species again suffering the greatest losses and possible local extinctions.

2.2.2 The statistical nature of fish-habitat relationships

Despite the large amount of empirical data on fish-habitat relationships, their statistical nature remains poorly explored. Relationships between habitats and aggregate measures of the fish community are mediated by the species that form the community. These relationships change through time, and depend on the absolute values of the variables observed (Jones and Syms 1998). In addition, it is likely that numerous interaction effects exist between the relevant variables, one variable moderating the relationships between two other variables. For example, depth may have significant influence on fish species richness through its covariance with wave action, shelter space and food availability on the reef slope where high coral cover typically exists, but similarly may have no influence on fish species richness in a sandy lagoon area with typically low or no coral cover.

Despite this obvious complexity of relationships, most studies have assumed simple linear relationships between the studied variables, using methods of data analysis such as canonical correlation analysis (McCormick 1994), discriminant analysis (Ormond et al. 1996) or various forms of linear models (LM) (Chapman and Kramer 1999; Friedlander and Parrish 1998). Non-linearities in the relationships, as predicted theoretically by intermediate disturbance theory (Connell 1978) and shown empirically by Knudby et al. (2008), are dealt with by log or root transformations of independent or dependent variables when necessary, e.g. Kuffner et al. (2007). However, there is no theoretical basis for assuming linear relationships between habitat and fish variables in the first place (Jones and Syms 1998), and more complex models may therefore both provide a more realistic description and deeper ecological insight, and provide lower prediction error in predictive models. One example of a statistical model better suited to deal with non-linearities is the general additive model (GAM), which allows the additive use of different statistical models (e.g. linear, power, log, smoothing splines) (Knudby et al. 2008). However, the statistical models, both LM and GAM, are both unable to deal effectively with interaction effects, which is likely to limit their power to model ecological relationships on coral reefs.

In addition to statistical models, a new suite of algorithmic models are becoming available, most of them developed in the field of Machine Learning. These models differ from those described above in that the nature of the modelled relationships is not pre-supposed through model selection, but rather learned through a set of training data. Algorithmic models, such as Artificial Neural Networks (ANN), Support Vector Machines (SVM), and a variety of Regression Trees (RT), have only recently been used to model fish community variables (Pittman et al. 2007; Pittman et al. 2009), and a thorough evaluation of the different models and their ability to produce interpretable results and useful predictive models is pending.

Having determined the statistical nature and specific attributes of relationships between habitats and fish community measures for a given coral reef area, spatially distributed information on habitats will allow the distribution of these measures to be predicted in the form of a map. To derive spatial information on coral reef habitats, remote sensing has been the tool of choice since the first application of Landsat data (Smith et al. 1975).

2.3 Remote sensing of coral reefs[1]

Most coral reef remote sensing research has not been carried out in the context of biodiversity, but has focused on mapping geomorphologic zones (Andréfouët and Guzman 2005; Smith 1975) or substrate types (Andréfouët et al. 2003; Mumby et al. 1997b). Nevertheless, the existing research has built a solid foundation from which biodiversity studies can benefit.

Geomorphologic zones mapped with remote sensing typically include the forereef, reef crest, lagoon, backreef and patch reefs, as well as coral reef associated habitats such as seagrasses and mangroves; the classes used in a particular study depend on the site and the desired level of detail. Despite developments of semi-automated systems (Suzuki et al. 2001), geomorphologic zones are typically mapped manually - outlined on a plot of original or classified data by an expert user (Andréfouët et al. 2001; Andréfouët and Guzman 2005).

More automation has been possible for mapping substrate types. Based on the different spectral reflectance properties of substrate types such as coral, sand, algae, and seagrass, multi- and hyper-spectral instruments have been able to map these substrates to depths of 15-30 m in clear water (Mumby et al. 2004c). The level of detail that can be obtained, expressed as the number of classes that can be discriminated combined with the accuracy of the classification, depends on the platform and sensor type, and on environmental factors such as water depth and turbidity, the state of the sea surface, and the atmosphere (Mumby et al. 2004c). Early studies used Landsat TM and SPOT HRV sensors, which typically only allowed broad categories such as coral, sand, seagrass and algae to be discriminated. The better spatial resolution of the IKONOS and Quickbird satellites (Andréfouët et al. 2003; Mumby and Edwards 2002), and developments of airborne and satellite-based hyperspectral instruments (Kutser et al. 2003; Mumby et al. 1997b), have enabled mapping of more detailed classes while retaining satisfactory mapping accuracy (Mumby et al. 2004c). However, the dominant features of the spectral

[1] This section deals exclusively with passive airborne and spaceborne remote sensing. Shipborne acoustic remote sensing has found application on optically deep coral reefs, however, due to its very limited spatial coverage; acoustic instruments are usually not a cost-efficient alternative for coral reef studies. Active optical lidar instruments have also found application in mapping both depth and water optical properties, but are currently only available at a cost that precludes their general use.

signatures used to discriminate between typical coral reef substrates are in the part of the visible spectrum where water penetration is at its lowest, 550-700 nm (Kutser et al. 2003; Mobley 1994), which reduces the number of distinct substrate types that can be distinguished as depth increases (Capolsini et al. 2003; Hochberg and Atkinson 2003; Holden and LeDrew 1999).

Only a few studies have related the mapped substrate types and geomorphologic zones to specific species or species assemblages. These studies do not infer the presence of particular species from the spectral reflectance of the area, but rather map classes of species assemblages in which particular species are known from field observations to be dominant (Purkis et al. 2006; Turner and Klaus 2005). The remotely sensed information thus functions more as geolocation of field observations than as the primary information source, and such studies require extensive fieldwork for each investigated site.

2.4 Mapping habitat variables

In addition to geomorphologic zones and substrate types, remote sensing has also proven its ability to map several of the habitat variables shown by field studies to influence the fish community, including depth, structural complexity, and live coral cover. Depth (Lyzenga 1978; Stumpf et al. 2003) and live coral cover (Joyce 2004a; Joyce et al. 2003) are routinely mappable using remote sensing, whereas the mapping of structural complexity only recently has been explored (Pittman et al. 2007; Purkis et al. 2008). Field measures of these variables are typically necessary to calibrate remotely sensed values, and the issue of disparate spatial scales remains, particularly for structural complexity (Knudby and LeDrew 2007).

2.4.1 Remote sensing of depth

Methods for remotely sensing depth rely on the wavelength dependency of light attenuation in water. Longer wavelengths attenuate more rapidly (Mobley 1994), hence substrates located in deeper water will show a greater proportion of reflected light in shorter wavelengths (Lyzenga 1978). Variation in substrate spectral reflectance introduces error, which can, at least in theory, be mitigated when using hyperspectral data (Hedley and Mumby 2003). Water optical properties and substrates with very low reflectance introduce additional complications (Philpot 1989; Stumpf et al. 2003), and depth mapping always requires *in situ* calibration.

2.4.2 Remote sensing of live coral cover

Mapping of live coral cover has suffered from low levels of accuracy due to problems of sub-pixel heterogeneity and high spectral similarity between corals and other substrate types such as algae and seagrass. This spectral similarity, and the complicating influence of variations in water depth and optical properties, have hampered efforts to use spectral unmixing (Hedley and Mumby 2003), routinely used in terrestrial environments, to separate live coral from spectrally similar substrates. High accuracy with this approach is thus dependent on near-perfect conditions, i.e. hyperspectral imagery with high spatial resolution (≤1 m), clear and shallow water, independently known depth, and absence of brown macroalgae (Hedley et al. 2004; Mumby et al. 2004b). Despite these problems, some studies have been able to demonstrate success in mapping live coral cover. Isoun et al. (2003) used a classification-based approach, with seven classes based only on percentage live coral cover, and achieved 77% overall classification accuracy with airborne hyperspectral imagery. Newman et al. (2007) achieved similar levels of accuracy with four classes based on percentage live coral cover, using IKONOS data. Using hyperspectral data from CASI-2, Joyce (2004a) used an index-based approach to investigate correlations between live coral cover and spectral reflectance ratios and derivatives. Results achieved a coefficient of determination (R^2) of 0.58, and showed that the optimum band ratio and derivative varied between ‘blue’ and ‘brown’ coral types (Hochberg et al. 2003b), and depended on resampling of the dataset, depth and water quality.

Other habitat variables with known influences on species richness (e.g. distance to reef edge) are also mappable. There is thus ample scope for further exploring the potential of this approach to predict the spatial distribution of biodiversity on coral reefs. The development of methods for mapping habitat variables using remote sensing is based on correlations between measures of the given variables derived *in situ* and derived using remote sensing data. The following sections will outline how *in situ* and remote sensing-based data are collected for each habitat variable, and discuss potential issues that could arise when relating the two kinds of data.

2.4.3 Remote sensing of structural complexity

The physical structure of coral reefs exists at a continuum of scales, ranging from the intricate structure of the coral skeleton, through the variety of structures formed by coral colonies, to the regional distribution of reef complexes. At scales available to remote sensing, structural complexity can be quantified using a variety of measures all calculated on the basis of a Digital Elevation Model (DEM), itself typically produced by the methods described above for remotely sensing depth. Once the DEM exists measures of structural complexity are limited only by the ingenuity of the investigator; existing measures include linear or triangulated rugosity, slope, curvature, fractal dimension and more (Brock et al. 2004; Kuffner et al. 2007; Pittman et al. 2009; Purkis et al. 2008). The spatial resolution of the remote sensor obviously determines the smallest spatial scale at which structural complexity can be resolved. Currently, no sensor can resolve structural complexity at the same spatial scales at which field studies have shown influence on fish biodiversity (Knudby et al. 2007), and results to date suggest that correlations between *in situ* and remotely sensed structural complexity depend both on the spatial scales compared and on the environment in question (Knudby and LeDrew 2007; Kuffner et al. 2007; Wedding et al. 2008).

2.4.4 Remote sensing of habitat variety

In addition to mapping variables shown by field studies to influence fish biodiversity, remote sensing enables the quantification of other aspects of habitat not easily obtainable through field studies. One example is habitat variety (note: see terminology for definition of '*habitat variety*'), which can be quantified at a range of user-determined spatial scales using habitat maps (Purkis et al. 2008). No standardized procedure exists, and quantifications of habitat variety and their usefulness in biodiversity studies will depend on the number and relevance of substrate classes mapped, the measure used, and the spatial scales at which it is calculated. However, it is important to note that remote sensing in this case does not only estimate the value of a habitat variable that could be more accurately measured *in situ*, such as live coral cover, but enables quantification of a variable that is practically immeasurable *in situ* at scales beyond a few metres.

2.5 Mapping fish biodiversity

Remote sensing thus offers an indirect approach to mapping the biodiversity of fishes on coral reefs. Relationships can be established between remotely sensed habitat variables and measures of fish biodiversity, allowing remotely sensed maps of habitat variables to extrapolate biodiversity values to areas not sampled in the field.

Studying benthic organisms, Adjeroud et al. (2000) used SPOT satellite images to map pinnacle density, surface area, and hydrodynamic aperture of nine atolls in French Polynesia, and found that these explained part of the between-atoll variation in species richness of investigated taxa (corals, molluscs, echinoderms and algae). Similarly, Andréfouët and Guzman (2005) found a weak (non-significant) correspondence between geomorphologic zones and the biodiversity of corals and octocorals, though only at specific spatial scales.

The mobile nature of reef fishes may make spatial predictions of the biodiversity less accurate than for benthic organisms. Nevertheless, some promising results have been obtained. Kuffner et al. (2007), working on patch reefs in Biscayne Bay, used lidar-derived rugosity to predict fish species richness and abundance, with statistically significant but very weak results. Wedding et al. (2008), obtained stronger results in a similar lidar-based study in Hanauma Bay in Hawaii. Both studies illustrated the influence of spatial scale, though they arrived at different optimum scales for rugosity measurements (5 m and 25 m, respectively). Purkis et al. (2008), using IKONOS data, found that both remotely sensed habitat variety (quantified as Shannon evenness) and structural complexity showed significant relationships with fish species richness at a reef complex in Diego Garcia, and also demonstrated relationships with other measures of the fish community, such as abundance or richness of specific size classes. Optimum scales were found at 8 m for rugosity measurements, and 40 m for habitat variety. Pittman et al. (2007), working in Puerto Rico and the US Virgin Islands, combined several measures of both structural complexity and substrate availability into an RT model, variables entered at a range of scales from 5 m (bathymetric standard deviation) to 325 m (rugosity, seagrass areas, hard-bottom area).

These studies demonstrate the possibility of predicting the spatial distribution of fish community variables. However, the sub-questions listed in chapter 1 remain to be answered for the tool to become operational. The research presented in this thesis seeks to answer these questions, through a study conducted at two reefs, one protected and one unprotected, both located in Zanzibar, Tanzania, East Africa.

2.6 Summary

Numerous studies have established statistical relationships between coral reef fish communities and their habitat. However, the exact relationships change with time, through space, and with the absolute values of the variables in question. A range of model types exist to describe these complex relationships. Some of these model types allow modeling of non-linear and non-smooth relationships, and some allow modeling of interaction effects. Remote sensing data can be used to produce maps of coral reef habitats, and to estimate some aspects of habitat, such as depth, structural complexity, live coral cover and habitat diversity, that have all been shown to influence the fish community. Some studies have shown that it is possible to use remote sensing data to make predictions about the spatial distribution of fish community variables, but the development of methods and models is still in its infancy, and has not yet moved beyond the research community.

In this thesis, we present research that contributes to the understanding of fish-habitat relationships and the different approaches to quantify them, as well as exploring the potential to use these relationships and remote sensing data to produce maps of fish community variables. The research is based on two reefs in Zanzibar; one is heavily impacted by fishing, most of the other is protected by a strictly enforced no-take marine park.

CHAPTER 3: STUDY AREA

In this chapter we outline the context of the area in which the research took place, including the larger biogeographic context, some societal and political factors influencing the study site, and a more detailed description of the two reefs that are the focus of the study.

3.1 East Africa and the Western Indian Ocean

Zanzibar is located in central East Africa, at the extreme western end of the large Indo-Pacific biogeographic region. The climate is tropical, with two main seasons characterized by the prevailing winds. From November to March, northern '*kaskazi*' winds prevail, bringing sporadic rains and high temperatures. The period from March is characterized by heavy rains, which last until the southern '*kusi*' winds pick up in June, bringing cooler air and end to the rainy season. These southern winds in turn last until the next reversal of wind direction in November (Ngoile 1990; Ngusaru 2002).

East African waters are connected to the rest of the Indian Ocean by the South Equatorial Current, flowing westwards, connecting East Africa biogeographically to the Central and Eastern Indian Ocean. As such, the coral reefs of East Africa lie in the second most species rich region in the world, after the Western Pacific (Lieske and Myers 2001). The South Equatorial Current reaches the East African coast in the area around Northern Mozambique and Southern Tanzania, where it splits in two. The Mozambique current flows south, while the East African Coastal Current flows north, to Zanzibar and beyond (Ngusaru 2002) (Figure 3.1).The Zanzibar archipelago thus has a permanent northbound current, though currents in nearshore waters are often dominated by the tidal cycle (Ngoile 1990).

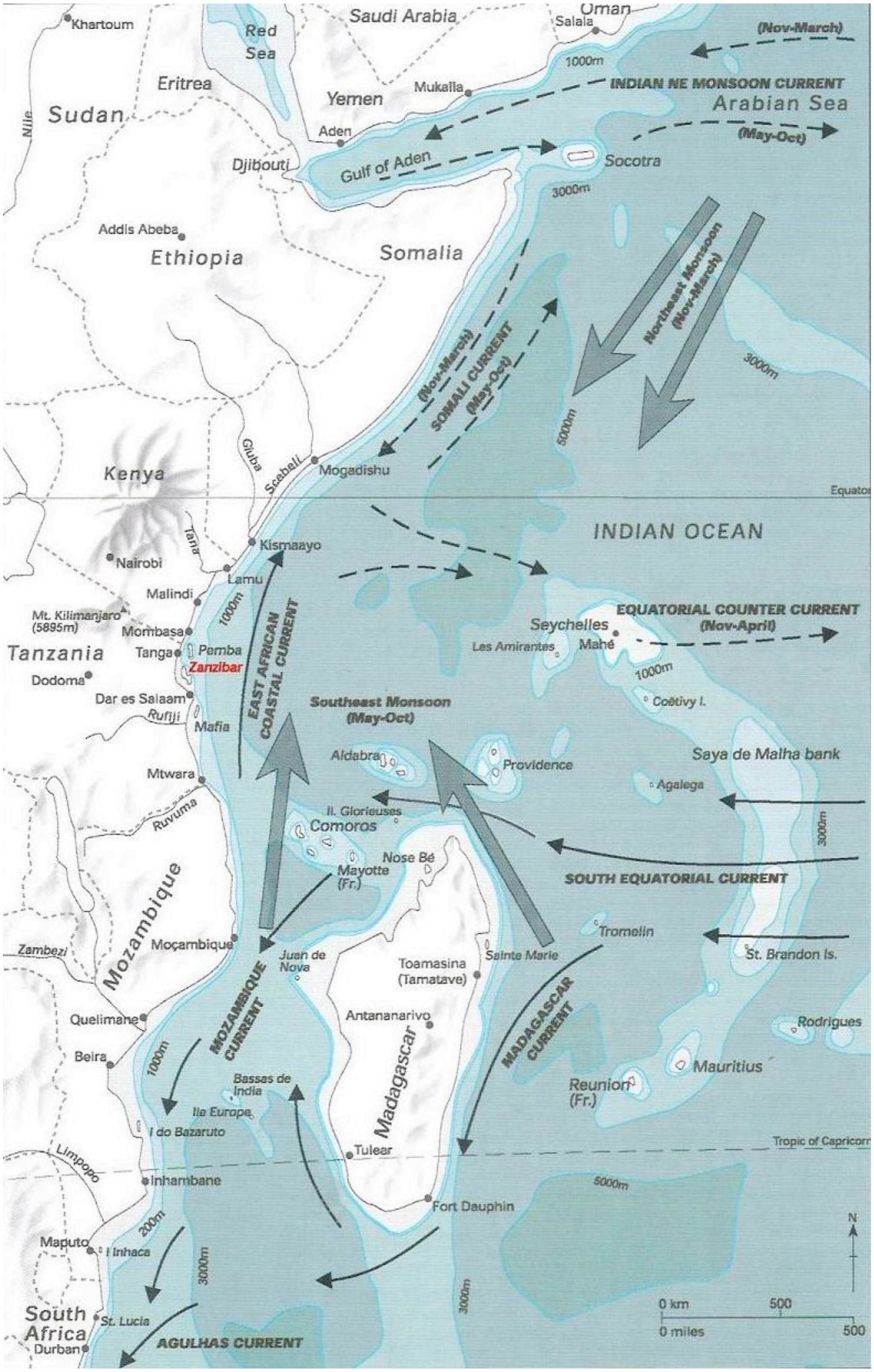

Figure 3.1: Overview of major ocean currents and monsoon winds in the Western Indian Ocean. Modified from Ngusaru (2002).

3.2 Zanzibar

Zanzibar consists of two main islands, Unguja and Pemba, with numerous smaller islands surrounding them. While Pemba is believed to be a part of the African continental plate that broke away 10 million years ago, Unguja and its surrounding islands are all raised Pleistocene reefs (Ngusaru 2002). Coral reefs around Unguja are dominated by the large fringing reef on the east coast, and a series of patch reefs and shorter fringing reefs around the islands and sand banks off the west coast (Horrill et al. 2000).

3.2.1 Fishing

Fishing is mostly small-scale in Zanzibar, with fishermen operating close to shore from small vessels, selling their fish immediately upon return to the local fish market (Jiddawi and Ohman 2002). The latest census for 2007 found 34,269 fishermen in Zanzibar, or 3.5% of the population, with almost half of the vessels used being dugout canoes holding one or two people (Jiddawi and Khatib 2007). Nevertheless, fisheries contribute 6% of the GDP, provide 60% of the protein consumed by local communities (Cesar et al. 2003), and form an important cultural part of coastal livelihoods (Grootenhuis and Lopez 2003). A wide range of fishing techniques are used, including traps, hook and line, a variety of nets, and several illegal but widespread techniques using bottom seine nets, spear guns, poison or dynamite. The limited data available point towards the reef fishes in Zanzibar being over-exploited, at least in areas close to settlements from which fishing pressure declines with increasing distance (Jiddawi and Ohman 2002).

3.2.2 Conservation

Although some limited traditional management practices were in place before the revolution in 1964 (Horrill et al. 2000), formal marine conservation has had a short and mixed history in Zanzibar. The first steps were taken in 1989, when an agreement for exclusive use of Mnemba Island, close to north-eastern Unguja, was reached between the Zanzibari government and a private tourism developer (EcoAfrica 2005a). The exclusive use was formalized by a lease agreement in 1992, which included a no-take zone extending 200 m from the mean high water mark of the island. In 2002 the no-take area was extended to 200 m beyond the reef crest and gazetted formally as an MPA. However, with only periodic enforcement, conservation effects have not materialized, and reports of repeated fishing, appear regularly (EcoAfrica 2005a). In 1994 the area stretching 300 m west of Chumbe Island, close to south-western Unguja, was gazetted as another no-take MPA and management authority given to a private company, Chumbe Island Coral Park. Management is carried out by rangers stationed on the island, and after some initial difficulties the no-take zone is effectively managed (Muthiga et al. 2000). Then in 1997, a large area south of Unguja, Menai Bay, was gazetted as MPA. It is, however, not a no-take area, and restrictions on fishing are limited to a prohibition of '*dago*', camping overnight on islands to continue fishing the next day. Menai Bay has also suffered from very limited enforcement, and conservation effects have not materialized (EcoAfrica 2005b).

Pemba's first MPA was established as a multiple-use marine reserve in 1998, which includes a no-take core zone. As on Chumbe Island, enforcement has been carried out by rangers based on the island, and despite initial difficulties the core zone has been effectively protected since 2001 (Tyler 2005). More recently, protection along Pemba's entire west coast has been established through the creation of the Pemba Channel Conservation Area (PECCA), a large multiple-use marine reserve created in 2006.

3.3 Two reefs, one protected and one unprotected

Data for the research presented here were collected on two fringing reefs located near islands immediately west of Unguja, Chumbe and Bawe. The reefs are similar in many aspects; they are both fringing reefs surrounding raised coral islands, with well-developed geomorphologic structures, and coral growth reaching to depths of 10-12 m (Knudby, pers. obs.) Both islands have tourism development in the form of low-volume, high-end resorts. However, their management histories are very different.

3.3.1 Chumbe

The reef on the western side of Chumbe, as described above, has been effectively protected from fishing since 1994, by rangers stationed on the island. Before its designation as an MPA this area was used by the military and fishing was not allowed, although anecdotal evidence suggests that enforcement was weak and fishing was more restricted by the reefs distant location in relation to local fishing communities (Muthiga et al. 2000). In addition to enforcing the MPA, Chumbe Island Coral Park has operated an eco-resort on the island since 1998, and conducted guided snorkel tours on the reef for tourists and local schoolchildren. The resort has implemented several measures to reduce the tourism's impact on the reef, including filtering of waste water and the collection and transportation of all trash to Unguja for proper disposal. The best developed part of Chumbe's coral reef lies in the protected zone along the island's western side, but the reef continues several km south of the island, before turning north again and forming a large lagoon on Chumbe's east side.

3.3.2 Bawe

The reef around Bawe is slightly less developed geomorphologically, the reef on the island's northern and eastern side consisting of a string of bommies rather than a well-formed reef crest. Bawe has never been protected against fishing, and fishermen are regularly observed on or off the reef, using both legal and illegal fishing gear (Knudby, pers. obs.) A resort also exists on the island, but no known measures are taken to limit the impact of tourism on the reef. The location of both islands, relative to Unguja and the Tanzanian mainland, is illustrated in Figure 3.2.

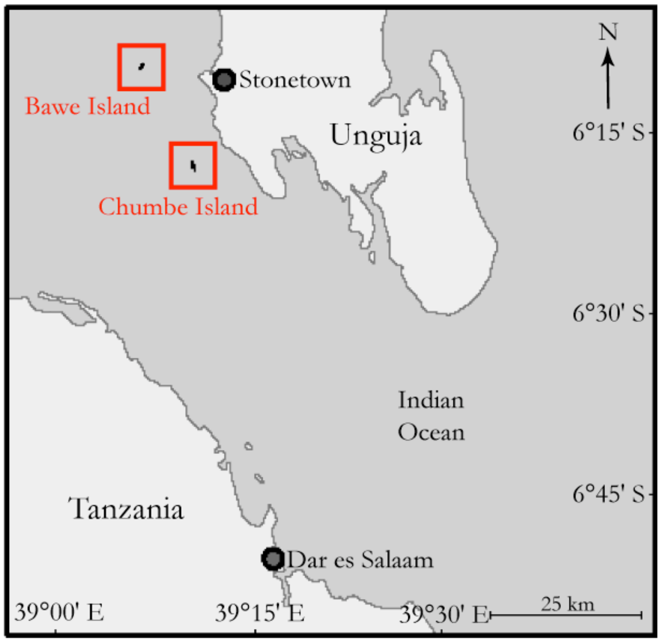

Figure 3.2: The location of Chumbe and Bawe islands in relation to Unguja and the Tanzanian mainland.

3.4 Summary

Reef fish are an important ingredient in the mixed livelihoods of the communities in Zanzibar, and specific results of relevance for marine management may therefore be of real benefit in the area. The choice of Chumbe and Bawe reefs as study sites for this research allows for a comparison between two important types of coral reef ecosystems: the exploited reef near human settlements whose fish fauna is heavily impacted by fishing and other human activity (Bawe), and the near-pristine reef that is currently under protection from local influences but remains open subjected to global-scale impacts such as increasingly warm and acid waters (Chumbe). This allows us to investigate whether the fish community can be modeled more accurately in one or the other of these ecosystems, and also allows us to assess the importance of protection for the fish community.

CHAPTER 4: DATA AND RESEARCH METHODS

In this chapter we outline the research methodology, including methods used for both the collection and analysis of the data used.

4.1 Research approach

In order to answer the research question of this thesis, each of the sub-questions is answered in sequence. To answer sub-question A, a range of models are developed to predict the three fish community variables from the *in situ* habitat data. The accuracy of predictions from each model type is compared, and the most important habitat variables identified. To answer sub-question B, habitat variables from the remotely sensed data are extracted at a range of spatial scales, and characteristic scales of response, scales where the correlation between fish and habitat variable were strongest, are derived. To answer sub-question C, a reduced range of predictive models is subsequently developed to predict the fish community variables from the remotely sensed habitat data, and the identification of important habitat variables is repeated with this dataset. Two spatial scales are chosen for model development with remotely sensed habitat data, one using data at the finest scale offered by IKONOS data, the other using a subset of the IKONOS data that is limited to the spatial scales that would be available from Landsat TM data. The data collection, along with processing and analysis, is described below. Field and remote sensing data were collected for both Chumbe and Bawe reefs, using identical methods, so the two datasets are treated as one in the processing and described as such in the following.

4.2 *In situ* data

The methods used to answer sub-question A are described below in four sections. 4.2.1 and 4.2.2 describe the data on fish community and habitat, respectively, used in this study, and section 4.4 describes the statistical and algorithmic models developed, and their comparison. All *in situ* data were collected during a period of 3 months, between mid-September and mid-December 2007. The complete data set can be obtained by contacting the chair of the Department of Geography and Environmental Management at the University of Waterloo.

4.2.1 Fish data

Data on the fish community at each site were collected using a modified version of the point count method of Bohnsack and Bannerot (1986), yielding geolocated data of the abundance and minimum, mean and maximum fork length for each species present. Fieldwork at sites with water depth <8 m was carried out snorkelling; SCUBA was used for deeper sites. At each site, the observer rests passively on the surface (or sits passively on the bottom if using SCUBA) at the centre of the site. The observations are separated into two five-minute intervals. While approaching the site, and for the first five-minute interval, during which the observer slowly rotates to look in all directions, all fish species observed within a radius of 5 m of the centre of the site are noted on a dive slate. Only fish with fork length >5 cm were noted. During the second five-minute interval the number and average size (fork length) of individuals is noted for each species. If a species was observed as present during the first five minutes but cannot be found during the second five minutes, the number and average size is retrieved from memory. If a species is observed only during the second five-minute interval, it is not recorded. The location of each site was initially found by snorkelling in a random direction for a random number of fin kicks from the previous site. After more than 80% of the data had been collected on Chumbe, the locations of sites were plotted on a satellite image, and conspicuous habitats with no or few data points were specifically targeted for the remainder of the sites, along with the highly variable habitat along the reef edge. Logistical constraints for fieldwork on Bawe did not allow such planning, and field sites were clustered around anchoring sites near the reef, although care was taken to represent all major substrate types in the field data. This procedure results in a sampling design that is deliberately (but sub-optimally) stratified by substrate type, reef, and geomorphologic zone, habitat variables that are easy to derive from the satellite image. However, at the same time the sampling design is not deliberately stratified along other possible (and possibly more important) axes such as depth, live coral cover, structural complexity etc. The ultimate distribution of field sites, representing the compromise between ideal sampling design and logistical constraint, is illustrated in Figure 4.1 and Figure 4.2. Fish surveys were practised for a period of two weeks before data collection began, to become familiar with the methods and the fish species in the region. At any time, fieldwork was aborted if visibility fell below 8 m. Based on this dataset, the biomass for each species was calculated using the mean length and the equation:

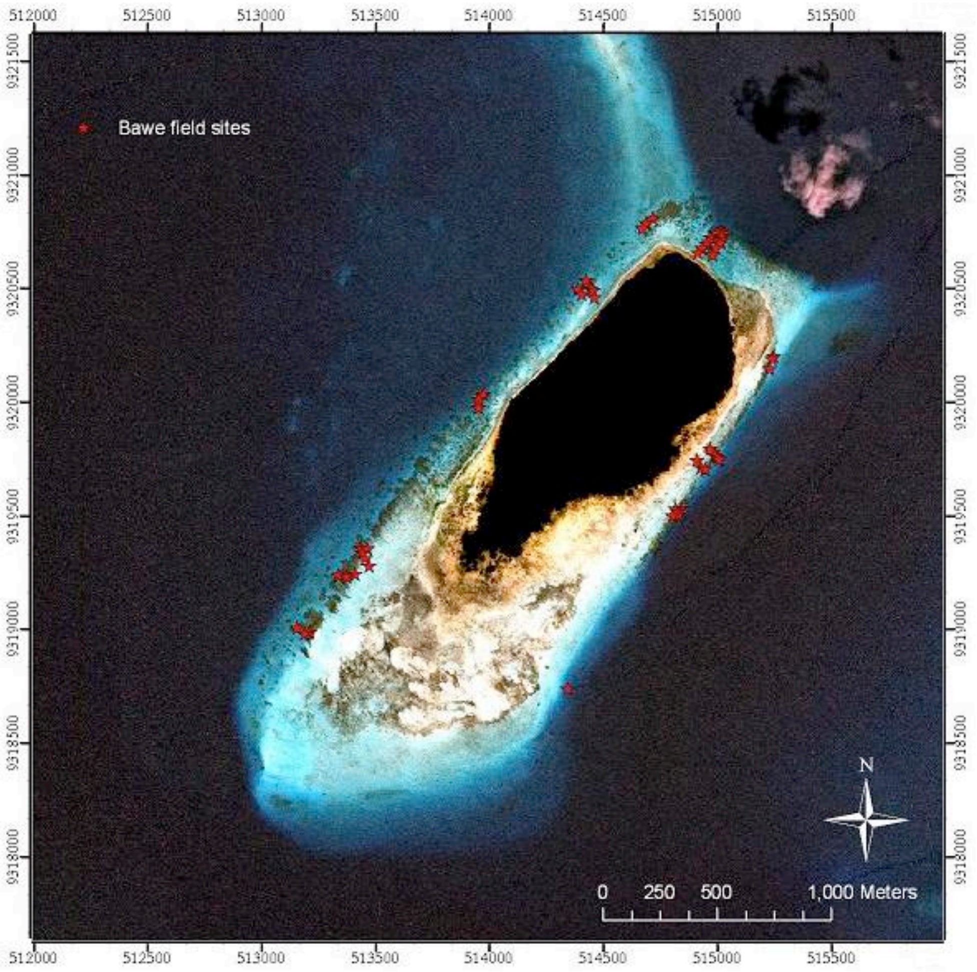

Figure 4.1: Field sites distributed in clusters on the reef area around Bawe Island.

$$W = A * L^{B} \text{ (eq. 1)}$$

where W is the weight of the fish in grams, L is the length of the fish in cm, and A (g/cm^3) and B (unitless) are parameters derived for each species from published values on Fishbase (Froese and Pauly 2008). When species-specific values for A and B were not available, average values from other species within the genera, or family, were used. Three often used measures of the fish community were then derived for each site. Total **biomass** was derived by adding the biomass values of each species, **species richness** was calculated as a simple count of species, and the **diversity** was derived by calculating Shannon's Diversity Index, expressed as:

$$H' = -\sum_{i=1}^{s} p_i * \ln(p_i) \text{ (eq. 2)}$$

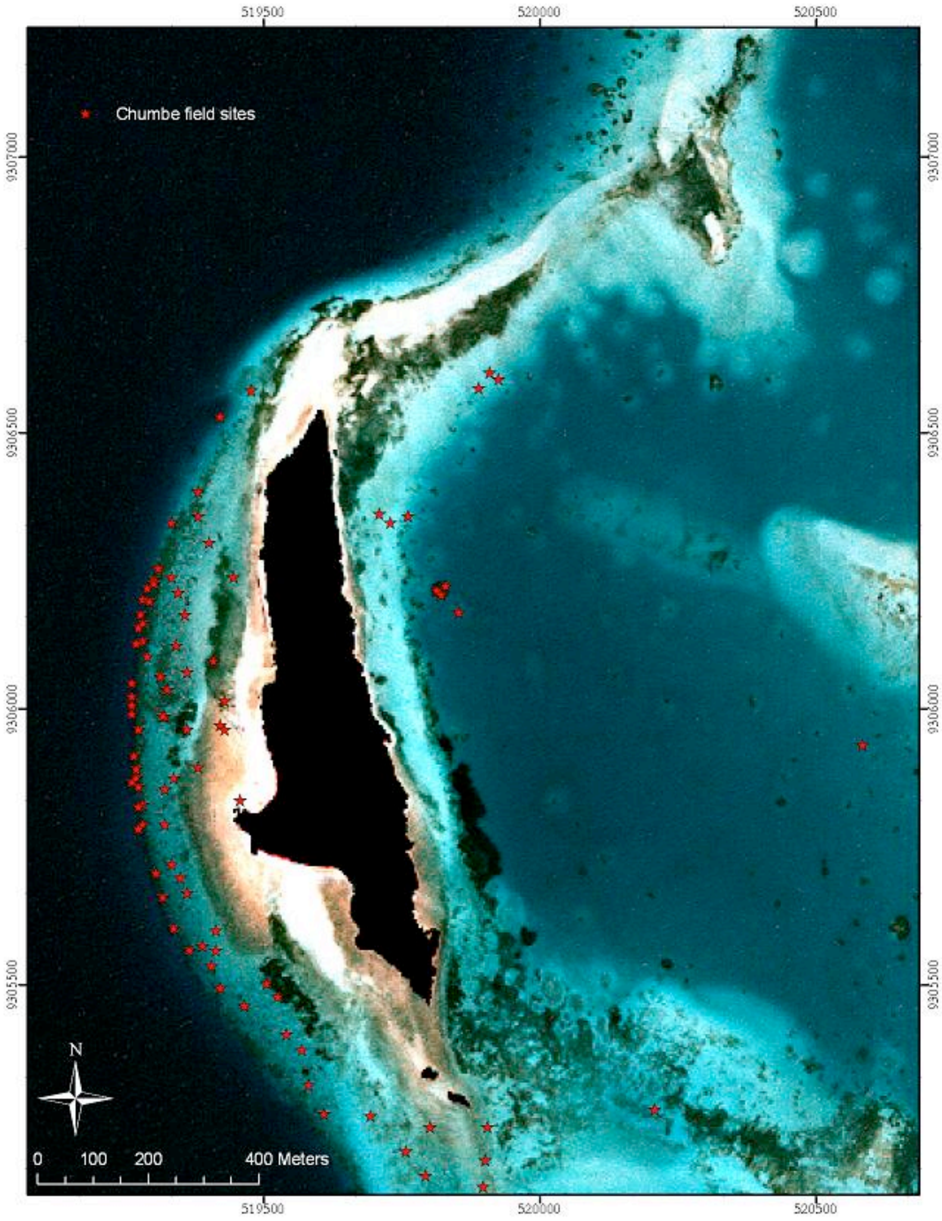

Figure 4.2: Field sites distributed on the reef around Chumbe Island.

where p_i is the proportion of the total biomass in each species, i is the species, and s is the total number of species (Ludwig and Reynolds 1988). Each species was represented by its biomass rather than the traditionally used abundance in equation 2, in order to improve the ecological relevance of the diversity measure (Wilhm 1968), and to reduce bias in the calculation, caused by schools of *Chromis spp.* with hundreds of individuals.

These three measures by no means constitute a full description of the fish community, and more targeted measures, e.g. presence/absence of species of interest, biomass and diversity of functional groups etc., can be derived as necessary in future studies. The flow of data processing is illustrated in Figure 4.3. A visual assessment of frequency distribution was carried out for each fish variable and log-transformations were applied as necessary to avoid extreme distributions.

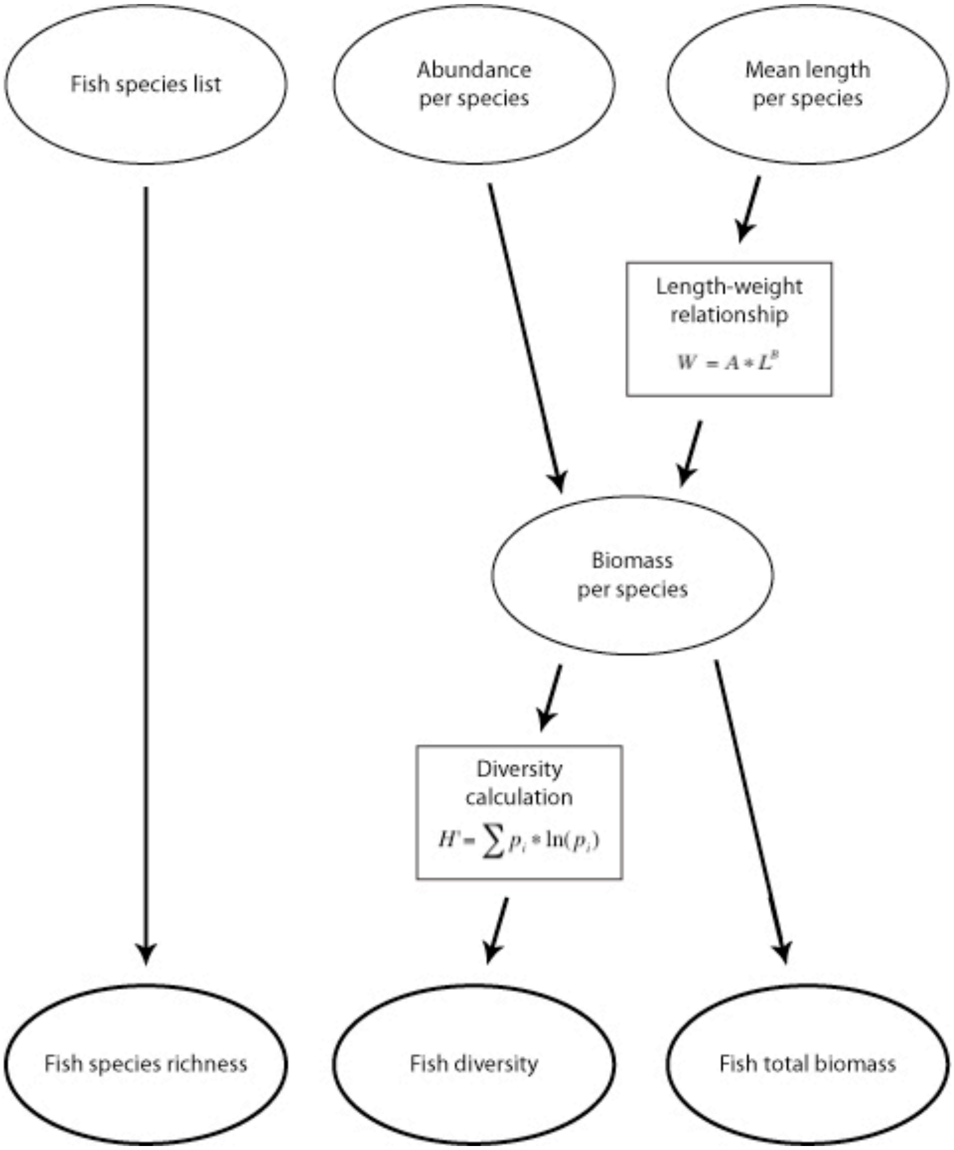

Figure 4.3: Processing flow for fish data. Data are in ellipses, processing steps in boxes. Some simple processing, e.g. counting the number of species in a list, has been omitted. Variables used in the subsequent analysis have bold borders.

4.2.2 Habitat data

At the sites of the fish point counts, a number of habitat data were also collected, following the fish survey. For each site, maximum and minimum depth was measured by placing a dive computer on the substrate for five seconds before recording the value. A Garmin Etrex GPS on a float was used for geolocation, and rugosity was assessed using a visual scale of 0-5 following the method of Wilson et al. (2007) with one modification. Two scales were adopted, the **coarse rugosity** scale focusing on rugosity caused by large reef elements such as small patch reefs and large boulder corals, the **fine rugosity** scale focusing on rugosity caused by small reef elements such as branching and digitate corals. Substrate photos covering the 5 metre radius were then taken, and time, visibility and current strength and direction were noted.

Depth data were transformed to depth at mean tide level by deriving the tidal stage at the time of data collection from local tide tables provided for Zanzibar port, and applying a simple correction. Based on the corrected depth data, the **average depth**, as well as the **depth range**, was calculated.

The substrate photos were then processed in CPCe (Kohler and Gill 2006), to derive the percentage cover of the following substrate types: **branching coral**, **digitate coral**, **massive coral**, **encrusting coral**, **foliose coral**, **turf algae**, **macroalgae**, **dead coral**, **sand**, **seagrass**, **rubble**, **pavement**, and **other** (mostly sponges). Based on the values of these variables, the total **algae** and total **live coral cover** of any growth form was calculated, the total **number of coral growth forms** was recorded, and the **substrate diversity** was calculated following equation 2, with p_i representing the proportion of each substrate type. In addition, **substrate evenness** was calculated as:

$$J = \frac{H'}{\ln(S)} \quad \text{(eq. 3)}$$

where S is the total number of substrate types present and H' is the Shannon diversity. Two further variables were added to the dataset, the **reef** variable ('Chumbe' or 'Bawe') describing where the data were collected, and the **conservation status** variable ('protected' or 'unprotected'). The flow of data processing is illustrated in Figure 4.4. A visual assessment of frequency distribution was carried out for each habitat variable and log-transformations were applied as necessary to avoid extreme distributions.

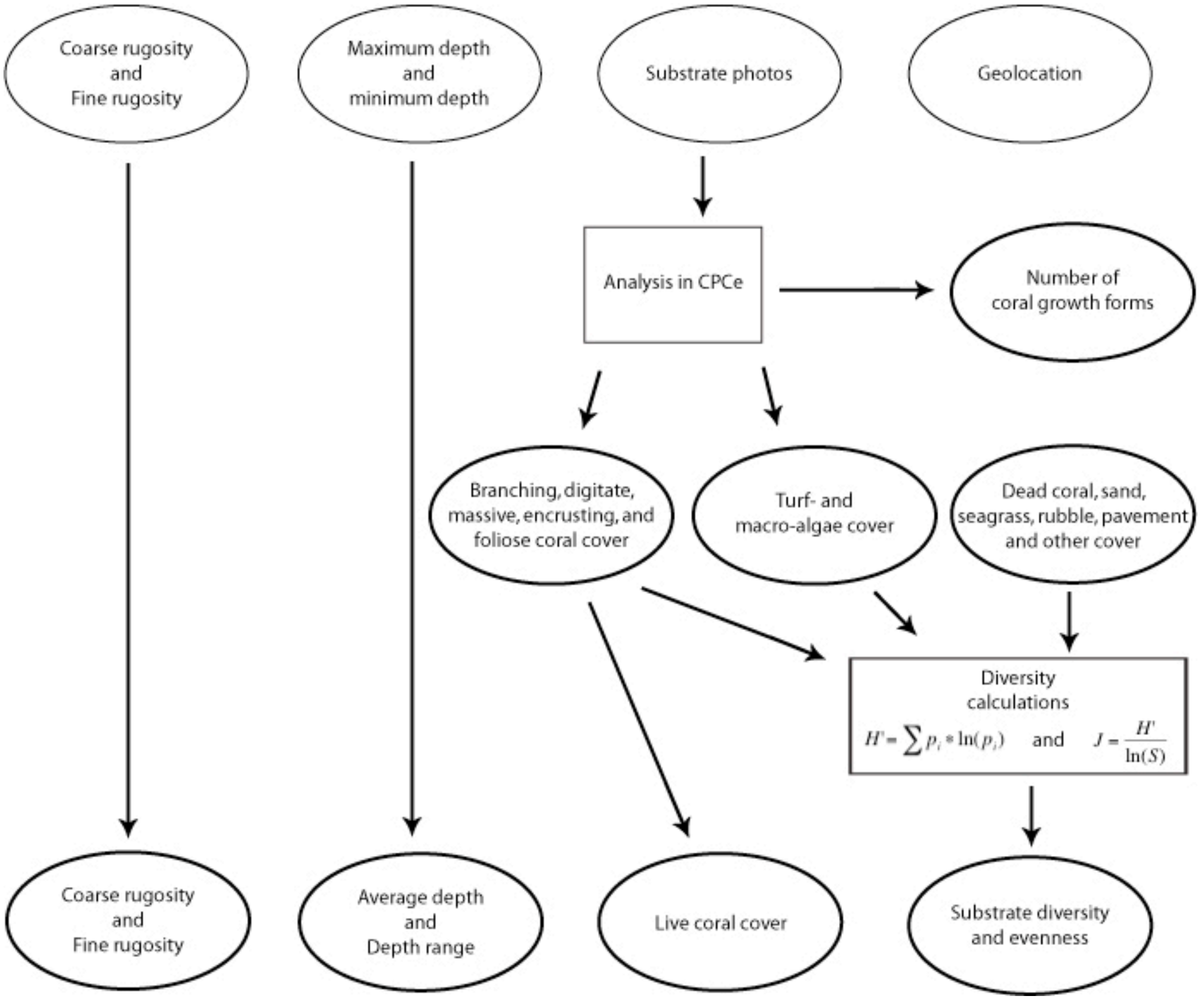
Coarse rugosity and Fine rugosity
Maximum depth and minimum depth
Substrate photos
Geolocation
Analysis in CPCe
Number of coral growth forms
Branching, digitate, massive, encrusting, and foliose coral cover
Turf- and macro-algae cover
Dead coral, sand, seagrass, rubble, pavement and other cover
Diversity calculations
$H' = \sum p_i * \ln(p_i)$ and $J = \frac{H'}{\ln(S)}$
Coarse rugosity and Fine rugosity
Average depth and Depth range
Live coral cover
Substrate diversity and evenness

Figure 4.4: Processing flow for habitat data. Graphic conventions are similar to those used in Figure 4.3.

4.2.3 Note on geolocation

The Garmin eTrex used for geolocation of the field sites has an estimated horizontal accuracy of 5-7 m. Many of the shallow-water areas in the study area have substrates with small patch sizes compared to this accuracy. The GPS-based geolocation of all field sites was therefore assessed against the true-colour composite of the IKONOS image, and corrected if possible or discarded if necessary (Phinn et al. 2008). Correction was carried out where an obvious fit between field and satellite data could be obtained by moving the field site no more than 6 m so the error was clearly identifiable and the correction justified. Examples include a field site GPS-located in the deep sandy area just off the reef edge while notes state that the site was in a coral-dominated area on the reef edge, or a site with 100% seagrass cover, GPS-located a few metres away from a dense seagrass bed on the imagery. In order to avoid having to make such corrections it is suggested that, whenever possible, a differential GPS system is used to geolocate field data in similarly patchy environments.

4.3 Remote sensing data

The following sections describe the satellite-derived data used, in combination with the field data, to answer sub-questions B and C. The satellite data were used to produce a set of remotely sensed habitat variables, which can function as independent variables in predictive models. Each of these new independent variables will be outlined in the following section.

IKONOS data were used in this study because they combine relative affordability (as opposed to airborne hyperspectral data) with high spatial resolution (as opposed to Landsat data). IKONOS data would therefore be a reasonable choice of remote sensing data for organisations or project with a serious desire to map one or more reefs in great detail (Andréfouët et al. 2003; Mumby and Edwards 2002). In addition to producing the set of remotely sensed habitat variables at the best spatial resolution IKONOS data offer, a subset of the IKONOS data limited to the spatial scales that would be available with Landsat TM data, was created by a simple coarsening of the spatial scale of each habitat variable to match the spatial resolution of Landsat data. The coarsening of the spatial scale was applied after derivation of each habitat variable, and therefore does not take into account the different classification accuracy, spectral resolution, or atmospheric and geometric accuracy that would have been achievable with a real Landsat dataset. It is thus likely to produce optimistic estimates of what would be obtained from real Landsat data. In the following, this subset of the IKONOS data, at Landsat TM spatial scales, is referred to as “simulated Landsat data”.

4.3.1 Satellite imagery

For each of the habitat variables measured *in situ*, it was considered whether it would be possible to estimate the variables using IKONOS data. Some variables, e.g. coral growth forms, have been shown by previous investigations to influence the reflectance spectra when these are measured *in situ*, probably as a result of internal shading (Joyce and Phinn 2002; Minghelli-Roman et al. 2002). However, the internal shading is not always detectable (Holden and LeDrew 1999), and it is not clear how upscaling to airborne or satellite measurements could be achieved. Other variables, as described in chapter 2, are more easily mappable with remote sensing. This section describes the satellite imagery used in this study, the image processing applied, and the habitat variables derived from the imagery.

Two IKONOS images were used for this study. The image covering Bawe Island (henceforth: Bawe image) was acquired on 31 October, 2005, at a 19º off-nadir angle, with nearly cloud-free conditions in the area of interest. The image covering Chumbe Island (henceforth: Chumbe image) was acquired on 20 October, 2007, at a 20º off-nadir angle, with perfectly cloud-free conditions in the area of interest. Both images were provided in GeoTIFF format, at full radiometric resolution (11 bit), at a pixel size of 4 m, re-sampled from the original data using the cubic convolution method. Both images are cloud-free in the areas of interest, and recorded on days of good water clarity (features known to be at depths >10 m are distinguishable in both images).

The Bawe image predates the field data by 2 years, whereas the Chumbe image was acquired during the period of field data collection. The coral reef on Bawe, and its fish community, may have changed during this two-year period, introducing an error when mapping the habitat variables of the reef, sampled in 2007, using this image. However, change on reefs happens slowly in the absence of major disturbances or phase shifts. No major disturbance, such as a severe storm, a mass bleaching event or a Crown-of-Thorns Starfish (COTS) outbreak, has been observed on Bawe in the period between October 2005 and October 2007. This absence of substantial disturbance justifies the use of the image in this study.

4.3.2 Processing

The two images were processed using very similar procedures, described in detail below.

4.3.2.1 Datum, projection, and geometric correction

The choice of datum for this study was difficult. The Zanzibari government uses the Arc1960 datum, and therefore this datum is also adopted by Zanzibari research institutions and research produced by development projects carried out in collaboration with Zanzibari government partners. Most research by foreign research groups, however, is carried out using the WGS-84 datum, and this was also the datum used for the original satellite imagery. ArcGIS provides a tool for conversion between the two datums, but simple tests using this tool showed that it produced large errors in the areas of interest to this study. Ultimately, the difficulty with conversion meant that it was necessary to use the WGS-84 datum for this research. UTM was chosen as the projection and coordinate system.

Geometric correction of both images was performed in ENVI (ITT Visual Information Solutions 2007), using more than 30 field-collected ground control points (GCPs) for each image. These were collected with a Garmin eTrex handheld GPS, with a typical absolute accuracy, as estimated by the GPS unit, of 5-7 m. For the Chumbe image, the panchromatic band was available in addition to the four monochromatic bands (Blue, Green, Red, Near-Infrared), and the higher resolution of the panchromatic band allowed more small and recognisable features to be used for GCPs, ultimately producing a very high rectification accuracy of RMSE=0.96 m. For the Bawe image, only the monochromatic bands were available, and a rectification accuracy of RMSE=2.52 m was achieved. Considering the absolute accuracy of the GPS unit, these rectification accuracies are very acceptable.

4.3.2.2 Atmospheric correction

Both images were atmospherically corrected to produce surface reflectance values using the Atcor2 algorithm as implemented in Geomatica (PCI Geomatics 2003) with the latest available calibration coefficients (Spaceimaging 2001). The use of advanced atmospheric correction algorithms (Atcor2 is based on MODTRAN code) can be problematic when lack of information about the state of the atmosphere requires the use of standard atmospheric models which may or may not be appropriate. For example, in some cases inappropriate parameterization can lead to obvious errors such as negative surface reflectance values. Simpler methods such as dark pixel subtraction may therefore be preferable. In this study, Atcor 2 was implemented using the “Tropical Maritime” standard atmosphere, 20 km visibility. The "tropical" is an obvious choice, and the coastal location of the study site, on the border between the Indian Ocean and the African landmass, justifies the "maritime" choice. The 20 km visibility was based on experience from the area (from Chumbe Island the African mainland is 25 km away and can be seen on rare occasions). The adjacency value was left at its default 1 km value. The resulting surface reflectance values were checked for negative values (there were none) and evaluated against typical values of recognizable substrate types, with which they fit well.

4.3.2.3 Deglinting

The images were then corrected for specular reflection of sunlight off the sea surface, also called sun-glint, a serious confounding factor for remote sensing when the sea surface is not flat. The method of Hedley et al. (2005) was chosen for its robustness and ease of application. This method is a modification of an original method (Hochberg et al. 2003a), which relies on the high absorption in water of Near-Infrared (NIR) radiation. The method assumes that NIR radiation recorded by the sensor is composed of specular reflection off the water surface, in addition to a small amount of ambient "noise", and that the amount of specular reflection in other wavelengths is proportional to that found for NIR radiation. The exact proportional relationships between specular reflection in the NIR and other regions is found by calibration in an area of deep water, where no influence from the substrate is ensured, and then used to correct reflectance values in the non-NIR bands in the entire image.

4.3.2.4 Water column correction

As sunlight passes through the water column, both before and after its reflection off the substrate, it is attenuated by dissolved and particulate matter in the water, and by the water itself. This attenuation depends on the water constituents and the depth. Attenuation reduces the intensity of the light, and because it is wavelength-dependent it also changes the light's spectral composition. Because the focus in remote sensing of coral reefs is typically the substrate, not the water itself, this attenuation needs to be corrected for.

A simple and useful method to perform this correction was developed by Lyzenga (1978; 1981). The intensity of light will, according to Beer's Law, decay exponentially with increasing depth. Lyzenga's method applies a natural logarithm to reflectance values recorded by the sensor, in order to linearise the relationship between reflectance and depth (radiance values could be used instead of reflectance values to achieve the same results). A well-defined substrate is then chosen, typically bare sand, and log-transformed reflectance values are derived, from two bands, from a range of depths. These values are then used to create a bi-plot, where the values from the band with the longest wavelength are put on the x-axis and values from the other band on the y-axis. If the water has been homogeneous in the area chosen, and the substrate well-defined, the points will form a line. The slope of this line equals the ratio of attenuation coefficients for the two chosen bands, and this ratio can be used to calculate Lyzenga's depth-invariant index (DII) of bottom type for all pixels in the image, using the following equation:

$$DII_{ij} = \ln(R_i) - ((\frac{k_i}{k_j}) \times \ln(R_j) \quad \text{(eq.4)}$$

where R is reflectance values (unitless), i and j are the two bands in question, and k_i/k_j is the ratio of attenuation coefficients (units cancel out). The DII has been shown to improve the discrimination of bottom types on coral reefs (Mumby et al. 1998), and has become a standard image processing component in remote sensing studies of nearshore environments (Green et al. 2000). The image-wide application of the DII assumed a homogeneous water body, an assumption that is clearly violated if turbidity varies, e.g. due to inputs from rivers or urban areas. However, without spatially distributed measurements of water constituents at the time of image acquisition, remote sensing of the benthos in areas of varying depth relies on this assumption, and it is up to the investigator to assess the feasibility of the approach. In this study, no heterogeneity was observed in the water for either image, and the water column correction was applied as described above.

4.3.3 Substrate classifications

Substrate classifications, often called habitat maps, were developed for each image using both the field data described above, and an additional set of field data collected purposely for substrate classification. 302 and 425 field observations were available for classification of the Bawe and Chumbe images, respectively. For each image, half of the field observations were used to develop a Maximum

Likelihood classifier (Figure 4.5 and Figure 4.6), and the other half was used to assess the accuracy of the resulting classification. The Maximum Likelihood classifier was chosen based on a comparison with Minimum Distance and Parallelepiped classifiers in which it produced the highest overall classification accuracy (79.5% vs. 72.1% for Minimum Distance and 70.8% for Parallelepiped, for Chumbe Island).

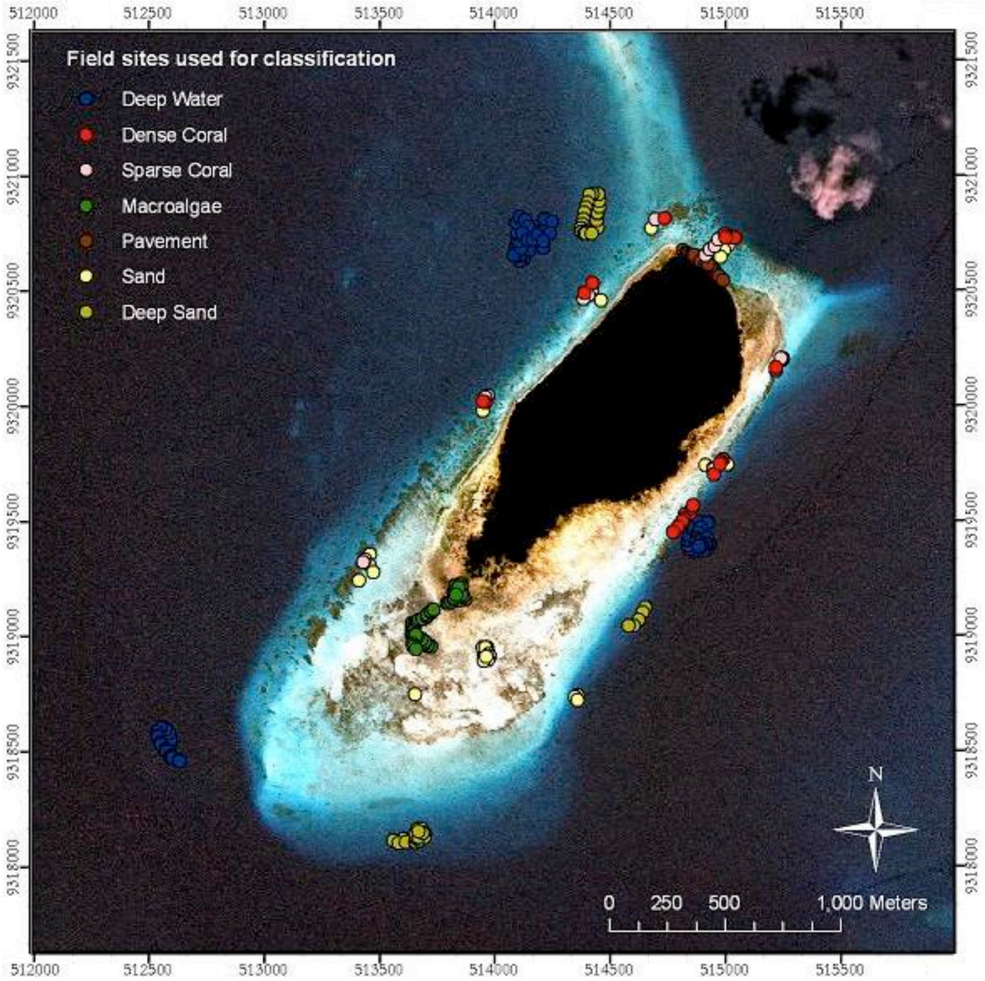

Figure 4.5: Field sites used for classification of the Bawe image.

The classes used for each image classification differ slightly, to reflect the difference in dominant substrate types on each reef. Both classifications used 'Deep Water', 'Dense Coral' (>40% coral cover), 'Sparse Coral' (5%-40% coral cover), 'Pavement', 'Sand' (depth<5m) and 'Deep Sand' (depth>5m). In addition to these classes, 'Dense Seagrass' (aboveground seagrass biomass>250g/m^2) and 'Sparse Seagrass' (aboveground seagrass biomass 5-250g/m^2) were added to the

Chumbe classification, as seagrass beds of varying density are found all around Chumbe Island. In addition, ‘Exposed Sand’ was added to the Chumbe classification, as the spectral signature of these areas differed substantially from sandy areas covered by even very shallow water. Neither seagrass nor exposed sand was found on Bawe, where instead the ‘Macroalgae’ class was included, as brown erect macroalgae were found to dominate the substrate in large areas. These classes were used to cover all ecologically important substrate types, and at the same time ensure a high spectral separability between the classes.

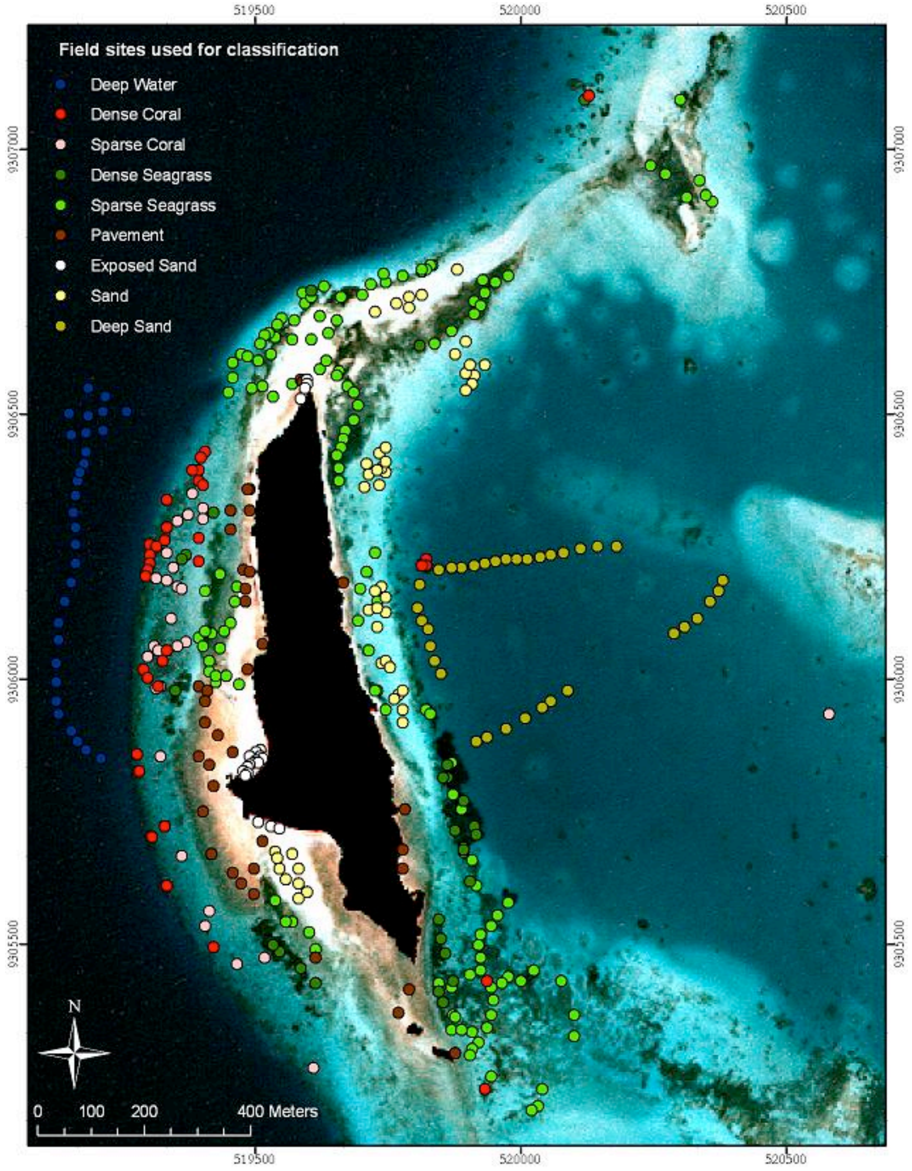

Figure 4.6: Field sites used for classification of the Chumbe image.

Despite the use of the DII as the basis for the classifications, which would ideally avoid the need for a separate class for sand at greater depth, the 'Deep Sand' class was found to improve the identification of sandy substrates at depths greater than 5 m. After classification both this class and the 'Exposed Sand' class were combined with the 'Sand' class to form a combined class for sand substrates regardless of depth. Based on the substrate classifications, a new independent variable was developed for each field site, to designate the substrate type at the site.

4.3.4 Determination of geomorphologic zones

The degree of reef development varies between different areas on the two reefs, but both reefs have one section of somewhat developed fringing reef, with areas that can be separated into those on or very near the reef crest, and those on the reef flat. On Bawe, this area extends along the northern and western sides of the island, and on Chumbe along the western side. A narrow forereef also exists, but coral growth ceases very rapidly with increasing water depth and the developed part of the forereef is typically only a few metres wide. It was therefore not included as a geomorphologic zone in this study. The reef could therefore be separated into three distinct zones, the reef crest, the reef flat, and areas with no recognisable reef zones. An independent variable was developed, to designate the geomorphologic zone that each field site was located within.

4.3.5 Estimation of live coral cover

Most studies involving remote sensing of coral reefs have mapped coral cover through a straight-forward classification separating "coral" or "coral-dominated" areas from other areas. Though such maps may be useful for some purposes, the inability to differentiate between areas with widely differing live coral cover would be a drawback in this study. Other studies have sought greater detail by defining multiple classes on the basis of coral cover. Isoun et al. (2003) used seven such classes and achieved a remarkable classification accuracy (77%), though this was only possible using narrow bands (10 nm FWHM) optimized for discrimination of reef benthos and captured from a low-flying aircraft. Newman et al. (2007) used four broader classes and two IKONOS images, and achieved classification accuracies of 78% and 81%. However, for this study a more detailed discrimination of live coral cover was desirable. Joyce (2004b) obtained such detail by developing correlations between live coral cover and reflectance ratios/derivatives obtained from airborne hyperspectral data. She found that the optimum band ratio/derivative, i.e. the one producing the highest correlation with live coral cover, varied between blue and brown coral types (Hochberg et al. 2003b), but obtained a (Pearson) correlation coefficient of R=-0.76 between the optimum band ratio and live coral cover of both colour types, in shallow areas of Heron Reef, Australia. However, a band-ratio approach may suffer from the influence of several spectrally similar non-coral substrates such as seagrasses and algae. Spectral unmixing of hyperspectral data can in theory enable differentiation of these substrate types and produce a live coral cover estimate for individual pixels (Hedley and Mumby 2003; Hedley et al. 2004), but problems if depth variation and noise remain (Mumby et al. 2004b). Even if these are solved, spectral unmixing of these substrate types is only feasible using hyperspectral data, which are not currently available to most users.

In order to achieve the highest level of detail in the mapping of live coral cover, while being limited to IKONOS data, this study applied an approach similar to that employed by Joyce (2004b). To produce a spatially distributed estimate of live coral cover, the DII was extracted from each field site where coral cover was higher than 5% (i.e. sites that formed the basis for the 'Dense Coral' and 'Sparse Coral' substrate classes described above). A linear model was then fitted to the two variables using half the data points, and the other half of the data points were used for accuracy estimation. The model was then used to produce estimates of coral cover for all pixels classified in the two coral classes, while pixels classified in non-coral classes were excluded. This was carried out for the two reefs independently, as the relationship between the DII and live coral cover depends on the apparent optical properties of the water overlying the reef, and so must be calibrated individually for each image.

The spatial scale of the live coral cover estimates was then varied by applying averaging filter of varying kernel sizes to the DII values on each reef. The remotely sensed live coral cover estimations at a range of spatial scales were added as independent variables in the set of remotely sensed habitat variables. Although these measurements are based on square pixel windows, the term "radius" has been used for consistency of description between variables. The "radius" of these windows is considered to be half the side length of the square.

4.3.6 Estimation of depth and rugosity

Depth was calculated for each pixel using the ratio algorithm of Stumpf et al. (2003). This method, like the calculation of the DII, relies on the wavelength dependency of water attenuation. Longer wavelengths attenuate more rapidly (Mobley 1994), hence substrates located in deeper water will show a greater proportion of reflected light in shorter wavelengths. This property is used to develop a linear transformation of the ratio between log-transformed reflectances, using known depth values for calibration (Stumpf et al. 2003):

$$Depth = m_1 \frac{\ln(nR_i)}{\ln(nR_j)} - m_0 \text{ (eq.5)}$$

where R is reflectance (unitless), i and j are the two bands used for the depth estimation, and n, m_o and m_1 are manually tuneable constants. n is chosen to ensure only positive logarithms and a linear response with depth, m_0 and m_1 are optimized iteratively to minimize error. The ratio algorithm has been developed and tested to work on bottom types with different albedos, and is the state-of-the-art method for bathymetric mapping with multispectral remote sensing data. In this study we used depth measurements from the same field observations used for the classification and its accuracy assessment to tune the constants. The remotely sensed depth estimation was added as an independent variable in the set of remotely sensed habitat variables.

The structural complexity in different parts of the reef can be calculated by using the spatially distributed depth estimates produced by equation 5. In this study, structural complexity was quantified as area-based rugosity, calculated as the actual surface area divided by the area of a hypothetical flat surface covering the same area. This is a straight-forward three-dimensional extension of the two-dimensional rugosity typically calculated *in situ* using the chain method (Luckhurst and Luckhurst 1978; Risk 1972). NOAA's Benthic Terrain Modeler, based on algorithms by Jenness (2002), was used for the calculations. The Benthic Terrain Modeler builds a Triangulated Irregular Network (TIN) from the depth estimates produced by equation 5. For a given point in the TIN, the actual surface area is calculated using a 3x3 pixel window by adding the surface area of all portions of individual triangles that fall within the boundaries of the centre pixel (Jenness 2002) (Figure 4.7).

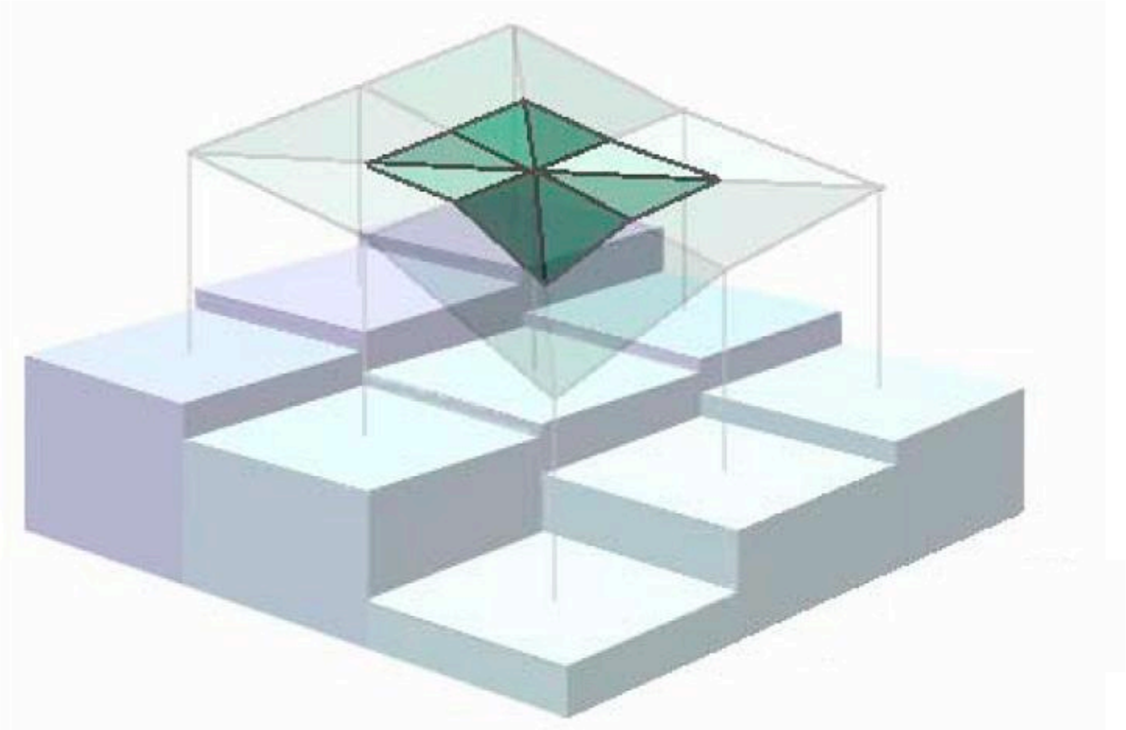

Figure 4.7: A graphical illustration of the principle behind the rugosity calculations of the Benthic Terrain Modeler (Jenness 2002).

The spatial scale of the rugosity estimates was varied by gradually coarsening the spatial resolution of the depth estimations. Rugosity estimates from each spatial scale were added as independent variables in the set of remotely sensed habitat variables. As for live coral cover, the term "radius" has been used for consistency of description between variables, considering the "radius" to be half the side length of the square.

4.3.7 Estimation of habitat variety

Habitat variety was derived from the substrate classifications described above, using two separate metrics, one simply calculating the number of substrate types within a given radius of the point ("habitat richness"), the other calculating Shannon's diversity index (eq. 2) with p_i being the proportion of the substrate type 'i' within a set radius around the data point ("habitat diversity"). Focal statistics in ArcGIS were used to calculate habitat richness, and NOAA's Diversity Calculator, an extension to ArcGIS, was used for the point-calculations of habitat diversity (Buja 2008). For the habitat diversity calculations, the substrate classifications were first transformed into vectorised polygons, with contiguous pixels of the same substrate class forming each polygon. Figure 4.8 illustrates the geometry of both measures of habitat variety. Unfortunately the Diversity Calculator was unable to calculate the habitat diversity measure for image layers, so a custom Python script was developed for this purpose. This script is available from the author on request (knudby@gmail.com).

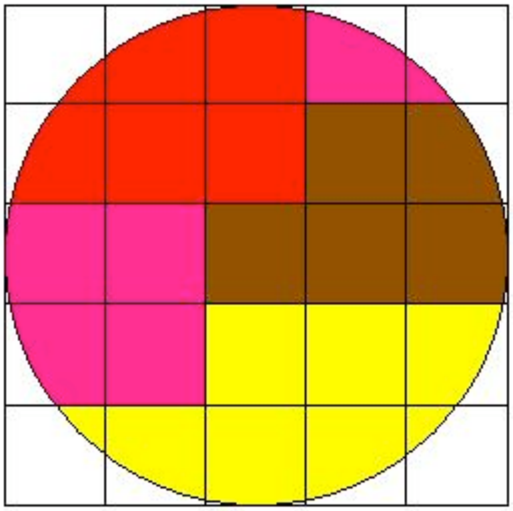

Figure 4.8: The geometry behind habitat variety calculations. Each colour represents a separate substrate type. With an IKONOS pixel size of 4 m, this example illustrates the calculation with a radius of 10 m.

The values of both habitat variety metrics depend on the level of detail in the substrate classification used as input, as more classes will lead to higher values in both metrics. In addition, more detailed breakdown of density classes within a given substrate type (e.g. if five coral classes were used instead of two) will increase output values in areas with such substrate types (e.g. coral areas). Much manipulation is therefore possible when designing the substrate classification that forms the input of the habitat variety calculations. For this study, the substrate classes were chosen partly on the basis of spectral separability, without which the accuracy of the habitat maps decreases and all subsequent calculations similarly lose accuracy. In addition, classes were chosen to cover what was considered the ecologically important substrate types in the habitat maps, which can be considered necessary for subsequent derivation of ecologically meaningful habitat variety indices.

Due to the different number, and kind, of classes used in classifications of the two reefs, the variety indices are not perfectly comparable (Kendall and Miller 2008). However, the different number of classes reflects a real difference in the number of major substrate types present on each reef, and calculations based on the substrate classifications described above are therefore justified. The different kinds of substrates (macroalgae on Bawe and seagrass on Chumbe), will similarly influence the relationship between the variety indices and the fish community, as the two habitat types provide different contributions to the ecology of the reef. This issue has not been addressed here, but is worthy of further exploration. Neither has the uneven difference between substrate types been addressed (e.g. 'Dense Coral' is more different from 'Exposed Sand' than it is from 'Sparse Coral') (Mumby 2001; Pittman et al. 2007).

Table 4.1: List of variables derived from field data. Log-transformations indicated by *.

Fish community	*In situ* habitat	Remotely sensed habitat
Species richness	Branching coral cover*	Substrate class
Biomass*	Digitate coral cover*	Geomorphologic zone
Diversity	Massive coral cover*	Depth
	Encrusting coral cover*	Live coral cover, 2-26 m radius
	Foliose coral cover*	Rugosity, 6-300 m radius*
	Live coral cover	Habitat richness, 5-80 m radius
	Dead coral cover	Habitat richness, 90-200 m radius*
	Number of coral growth forms*	Habitat diversity, 5-10 m radius*
	Turf algae cover*	Habitat diversity, 20-60 m radius
	Macroalgae cover*	Habitat diversity, 70-200 m radius*
	Total algae cover*	
	Sand cover*	
	Seagrass cover*	
	Rubble cover*	
	Pavement cover*	
	Other cover*	
	Coarse rugosity	
	Fine rugosity	
	Average depth*	
	Depth range*	
	Substrate diversity	
	Substrate evenness	
	Reef	
	Protection status	

The spatial scale of the habitat variety estimates was varied by changing the radius used for their calculation. Habitat variety estimates at a range of spatial scales were then added as independent variables in the set of remotely sensed habitat variables.

A visual assessment of frequency distribution was carried out for each remotely sensed habitat variable and log-transformations were applied as necessary to avoid extreme distributions. A complete list of the variables used in this study, and their transformation, is found in Table 4.1.

4.4 Predictive models

Two types of models were developed, some with explicit assumptions about the nature of the statistical relationship between independent and dependent variables (LM and GAM), others based on algorithmic model development with no prior assumptions about the nature of the relationship (tree-based models and the support vector machine) (Breiman 2001b). The models are described below. All models and related data processing were implemented in the free statistics software package "R" (R Core Development Team 2008), and its contributed packages.

4.4.1 Linear Model (LM)

As discussed in chapter 2, linear models are the most prominent in the literature relating measures of fish communities to their habitat. The linear model was here developed as a multiple linear regression model, as implemented in R's 'stats' package (R Core Development Team 2008). The model is developed through combined forward and backward variable selection, and the Akaike Information Criterion (AIC) is used to determine variable inclusion/exclusion (Akaike 1974).

4.4.2 General Additive Model (GAM)

The general additive model is an extension of the LM, allowing individual independent variables to be transformed before addition to the model. Any transformation can theoretically be included in a GAM, however, in this study only cubic smoothing splines have been used. The transformations can improve the predictive model when relationships between habitat variables and the fish community are non-linear. Also here the model is developed through combined forward and backward variable selection, and the AIC is used to determine both variable inclusion/exclusion and, if a variable is included, whether it should be included linearly or non-linearly. The transformations applied in this study are local spline smoothers of two equivalent degrees of freedom. Due to computational limitations, the number of independent variables for the GAM models had to be reduced to a maximum of 17. Exclusion of *in situ* variables was based on suspected co-linearities in the variables, and on results from the existing literature.

Table 4.2: List of *in situ* and remotely sensed habitat variables excluded from GAM models.

***In situ* and remotely sensed habitat variables omitted from GAM models**	
Foliose coral cover	Live coral cover (6, 14, 22 m)
Encrusting coral cover	Rugosity (12, 18, 24, 30, 60, 75, 112.5, 225, 300 m)
Algae cover	Habitat diversity (5, 20, 30, 40, 60, 70, 90, 100, 150, 175, 200 m)
Pavement cover	Habitat richness (all scales)
Seagrass cover	
Other cover	
Substrate evenness	

The excluded remotely sensed variables were chosen to allow similar spatial scales to remain in the model. Remotely sensed habitat richness was completely excluded from the models to allow for a variety of habitat diversity scales to be included. The variables listed in Table 4.2 were excluded from this model type. The GAM model was implemented in R's 'gam' package (Hastie 2008).

4.4.3 Tree-based models

A number of tree-based models have also been used in this study. Regression trees are constructed by recursively splitting the dataset into two subsets using any possible split, according to a decision rule such as achieving maximum homogeneity, typically defined as maximum reduction of RMSE. The partitioning is continued until a stopping criterion is met or until a partition consists of only one observation. The mean response value for each group is then assigned as the predicted value for all observations in the group. The ultimate size of the groups, and the number of splits, are determined by the model developer (De'ath and Fabricius 2000). Because of their structure, tree-based models deal efficiently with non-linearities including non-smooth functions, and they are also well suited to model interaction effects. However, they are also sensitive to small changes in the training data (Hastie et al. 2001), so a variety of methods have been used to stabilize them. Three of these have been employed in this study.

4.4.3.1 Bagging

The first tree-based method used here, called bootstrap aggregating (Bagging), trains multiple regression trees using bootstrap samples as training sets, and then averages the predictions from each tree to arrive at the bagged predictions (Breiman 1996). The Bagging model is implemented R’s ‘ipred’ package (Peters and Hothorn 2007), with 100 individual trees grown until their nodes are pure (i.e. consisting of a single observation).

4.4.3.2 Random forest

Another tree-based approach is the Random forest, which also trains multiple regression trees using bootstrapped training sets. However, Random forests differ from Bagging in that each split is determined using only a random sub-sample of the available independent variables (Breiman 2001a). The Random forest model is implemented in R’s ‘randomForest’ package (Liaw and Wiener 2002) with the default settings (500 trees, one third of the explanatory variables examined in each split).

4.4.3.3 Boosted trees

The last tree-based approach used in this study, Boosted trees, develops multiple regression trees by iteratively fitting new trees to the prediction errors of the existing tree assemblage. Existing trees are not changed through iterations, and the final model is a linear combination of all the trees in the assemblage (Elith et al. 2008). Hyperparameters for the Boosted trees (number of trees used and the ‘shrinkage’ variable) were tuned with internal cross-validation, and the Boosted trees model is implemented using R’s ‘gbm’ (Generalized Boosted regression Models) package (Ridgeway 2007).

4.4.4 Support Vector Machine (SVM)

The last model developed was the SVM, a novel machine-learning technique that can represent nonlinear effects and interactions between variables. It projects the explanatory variables into a higher-dimensional feature space, where the prediction problem has a linear solution (Moguerza and Muñoz 2006). The γ and ε parameters of the support vector machine are tuned automatically using an internal cross-validation procedure. We implement a complete grid search over (ε,γ), evaluating each parameter in 7 exponential steps from 0.001 to 1.0. The SVM model is implemented with libsvm in R's 'e1071' package (Chang and Lin 2009).

4.5 Model comparison

Models were processed for each dependent variable at the three spatial scales, using *in situ* data, IKONOS data, and simulated Landsat data. In addition, one set of models was developed using a combination of all datasets. The models were compared in terms of their prediction accuracy and precision. Accuracy was quantified as RMSE, and precision as standard deviation around the mean error. The models are known to be different because of their different structures, and a statistically significant difference between model accuracies would thus always be obtainable by increasing the number of model runs (Brenning 2009). Significance tests were therefore applied only to test whether the differences between models observed with the 100 repetitions used in this study were non-random. Paired *t*-tests with unequal variances were used, and the Simes procedure was applied to keep the false-discovery rate (FDR) at the 5% level (Benjamini and Hochberg 1995; Simes 1986). In addition, the models were compared in terms of their interpretability, specifically their ability to identify important independent variables. 3 data points, the only ones classified in the 'pavement' substrate class, had to be removed from the processing based on remotely sensed data to avoid processing errors.

4.5.1 Resampling methods

Two resampling methods were used for accuracy estimation, boot-strap and cross-validation (Efron and Tibshirani 1993). Resampling methods are used to provide most honest estimates of the prediction error that can be expected when a given model is applied to new data – data that were not used to developed the model. Due to the issue of resampling with replacement in boot-strapping, which is a greater problem in relatively small datasets (here n=144), the results obtained using cross-validation are used as primary results, though results obtained using boot-strapping will be discussed when differing from the primary results. Estimates of prediction accuracy were calculated using methods as similar as possible given this range of model types, as explained below.

4.5.1.1 Bootstrapping

Boot-strapping addresses the problem of estimating how a predictive model will perform on a future, yet unknown, data set. This is done by generating independently samples training and test data sets, randomly drawing points from the original data set, with replacement, to until both training and test sets contain the same number of observations as the original data set. The training set is then used for fitting a model, and tested on the test set. The procedure is repeated a number of times, here 100, and the performance measure from each repetition is averaged to estimate the model's predictive performance. In this study, predictive performance is quantified as the Root Mean Square Error (RMSE) of predictions. In each repetition, we use the same training and test samples for all modelling techniques so detecting pairwise differences in the performances of different techniques is a paired-sample problem.

4.5.1.2 Cross-validation

Because of the drawing with replacement involved in boot-strapping, a number of data points will be included in both the training and test sets, which is likely to overestimate the performance of the model on a future dataset. Another method, cross-validation, designates the training and test sets differently, with the aim of providing a more realistic estimation of the predictive performance of the model. In cross-validation, the dataset is split randomly into a number of groups (k) of equal size, here k=10 is used. A training set is then formed by combining all except one of these groups, with the last group forming the test set to derive performance measures (Efron and Gong 1983; Efron and Tibshirani 1993). In this way, the training and test sets contain completely different data points. Cross-validation thus works similarly to the approach typically taken with large data sets, separating the data into training and test sets, but it does so in a manner that retains 90% of the total data for model development. This makes it less “wasteful”, particularly important to use with small data sets such as the one used in this thesis (n=144). As in bootstrapping, the procedure of splitting the data set into training and test sets and assessing prediction error is repeated 100 times, and the performance measure (RMSE) is averaged to get a true estimate of the performance of the model.

4.5.2 Spatial autocorrelation

Both fish and habitat data from coral reefs are almost certain to be spatially autocorrelated. If the spatial autocorrelation remains in the residuals of predictive models the assumption of independent and identically distributed residuals, fundamental to many statistical techniques, is violated (Dormann et al. 2007). This may bias parameter estimates for the individual models, and possibly relative model performance. In this study, we used semi-variograms to analyze the spatial autocorrelation of all model residuals, and found that is was negligible for all models.

4.6 Identifying important habitat variables

For ecological interpretation and practical conservation use of the fish-habitat models, it is important to identify the most influential habitat variables, i.e. those that have high predictive power in the models. These important aspects of habitat can then become targets for conservation management, or the subject of further ecological studies. However, there is no standardized measure of variable importance across the range of model types investigated here. For linear models, coefficients in the model can give a picture of the importance of each model in relation to its range of variation, but such measures are not available for other models. When using resampling methods as in this study, the frequency with which a given variable is included in the resulting model can be considered a measure of variable importance, which is applicable across the range of models (Brenning 2009). However, the frequency of selection does not indicate the influence the variable has on model predictions. Several complications exist for an unbiased identification of important variables across this range of model types.

In multi-variable models, the selection of an individual variable is influenced by the other variables present in the dataset, particularly those highly correlated to the variable in question (Murray and Conner 2009). For example, three habitat variables in our dataset quantify an aspect of the physical structure of the reef ('depth range', 'coarse rugosity', and 'fine rugosity'). Though these variables quantify slightly different aspects of physical structure, they are highly correlated, and when one of them has been included in a model the others are less likely to be included as well. Variable selection also depends on model structure, as predictor variables that exhibit non-linear relationships with response variables are more likely to be included in GAM than in LM models, and variables which may only become important through interaction effects are only likely to be included in tree-based models. Finally, variable selection depends on the quantification of the variable, particularly in tree-based models where continuous variables are more likely to be selected than discrete or binary variables (Strobl et al. 2007), because of the much larger number of possible splits possible in continuous variables.

Keeping these limitations in mind, variable importance is assessed in this study by permuting individual variables and quantifying the reduction of prediction accuracy caused by the permutations. This method can be standardized for all model types, and provides a direct measure of the influence of the information contained in the original variable values. A limited set of model types was used for the identification of important habitat variables. The LM model was selected because it is the most commonly used, the GAM because it illustrates the effect of non-linearity, and the Bagging model was chosen as the median-performing tree-based model (see chapter 5).

4.7 Determining the characteristic scale of response

Although determining the characteristic scale at which fish variables respond to the habitat was not necessary as input for the development of IKONOS-based predictive models, it provides valuable insight on its own. The scale at which a given habitat variable is most predictive of the fish community is likely to indicate the scale of the ecological processes that influence the community, such as daily migration distances, recruitment patterns etc. (Holland et al. 2004). Identification of these spatial scales is therefore important. However, caution must be taken with interpretation of the predictive ability of habitat variables on the aggregate measures of the fish community used here, as these communities are a composite of different fish species with different life histories and behaviours (Pittman et al. 2007). The characteristic scale was derived by calculating correlations between the fish variables and remotely sensed habitat variables quantified at different spatial scales. The scale at which the remotely sensed habitat variables obtained the highest correlation with the fish variable was then designated the characteristic scale. Because many variables could not be transformed to approximate normality, Spearman rank order correlations were used for the calculations. Spatial autocorrelation of residuals was checked visually by examining semi-variograms. A list of variables used in this investigation is found in Table 4.3.

Table 4.3: List of variables used for investigations of the characteristic scale of response.

***In situ* habitat variables**	**Remotely sensed habitat variables**
Coarse rugosity **Fine rugosity** **Depth range**	**Rugosity (6-300 m radius)**
Live coral cover	**Live coral cover (2-26 m radius)**
Substrate diversity	**Habitat richness (5-200 m radius)** **Habitat diversity (5-200 m radius)**

A similar investigation was carried out using *in situ* habitat variables as response variables. Although the correlations here must be assumed to decline with increasing difference between the spatial scales of the two measurements, the ability of IKONOS data to produce estimates of habitat variables measured *in situ* is important in itself. Not all remotely sensed variables were comparable to an *in situ* measurement (e.g. 'Substrate class'), and not all *in situ* habitat variables had been estimated using the IKONOS data (e.g. 'Number of coral growth forms'). A list of variables used in this comparison is found in Table 4.4.

Table 4.4: List of variables used for comparison of remotely sensed and *in situ* habitat variables.

***In situ* habitat variables**	**Remotely sensed habitat variables**
Coarse rugosity **Fine rugosity** **Depth range**	**Rugosity (6-300 m radius)**
Live coral cover	**Live coral cover (2-26 m radius)**
Substrate diversity	**Habitat richness (5-200 m radius)** **Habitat diversity (5-200 m radius)**

4.8 Summary

The research presented in this thesis is based on the combination of three data sets and a wide range of methods for data processing and analysis, all of which have been described in this chapter. The three data sets include two that derive from fieldwork – the *in situ* fish data and *in situ* habitat data – as well as the remote sensing data which consist of two IKONOS images. Collection of both *in situ* data sets, as well as pre-processing of the IKONOS images, followed tried and tested methods. Fish data were collected at 144 sites using the point count method, and habitat data were collected for the same sites using a range of established methods including visual estimates of structural complexity and random point counts from substrate photos. From each of the raw data sets, a range of variables describing the fish community and the habitat was derived. The IKONOS-based estimation of depth, rugosity, and habitat richness and diversity was also based on established methods, while a linear regression approach, using Lyzenga's depth-invariant index, was developed and tuned for each image to predict live coral cover. The spatial scale of all IKONOS-based estimates of habitat variables was then varied to determine the characteristic scale of response for each relevant variable pair. A range of new machine-learning approaches and traditional statistical models were employed to quantify fish-habitat relationships, and their predictive power and identification of important habitat variables were investigated.

CHAPTER 5: FISH-HABITAT RELATIONSHIPS

In this chapter, we provide descriptive statistics of the fish communities and habitats of the two reefs, a comparison of the predictive models based on *in situ* data. We also derive habitat variables with relatively strong influences on the fish community, as well as investigate the influence variable type and model selection has on the identification of important variables. Together, this constitutes an answer to sub-question A.

5.1 Fish communities and habitats on Chumbe and Bawe

5.1.1 Fish communities

The fish communities on the two reefs show marked differences, as would be expected from their different conservation status and the heavy fishing pressure in the area. Average biomass values for Chumbe are more than double those of Bawe, and major target species of the local fishery (Jiddawi and Ohman 2002) show even greater differences between the two reefs, with some families completely absent from Bawe (Table 5.1). Though these results depend on the location of the specific sites from which data for this study were collected and therefore cannot be taken as true means of the reefs, they agree with previous findings for the same reefs (Lanshammar 2004; Persson and Tryman 2003).

Table 5.1: A comparison of average biomass values (g/100 m^2) of sites on the two reefs, for commonly fished species. The top six families are all commonly fished. Serranidae and Balistidae are families with desirable target species, though their current rarity means that they constitute a small percentage of the local fishery.

Family	Chumbe	Bawe	Ratio
Mullidae	**180.0**	**74.9**	**2.4**
Siganidae	**8.7**	**0**	**∞**
Lutjanidae	**190.2**	**24.2**	**7.9**
Scaridae	**359.3**	**56.4**	**6.4**
Lethrinidae	**184.6**	**83.4**	**2.2**
Acanthuridae	**402.3**	**13.1**	**30.6**
Serranidae	**191.5**	**17.4**	**11.0**
Balistidae	**89.9**	**0**	**∞**

Differences in species richness were smaller, with an average of 17.5 species per site on Chumbe vs. 14.8 on Bawe, and the Shannon diversity measure was found to be practically identical, with 2.53 on Chumbe and 2.55 on Bawe. These results also correspond to findings of previous studies in the area (Tyler 2005).

Table 5.2: Summary of minimum, mean and maximum values for each fish and habitat variable on the two reefs.

Variable name	Minimum (Bawe/Chumbe)	Mean (Bawe/Chumbe)	Maximum (Bawe/Chumbe)
Fish species richness	0.0/0.0	14.8/17.5	30.0/38.0
Fish biomass (g/100m^2)	0.0/0.0	1454/3763	7796/24476
Fish diversity	0.0/0.0	2.55/2.53	3.93/3.80
Branching coral cover (%)	0.0/0.0	1.3/5.8	18.0/86.4
Digitate coral cover (%)	0.0/0.0	12.8/8.4	62.4/64.0
Massive coral cover (%)	0.0/0.0	10.5/9.2	83.2/44.8
Encrusting coral cover (%)	0.0/0.0	1.4/2.4	8.8/24.8
Foliose coral cover (%)	0.0/0.0	0.3/1.2	8.0/52.0
Live coral cover (%)	0.0/0.0	26.4/29.8	83.2/92.0
Dead coral cover (%)	0.0/0.0	24.6/6.5	92.0/48.0
Number of coral growth forms	0.0/0.0	2.5/3.0	5.0/5.0
Turf algae cover (%)	0.0/0.0	0.0/3.6	0.0/16.9
Macroalgae cover (%)	0.0/0.0	0.4/0.8	11.2/20.0
Total algae cover (%)	0.0/0.0	0.4/4.4	11.2/20.0
Sand cover (%)	0.0/0.0	35.6/24.0	96.0/92.0
Seagrass cover (%)	0.0/0.0	0.0/1.8	0.0/68.0
Rubble cover (%)	0.0/0.0	7.2/5.2	44.0/57.6
Pavement cover (%)	0.0/0.0	1.0/26.1	12.0/88.0
Other cover (%)	0.0/0.0	4.8/1.6	83.2/18.4
Coarse rugosity	0.0/0.0	1.4/1.7	4.0/5.0
Fine rugosity	0.0/0.0	1.8/1.6	5.0/5.0
Average depth (m)	0.6/1.4	3.0/3.6	8.9/10.1
Depth range (m)	0.0/0.1	1.2/1.6	5.8/7.2
Substrate diversity	0.17/0.28	0.98/1.16	1.50/1.84
Substrate evenness	0.24/0.22	0.65/0.68	0.99/1.00

5.1.2 Habitats

A comparison of habitat variables also illustrates differences between the two reefs. The cover of live coral has marginally higher average values on Chumbe (29.8%) than on Bawe (26.4%); the opposite is true for dead coral cover. In addition, Chumbe has substantially higher average values of both pavement and seagrass, the latter of which is completely absent from Bawe. Minimum, mean and maximum values for all variables are tabulated in Table 5.2.

5.2 Prediction of the fish community from habitat variables

Results from the 100 runs of each model, predicting values of each of the dependent fish variables from *in situ* habitat data, are presented in Figure 5.1. Results will be discussed primarily with reference to those obtained using the cross-validation resampling method, as this tends to produce more honest accuracy estimates with relatively small datasets (as this one, n=144) (Efron and Gong 1983), though results from both resampling methods are reported.

5.2.1 Model accuracies

All models perform better on average than a simple predictor that predicts the average value of the variables for all sites (Table 5.3). However, there is a substantial difference between the performances of the individual models. The model accuracies, expressed as the average RMSE (Figure 5.1), differ between resampling methods in that bootstrap resampling results in a lower average and larger range of RMSE values. However, the relative accuracy of each model type is generally consistent for the two resampling methods, with only the predictive models for biomass, resampled using the cross-validation method, differing markedly.

Table 5.3: Average RMSE values for an 'average predictor' model.

	Bootstrap	**Cross-validation**
Fish species richness	**7.80**	**7.72**
(log) Fish biomass	**0.670**	**0.628**
Fish diversity	**0.811**	**0.792**

5.2.1.1 Description of model differences

For all dependent variables and both resampling methods, the GAM model performs better than the LM, which is outperformed by all the other models except for one case (SVM predictions of biomass using cross-validation). The poor performance of the LM, as opposed to the GAM, is an indication that non-linearity is common in the modelled fish-habitat relationships. This non-linearity is best illustrated by the individual relationships (Knudby et al. 2008), three examples of which are presented in Figure 5.2.

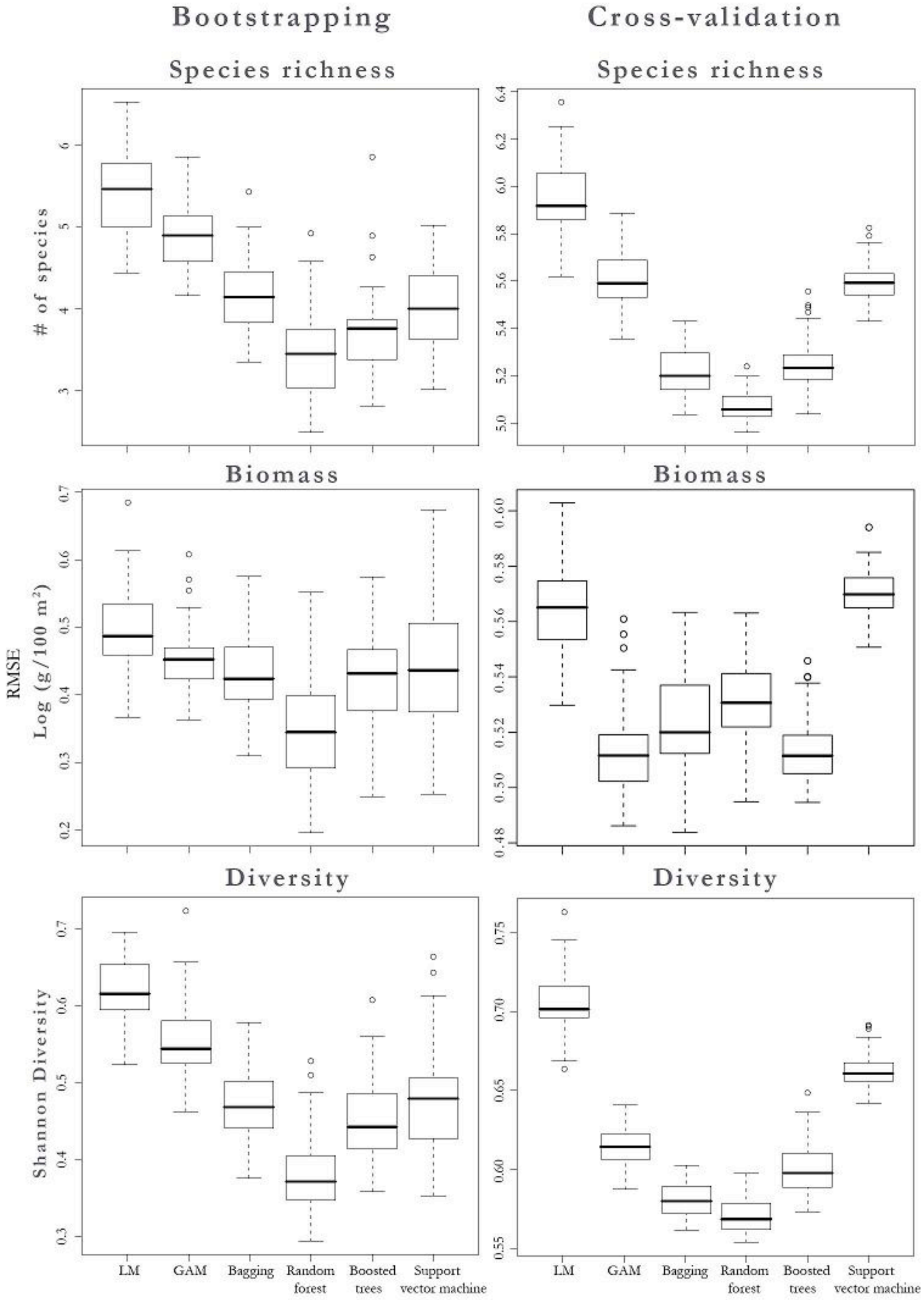

Figure 5.1: Results from 100 runs of the six models for each dependent variable. Results are shown using both cross-validation and bootstrapping resampling methods.

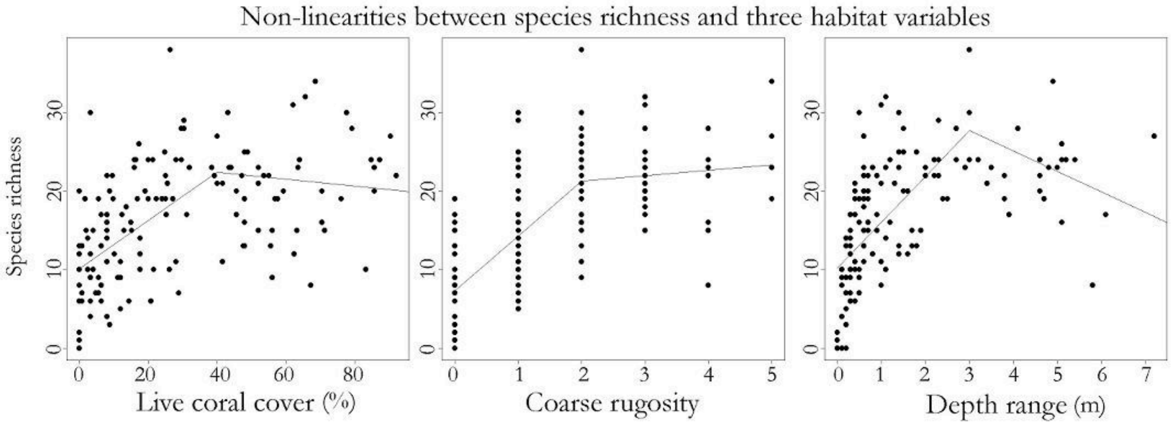

Figure 5.2: Examples of non-linear relationships between fish species richness and three habitat variables. Similar relationships exist for the other two fish variables. Lines are inserted to aid interpretation.

Except for the models predicting biomass using cross-validation resampling, all three tree-based models outperform both the LM and the GAM in terms of prediction accuracy. Tree-based models are known to be superior in dealing with interaction effects (Breiman et al. 1984; De'ath and Fabricius 2000), which are likely to occur in nearshore environments (Pittman et al. 2007). Two examples of interaction effects are illustrated here on partial plots, using the data from this study. At shallow depths (<4 m) there is no significant influence of conservation status on fish species richness, but at greater depths (>4 m) the effect is obvious and statistically significant (see Figure 5.3). This is probably partly due to the fact that most fishing (eliminated in protected areas) is conducted in the zone that naturally contains the greatest species richness – just off the reef edge where depths on the two reefs sampled here range from 4 to 10 m. A change with depth is also seen for the relationships between the variable 'coarse rugosity' and fish species richness (see Figure 5.4). At shallow depths (<4 m) there is a significant positive correlation between the two variables, but at greater depths (>4 m) this correlation weakens, with even a suggestion of negative correlation at the greatest depths (>6 m).

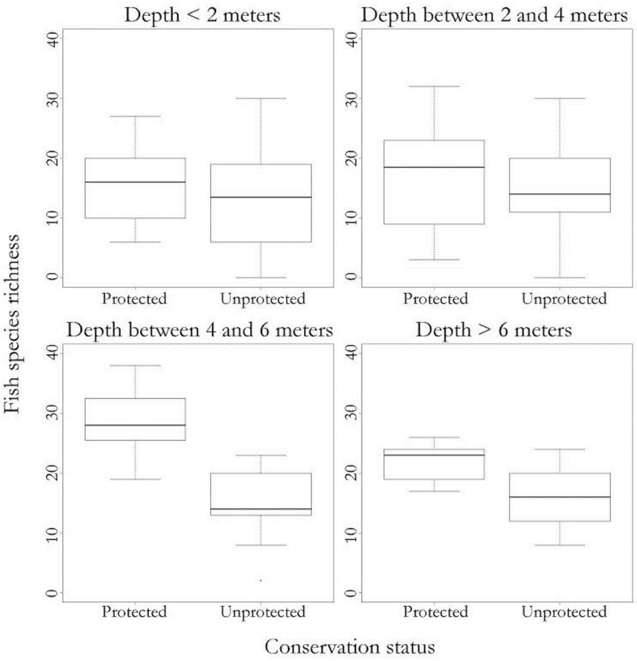

Figure 5.3: First example of an interaction effect. At shallow depths (<4 m) conservation status has no significant influence on fish species richness, whereas the effect is significant at greater depths (>4 m).

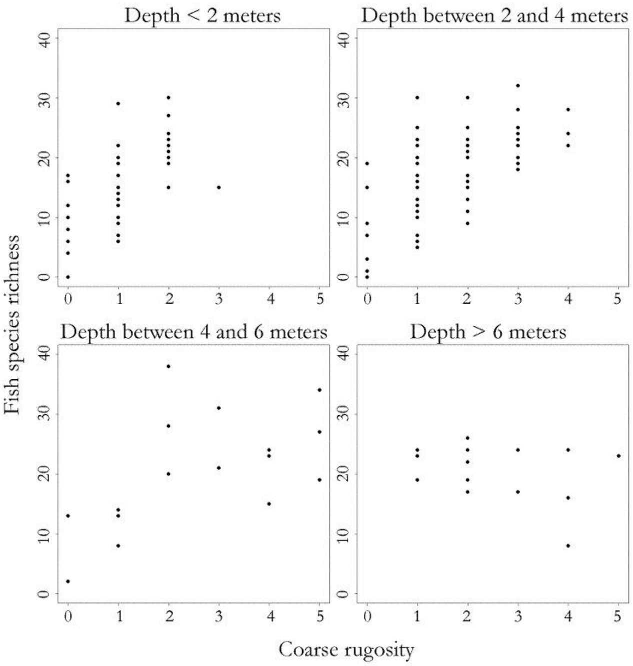

Figure 5.4: Second example of an interaction effect. The variable 'coarse rugosity' has significant positive influence on fish species richness at shallow depths (<4 m), but not at greater depths (>4 m).

Of the tree-based models, the Random forest model is superior to the others, except for predictions of biomass using cross-validation. Random forests have previously been shown competitive with both Bagging and boosting (Breiman 2001a), which is confirmed here. This is an important result because the use of Boosted trees recently has received attention in ecology (De'ath 2007; Elith et al. 2008), including for prediction of marine fish community variables (Pittman et al. 2007; Pittman et al. 2009). The SVM performed equally well or worse than the tree-based models for every dependent variable, and was consistently outperformed by the Random forest.

However, the GAM, together with the Boosted trees, outperforms the other tree-based models when predicting fish biomass. It is not clear why the GAM model performs better with fish biomass as response variable, or why the Boosted trees outperform the Random forest for this variable. Biomass, as opposed to the two other dependent variables, can be heavily influenced by a few large individuals at the site, or a large school entering the site during the period of observation. It is possible that the cubic smoothing splines applied in the GAM model are particularly suited to dealing with this issue, as opposed to the tree-based models' binary splits, which only employ "greater than" and "less than" decision structures. However, the superiority of the GAM model for this particular dependent variable and resampling method warrants more detailed examination if confirmed with other datasets. It is also surprising that the choice of resampling method can influence relative model performance so strongly.

5.2.1.2 Tests of difference between model prediction accuracies

Results from the t-tests show that all except one model pair produce significantly different prediction accuracies ($p <$ Simes-corrected α). Only the GAM and Boosted trees models for fish biomass were not significantly different.

5.2.2 Model precision

In addition to high accuracy, precision (stability) is a desirable characteristic for any predictive model. For prediction of fish species richness, the Random forest model achieved the lowest standard deviation of estimates (0.067), and can thus be considered the best predictive model both in terms of accuracy and precision. The SVM also achieved high precision (standard deviation of estimates=0.085). For prediction of biomass, the SVM achieved the lowest standard deviation of estimates (0.008), followed closely by the Boosted trees (0.011), but differences were small between all models with Bagging showing the highest standard deviation at 0.016. For fish diversity predictions, three models achieved very similar precision, with the SVM, Bagging and the Random forest all having standard deviations of 0.011. For all of the dependent variables, the most accurate model is also one of the most precise. Although outliers exist (see Figure 5.1) none are extreme, and model selection can therefore be based on accuracy without compromising on precision.

5.2.3 Finding the important habitat variables

As a basis for interpretation only, correlations of individual habitat and fish variables (excluding 'Conservation status' and 'Reef') are presented in Table 5.4. A summary of variable importance, derived from the permutation of individual variables in models predicting each fish variable, is shown in Figure 5.5.

Table 5.4: Correlation coefficients obtained between individual habitat and fish variables.

Variable	Fish species richness	Fish biomass	Fish diversity
Branching coral cover	0.18	0.01	0.21
Digitate coral cover	0.22	0.24	0.25
Massive coral cover	0.34	0.22	0.21
Encrusting coral cover	0.24	0.19	0.13
Foliose coral cover	0.04	0.05	0.10
# of coral growth forms	0.57	0.40	0.44
Turf algae	0.21	0.15	0.16
Macroalgae	-0.07	-0.03	-0.17
Depth	0.30	0.34	0.15
Depth range	0.52	0.53	0.36
Coarse rugosity	0.59	0.55	0.49
Fine rugosity	0.34	0.26	0.37
Substrate diversity	0.28	0.12	0.26
Substrate evenness	0.08	-0.02	0.07
Live coral cover	0.50	0.39	0.43
Dead coral cover	0.14	-0.11	0.30
Algae cover	0.14	0.11	0.04
Sand cover	-0.50	-0.34	-0.41
Rubble cover	-0.18	-0.08	-0.15
Seagrass cover	-0.16	-0.18	-0.17
Pavement cover	0.01	0.12	-0.12
Other cover	-0.02	-0.03	-0.02

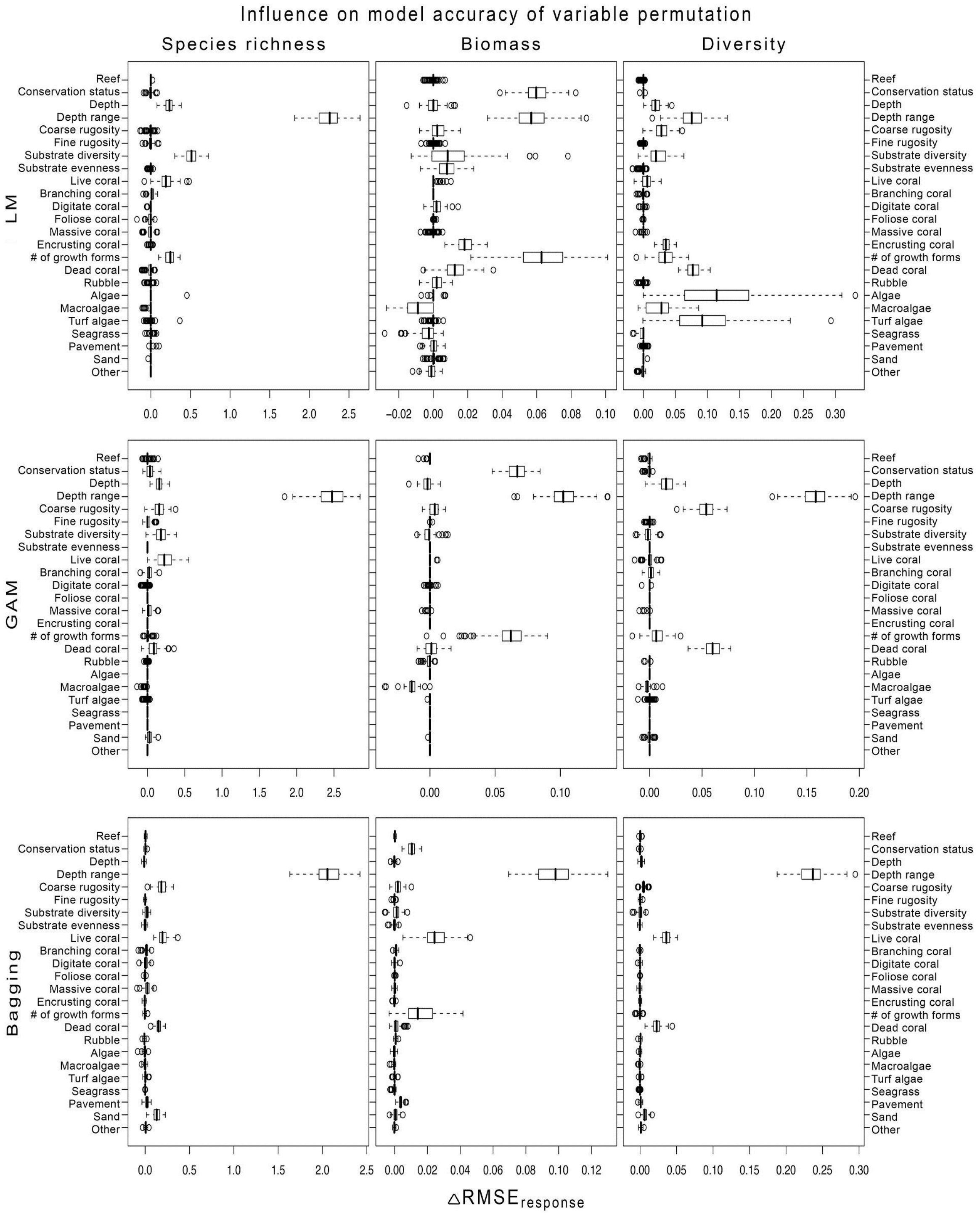

Figure 5.5: Summary of variable importance for all combinations of response variable, for LM, GAM and Bagging models. The boxes and whiskers in this and all similar figures are drawn to according to Tukey’s definitions. The centre-line marks the median value, the two ends of a box mark the first and third quartiles, whiskers mark end of 3/2 of the interquartile range, circles mark outliers (outside whiskers).

The ‘depth range’ variable is important for all combinations of response variable and model type, and indeed has the highest importance in 7 of the 9 combinations. The importance of other variables, however, differs substantially between model types and response variables. For all models of fish species richness, the ‘depth range’ variable is complemented by a series of variables with minor importance, including ‘depth’, ‘substrate diversity’, ‘live coral cover’, ‘number of coral growth forms’, ‘coarse rugosity’, ‘dead coral’, and ‘sand’.

For the fish biomass response variable, substantial difference is seen between the model types. In the LM model, ‘conservation status’ is important along with the ‘number of coral growth forms’, and a few other variables of minor importance. The importance of ‘conservation status’ for fish biomass is best explained by the local MPA effectively protecting large-bodied fishes, which are rarely seen outside the MPA boundary. The ‘number of coral growth forms’ is more likely to be a proxy-variable describing proximity to the reef edge, where greater variety of coral growth forms is seen along with greater fish biomass (Knudby, pers. obs.). The same variables are important in the GAM model, whereas ‘conservation status’ has lost importance in the Bagging model and been replaced by ‘live coral cover’. These two variables are themselves closely related, as live coral cover is much higher within the MPA than outside it.

Models of fish diversity show greater differences in variable importance. A number of unsuspected variables are important in the LM model, such as ‘algae’, ‘turf algae’, and ‘dead coral’, along with ‘depth range’ and a series of variables with minor importance. The two variables related to algae cover have most likely become important in the LM because their distributions are far from normal, and the permutations are therefore more likely to cause a dramatic effect on predictions, particularly with the RMSE used as measure of model performance. The algae variables lose importance in the GAM model, probably because its transformation of these skewed variables limits the impact of permutations, and in the Bagging model which is less sensitive to extreme values. In the GAM model, ‘depth range’ is again the most important variable, followed by ‘dead coral’ and ‘coarse rugosity’. The ‘depth range’ variable is even more dominant in the Bagging model, along with ‘live coral cover’ and ‘dead coral’ which both have minor importance.

5.3 Discussion

5.3.1 Model types

The comparison of model accuracies demonstrates the improvement in predictive performance that can be achieved with models that are able to incorporate non-linear relationships and interaction effects. The only difference between the LM and GAM models is the ability of GAM models to incorporate non-linear transformations of the input data, and this lead GAM to consistently outperform LM in this study. This is not surprising given known linearities in individual fish-habitat relationships (Knudby et al. 2008), but points to the importance of accepting and incorporating non-linearities when modeling fish-habitat relationships.

Similarly, the superiority of the tree-based models for predicting fish species richness and diversity suggests that incorporation of interaction effects is important and leads to higher prediction accuracy. The other feature that distinguishes the tree-based models is their reliance on binary splits rather than continuous functions. The models' relative performance suggests that this structure may be beneficial for modeling fish species richness and diversity, but not fish biomass. Compared to the other response variables, fish biomass has more extreme values, caused by the presence of a few large individuals or a school of medium-sized fishes passing through a site during data collection. The "larger than or smaller than" decision rules used in tree-based models are unable to successfully predict these extreme values, and the ability of the GAM model's smoothing splines to do so may be the reason for its superior performance in the prediction of the fish biomass variable. The reason for this difference may ultimately lie in the better ability of the less flexible GAM to extrapolate relationships learned on cross-validation samples with some spectrum bias towards smaller fish. Spectrum bias between training and test samples is more likely to occur in partitioning approaches such as cross-validation than in bootstrapping, i.e. simulating independent random sampling. This might help explain the differences in error estimation results between cross-validation and the bootstrap in the case of fish biomass.

To my knowledge, this study is the first to compare a range of predictive modelling techniques for coral reef fish community variables, but similar comparative studies have been conducted in forest environments. Moisen and Frescino (2002), studying a variety of discrete and continuous response variables describing forest state, found that GAM and Multivariate Adaptive Regression Splines (MARS) both outperformed regression trees in all performance measures, however, no ensemble technique was used for the regression trees. Another study by Moisen et al. (2006) found that boosted trees outperformed both GAM and individual regression trees for predictions of tree basal area. Prasad et el. (2006) evaluated single regression trees against bagging and random forests as well as mars, and found that both bagging and random forests outperform the other models for predicting basal area, with random forests outperforming bagging slightly. These studies reinforce the necessity of applying ensemble techniques to regression trees, but do not provide a strong foundation for general conclusions about the relative performance of model types. Several comparison studies also exist for species distribution modelling (e.g. Elith et al. 2006; Guisan et al. 2007). Although these are not directly comparable because of their binary response variables and different performance measures, they lend some support to the strong performance of tree-based ensemble techniques.

5.3.2 Important variables and model types

To my knowledge, this study is the first to use a permutation-based approach to assess variable importance across a range of multi-variable model types. This approach was inspired by the need to derive a measure of variable importance applicable to all the structurally different model types used in this study, some of which are gaining increasing popularity in the ecological modelling community. A similar approach is implemented in R's 'randomForest' package, and has successfully been applied to support vector machines (Taylor 2009). More specific variable importance measures exist for individual model types. For example, variable importance can be derived for linear models by comparing the coefficients of determination of individual predictor-response variable pairs, or through a range of methods designed for linear multi-variable models (Graham 2003; Murray and Conner 2009), however, these methods are not applicable to GAMs. For tree-based models, variable importance can also be measured as the reduction in prediction error achieved by the split at each node (Friedman 2001). This measure can be averaged for ensemble techniques, as implemented in R's 'gbm' package, but is only applicable to tree-based models. We propose a wider adoption of a permutation-based approach to assessing variable importance due to its transparency and applicability across model types.

Keeping in mind the limitations involved in determining variable importance, the comparison between response variables and model types reveals a difference in the number of variables each model identifies as important. The LM model identifies the greatest number of important variables. This is most likely due to its sensitivity to non-normally distributed variables, whose extreme values cause large areas when the variables are permuted and are therefore identified as important. The importance of 'algae' and 'turf algae' in determining fish diversity is neither supported by other studies, nor by the GAM and Bagging models, and must be considered artifacts arising from the combination of methods used to determine variable importance and the specific frequency distributions of the variables in the dataset used. The Bagging model identifies fewer important variables than the GAM, and seems to have a bias towards continuous (as opposed to binary or discrete) variables, as has been shown for Random forests (Strobl et al. 2007). The 'live coral cover' variable has probably been preferred over 'conservation status' in the Bagging model of fish biomass because it is continuous, as opposed to the binary 'conservation status' variable. This allows a large number of possible splits in the regression trees, as opposed to only one possible split for a binary variable. A similar situation is seen for the fish diversity models, where the importance of the 'depth range' and 'live coral cover' variables (both continuous) are increased in the Bagging model as opposed to the GAM, while the importance of the 'coarse rugosity' and 'number of coral growth forms' variables (both discrete) are decreased, along with the 'depth' variable. Identified variable importance is thus dependent on the set of variables included in the dataset, their frequency distribution, their scale of measurement and number of categories, and on the model type. These dependencies are rarely mentioned in the literature, possibly because a standardized tool for comparison, such as the permutation-based approach used here, has not previously been available. Given the potential use of "important variables" as conservation targets or objects of further scientific inquiry, the influence on model type on variable importance needs to be studied further.

The range of model types available and their relative predictive performance, as well as the intricacies of determining variable importance, are important for the practical use of predictive models. In the oceans, where large data sets are costly to obtain and the distributional patterns of species and ecological relationships between organisms and their environment must often be inferred from the limited available data, predictive models are crucial as input to conservation management (Leathwick et al. 2006). Our results, based on coral reefs with their high biodiversity and numerous interactions, provide strong support for the use of tree-based ensemble techniques when developing predictive models in such environments. Given the relative ease of developing these models in freely available software, practitioners are no longer forced to rely on simplistic linear models for modelling of their highly complex environments. The one problem caused by the use of ensemble techniques is a loss of interpretability, otherwise a strength of regression trees. This is only partly compensated by the variable importance measure, and it is suggested that individual regression trees be created and visualized for better interpretation of model structure and interaction effects.

5.3.2.1 The influence of conservation status

The positive effect of effective protection (e.g. through MPAs) on fish community variables has been shown in numerous studies (Halpern 2003). These studies use a variety of sampling designs, habitat and fish variables, and analytical models, and comparison between individual studies is therefore difficult. Previous studies conducted on the reefs around Chumbe Island and nearby areas show a positive effect of Chumbe Island Coral Park on both fish biomass and species richness (Lanshammar 2004; Tyler 2005), though substantial habitat differences between protected and unprotected areas make firm conclusions difficult to draw. The influence of protection on fish species richness is not supported by the results presented in this study, as the 'Conservation status' variable was not identified as important by any of the three model types. The influence on fish biomass, however, is supported, although the magnitude of the 'Conservation status' variable's influence on biomass is strongly dependent on model type (see Figure 5.5). The difference between the importance of this variable in the three models of fish biomass illustrates the complexity of assessing the importance of a single variable in an environment as complex as a coral reef, but it is worth noting that the importance of this variable is substantial in the most accurate model (GAM). Furthermore, the importance of 'Live coral cover' in the Bagging model can be considered an indirect support for the importance of protection, since the fishing methods used in the area influence live coral cover negatively, and the high coral cover inside the protected area is thus likely to partly be a result of protection from fishing.

5.3.2.2 Comparison with findings from other studies

Keeping in mind the limitations of the methods used to find important/influential habitat variables as well as the range of variables and methods used in other studies, comparison between the habitat variables identified as important in different studies is not straight-forward. Results from a number of studies point to three variables that have repeatedly been found to influence reef fish communities – depth, structural complexity, and live coral cover.

In this study, 'depth' was found to have a minor influence in LM and GAM models of fish species richness and diversity, but not biomass. This supports its importance for these two response variables (Friedlander and Parrish 1998; Huston 1994), but its exclusion from all Bagging models (which produce superior prediction accuracies for the two response variables in question) suggests that the inclusion of 'depth' in LM and GAM models may be a result of the relative simplicity of these model types, rather than a result of the importance of the 'depth' *per se*. Paradoxically, 'depth' is known to control the relationship between other variables (see Figure 5.3 and Figure 5.4), which should increase its importance in the Bagging model type. The geomorphologic structure of the reefs around Chumbe and Bawe islands may also explain the limited importance of depth in this study. Both reefs consist of a relatively extensive shallow reef flat area, a narrow reef crest, and a very limited fore reef. Most depth variation therefore exists between the sites on/near the reef crest and those on the reef flat, and several other variables, including those quantifying structural complexity and coral cover, are likely to also discriminate between these two parts of the reef.

Structural complexity, quantified most effectively in this study as 'depth range', was an important variable in all models for all response variables, which supports previous findings (Friedlander and Parrish 1998; Luckhurst and Luckhurst 1978; Risk 1972). Although not measured using the exact same methods and over the same spatial scales, the results presented here also support those of McCormick (1994), who found that substratum height difference, compared to other measures of structural complexity, correlated strongly with fish community variables. However, the continuous nature of the variable, as opposed to the two discrete variables used as alternative measured of structural complexity ('fine rugosity' and 'coarse rugosity') may also be the reason for the importance of the 'depth range' variable over the other two.

Live coral cover was identified as important, though weakly, in all models of species richness, and Bagging models of all response variables. The model-independent influence on species richness provides strong support for the influence of this variable, and is probably best explained by the live coral both providing food for corallivorous fishes (Garpe and Ohman 2003) and creating a fine-scale structure on the reef that provides shelter space for small fishes and juveniles of larger species (Lindahl et al. 2001). Its inclusion in Bagging models of both fish biomass and diversity, but not in the LM and GAM models for the same response variables, suggests that it is correlated with other variables and that the choice of variable to be included depends on model type, possibly on the basis of variable type (continuous, discrete or binary). Likely competing habitat variables are 'Conservation status' for the biomass models, and 'coarse rugosity' or 'number of coral growth forms' for the diversity models. These variables are all discrete or binary, and are all identified in the respective LM and GAM models, but not (or only with very limited importance) in the Bagging models.

In addition to these three variables, the 'number of growth forms' and the 'dead coral cover' variables were identified as important in some models. The 'number of coral growth forms' has previously been identified as important for the abundance of damselfishes (Ormond et al. 1996), whereas in the models presented here it was identified as influencing biomass. The causality behind this relationships is speculative, but it may be caused by the 'number of coral growth forms' functioning as a proxy for the distance to the reef edge, a zone where both the 'number of coral growth forms' and the 'fish biomass' variable have high values.

The variable 'dead coral cover' has been found to influence other aspects of the fish community such as the abundance of wrasses (Garpe and Ohman 2003), whereas it was identified as important for fish diversity in the results presented here. The causality behind this relationship is also speculative, and warrants further investigation.

Other variables were also identified as important in LM models (the 'algae', 'macroalgae' and 'turf algae' variables, as well as 'encrusting coral cover' and 'substrate diversity'). Although these variables may have real influence on the fish community, the fact that they are only identified as important in the LM models suggests that their importance is an artifact caused by their non-normal distributions and the LM model structure.

5.4 Summary

Sub-question A was "*What is the statistical nature of fish-habitat relationships?*". The results presented in this chapter show that the relationships are complex, non-linear and involves interactions between several habitat variables. However, they also illustrate that with data on the right habitat variables it is possible to develop predictive models of the fish community that significantly outperform an average-predictor. The complexity of the fish-habitat relationships leads tree-based ensemble technique to outperform the other model types tested here.

The chapter also presented an assessment of the influence each habitat variable has in predicting the fish community variables. Care must be taken when assessing the importance of individual habitat variables because of collinearity between them, particularly if using models that assume linear relationships or no interaction effects. Nevertheless, our results point to the importance of several variables also identified in other studies. In our bagging model, structural complexity quantified as 'depth range' is the main habitat influence on the fish community, although minor influences are seen from conservation status, coarse rugosity, live coral cover, dead coral cover, and the number of coral growth forms. 'Conservation status', of particular interest because MPAs are the primary management tool limiting human impacts on reef fish communities, has minor importance on fish biomass, and virtually no influence on species richness or diversity. 'Reef', the variable describing the location of a field site, has virtually no influence. This is encouraging because it suggests that extrapolation of predictions to reefs not sampled during fieldwork, although untested, could be feasible.

CHAPTER 6: REMOTE SENSING OF HABITAT VARIABLES

In this chapter we examine the accuracy with which some habitat variables measured *in situ* can be estimated with the use of IKONOS imagery. The results provide a first indication of the utility of remote sensing for spatial predictions of the fish community. Secondly, in order to investigate the optimum spatial scale at which these variables can be remotely sensed, the spatial scale of remote observations is varied and the influence of this variation on estimates of both *in situ* habitat variables and fish community variables is investigated. This provides an answer to sub-question B.

Only a few of the habitat variables measured *in situ* are investigated in this chapter, as several of them are not feasible to estimate with IKONOS data, or with any other currently available remote sensing data. Estimations may be obtained for some of these "difficult" variables (e.g. the number of coral growth forms), but only indirectly through correlations with other variables that lend themselves more to remote estimation (e.g. live coral cover). This chapter will focus on those habitat variables that have the potential to be estimated directly using IKONOS imagery. Although the classification of substrate types around the two islands was not in itself an important result of the study, it is also presented in this chapter because it forms the basis for both the mapping of live coral cover and habitat variety, both variables discussed later in the chapter. The substrate maps of Bawe and Chumbe that are the results of the classifications are shown in Figure 6.1 and Figure 6.2, respectively.

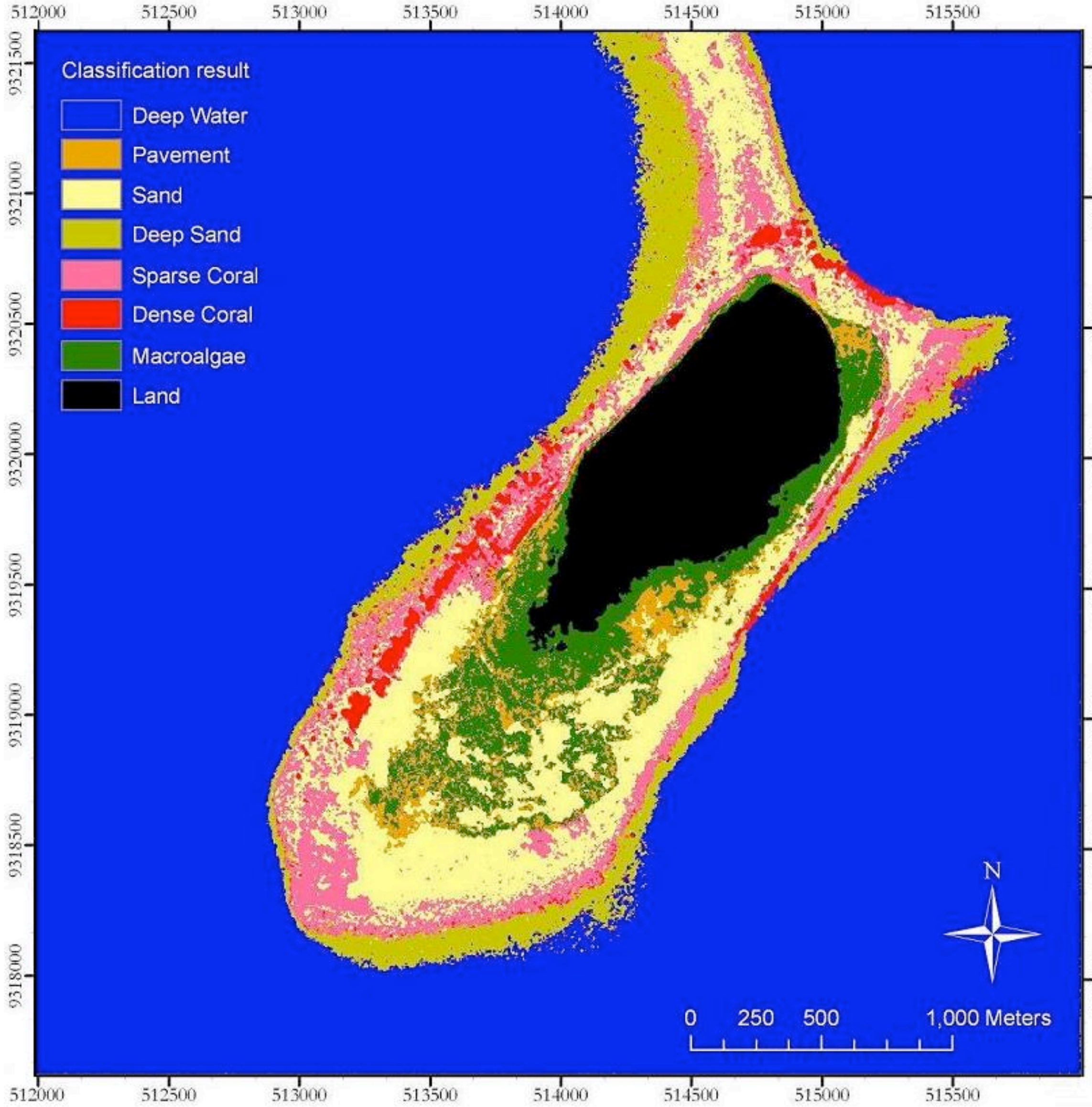

Figure 6.1: Result of Maximum Likelihood Classification, Bawe Island. Note scale difference between Figure 6.1 and Figure 6.2.

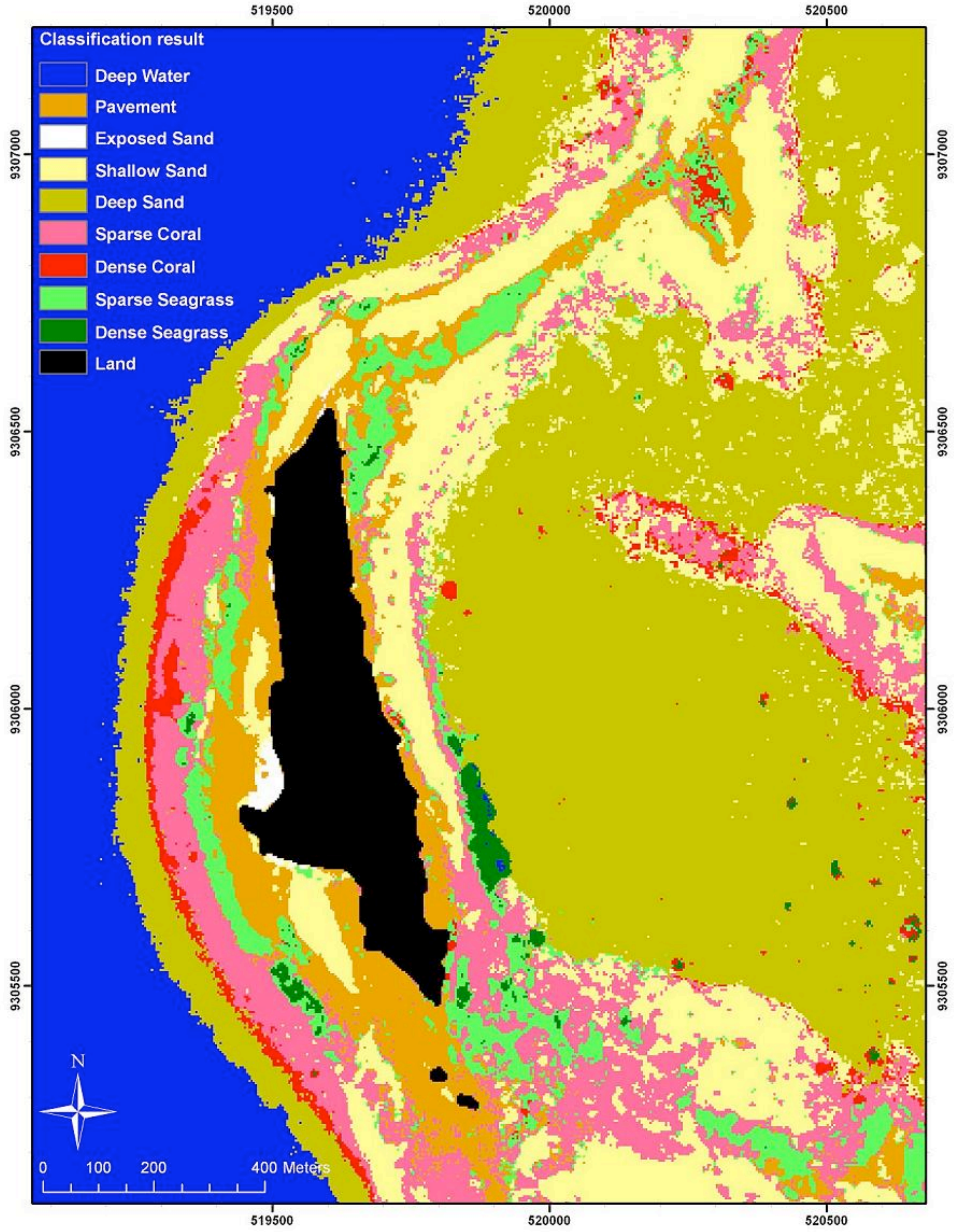

Figure 6.2: Result of Maximum Likelihood Classification, Chumbe Island. Note scale difference between Figure 6.1 and Figure 6.2.

6.1 Remote sensing of live coral cover

As outlined in chapter 4, a linear model was developed for each IKONOS image, and hence for each reef, predicting live coral cover values from the depth-invariant index. A scatterplot showing the relationship between these two variables for each reef is shown in Figure 6.3. It is clear from the figure that the difference in image attributes such as viewing geometry and water turbidity at the time of image capture causes a different specific relationship between the two variables. The two linear models were then used to predict live coral cover for each pixel classified as either 'sparse coral' or 'dense coral' on each reef, as shown in Figure 6.4 and Figure 6.5. Pixels not classified in either of those two categories were assumed to have no live coral cover. The use of linear regression with a response variable in percentage units may introduce bias, especially given the relative prevalence of values near 0%. However, we feel that the good fit, reasonable spread of live coral cover values in the field data, and the fact that the regression is only applied in areas classified as coral cover, justify the use of linear rather than logistic regression.

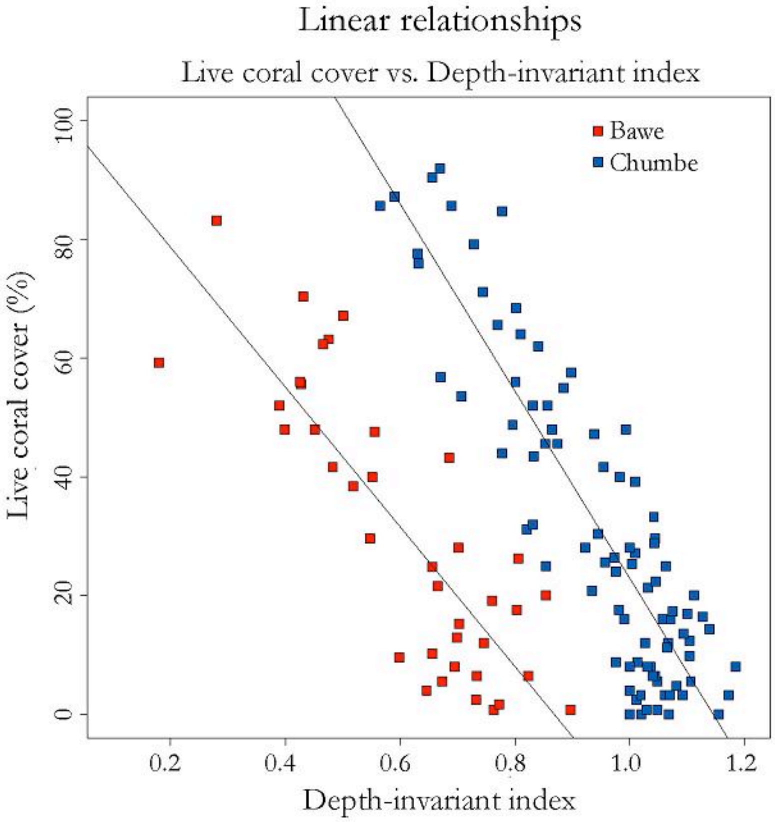

Figure 6.3: Linear relationships between the depth-invariant index and live coral cover on the two reefs.

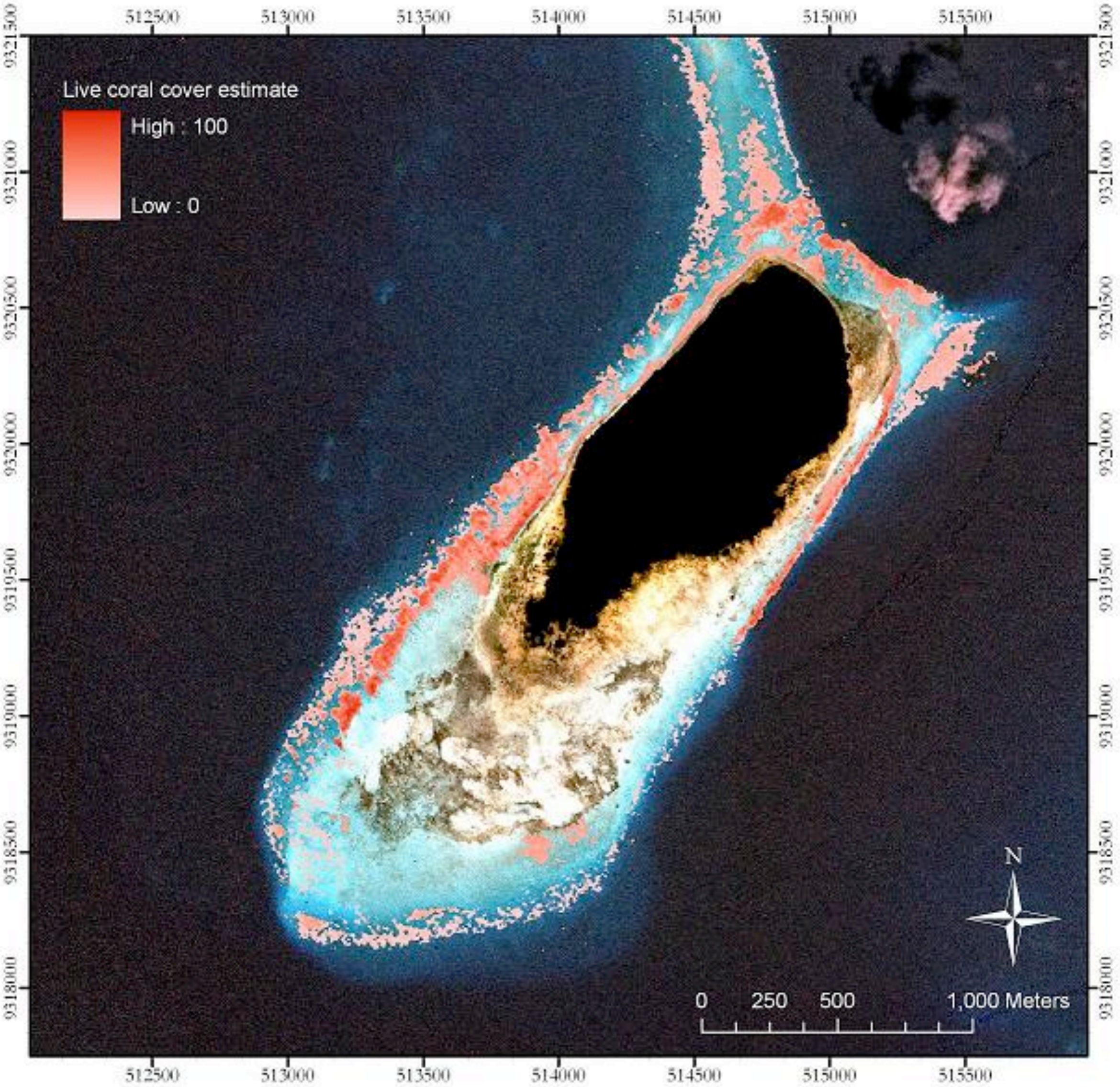

Figure 6.4: Spatially distributed prediction of live coral cover for Bawe Island. Coral areas are shown in red, overlaid on a true-color composite of the original data.

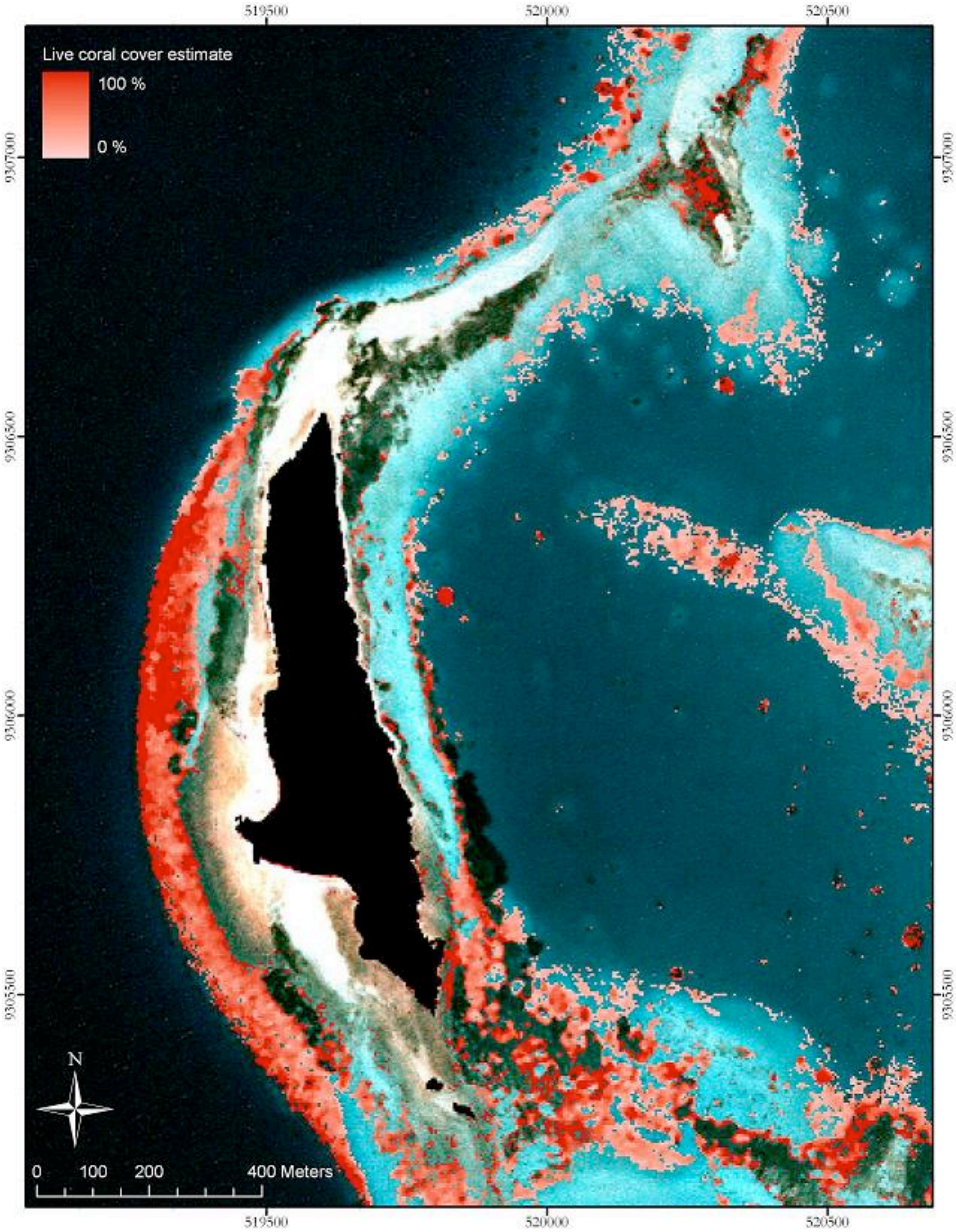

Figure 6.5: Spatially distributed prediction of live coral cover for Chumbe Island. Coral areas are shown in red, overlaid on a true-color composite of the original data.

6.1.1 Accuracy of IKONOS-based live coral cover estimates

In order to assess the accuracy of the live coral cover estimates, two separate issues must be taken into account. First the accuracy of the classification must be considered with respect to the coral areas, and secondly the quantitative estimate of live coral cover within the areas classified as coral must be estimated.

6.1.1.1 Assessing the classification accuracy

The confusion matrices for both classifications are presented in Table 6.1 and Table 6.2. As a function of limited time and equipment for fieldwork, the number of samples per class available for accuracy assessment is substantially lower than the 50 suggested as a rule of thumb by Congalton (1991). This questions the confidence one can have in the assessment. However, a reverse use of the two datasets (using the test data for classification and the classification data for accuracy assessment) provided similar results, in support of the accuracy assessment reported here. The relatively small study area also means that the number of data points collected provided a reasonable coverage of the area (see Figure 4.5 and Figure 4.6).

In order to assess the coral areas together, the two coral classes are combined into one, and the user and producer accuracies are calculated for the combined classes. On Bawe, the producer accuracy, i.e. the fraction of actual coral pixels classified in one of the two coral classes, was 87.0%. The user accuracy, i.e. the fraction of pixels classified in one of the two coral classes that actually contain coral, was 80.0%. On Chumbe, the values were 92.6% and 62.5% respectively, indicating an overestimation of the total coral area. The confusion matrix for Chumbe (Table 6.2) reveals that the low user accuracy for Chumbe is due to substantial confusion between sparse coral and sparse seagrass. Investigation of the 10 points that actual contain sparse seagrass but are classified as coral (italicized in Table 6.2) reveals that these points are all located in the area immediately southeast of Chumbe Island (see Figure 6.5), where the patchy nature of the substrates made contextual editing unfeasible. It is therefore likely that a substantial part of the area classified as coral in that part of the image is actually covered by seagrass. Given the dominance of chlorophyll *a* absorption from both corals and seagrasses, the relatively low cover values that both classes represent, and the range of the classes (5-40% coral cover and 0-250 g/m^2 aboveground seagrass biomass, respectively), the difficulty in discriminating between these two classes is not surprising. However, around Bawe, and in most areas around Chumbe, the areas dominated by coral have been mapped with an accuracy that compares favourably to other studies (Andréfouët et al. 2003; Mumby et al. 2004c).

Table 6.1: Confusion matrix for classification of shallow-water habitats around Bawe Island. Field data in columns, classification results in rows.

	Pave-ment	Shallow Sand	Deep Sand	Sparse Coral	Dense Coral	Macro-algae	Deep Water	Total
Pavement	7	0	0	1	0	4	0	12
Shallow Sand	0	20	0	2	0	0	0	22
Deep Sand	0	0	39	0	0	0	0	39
Sparse Coral	1	2	0	6	2	0	0	11
Dense Coral	0	0	0	2	10	0	2	14
Macroalgae	1	0	0	0	0	30	0	31
Deep Water	0	0	0	0	0	0	20	20
Total	9	22	39	11	12	34	22	149

Table 6.2: Confusion matrix for classification of shallow-water habitats around Chumbe Island. Field data in columns, classification results in rows.

	Pave-ment	Exposed Sand	Deep Sand	Sparse Coral	Dense Seagras	Sparse Seagras	Shallow Sand	Dense Coral	Deep Water	Total
Pavement	34	0	0	0	0	12	0	0	0	46
Exposed Sand	0	7	0	0	0	0	0	0	0	7
Deep Sand	0	0	22	0	0	0	0	1	0	23
Sparse Coral	1	0	0	11	2	10	0	4	0	28
Dense	0	0	0	0	9	2	0	0	0	11
Sparse	2	0	0	0	5	23	0	0	0	30
Shallow Sand	0	0	0	0	0	2	24	0	0	26
Dense Coral	0	0	0	1	0	1	0	9	0	11
Deep Water	0	0	0	0	0	0	0	0	28	28
Total	37	7	22	12	16	50	25	14	28	210

6.1.1.2 Assessing the accuracy of live coral cover estimations

Within the areas classified as coral, live coral cover was estimated with an RMSE of 21.65 percentage points on Bawe (Pearson R=0.49), and an RMSE of 18.32 percentage points on Chumbe (Pearson R=0.76). Though the results from Bawe show a significant correlation between real and predicted live coral cover (p=0.003), the predictive power of the model is only slightly better than that of an "average-predictor" (RMSE of 23.50 percentage points). This poor performance is mainly attributable to the inclusion of two data points from the area of dense coral immediately southwest of Bawe Island. At the time of fieldwork, this area was dominated by rubble and recently settled soft corals, suggesting that a recent disturbance had killed what hard coral was there, and re-colonisation of the substrate was ongoing. Although this area therefore was likely to have contained relatively dense coral cover at the time of image acquisition (2 years before fieldwork), the high values of coral cover predicted by the depth-invariant index were not present in the field data. Excluding these two points from the analysis would yield an RMSE of 16.77 percentage points for Bawe (Pearson R=0.69). The results from Chumbe compare more favourably with an "average-predictor" (RMSE of 29.96 percentage points). The results also compare reasonably well to those obtained by Joyce (2004b) (Pearson R=-0.76), especially considering that she used airborne data with spectral bands tuned to optimize water penetration and differentiation between typical coral reef substrates.

The most likely explanation for the ability of IKONOS data to predict live coral cover around Bawe and Chumbe islands, despite the acknowledged difficulties associated with multispectral satellite data for this purpose, is the low cover of spectrally similar substrates such as algae and seagrasses within the areas classified as coral. In addition, turbidity is low in both images, and the analysis is limited to areas with a maximum depth of 10.1 m due to the fact that coral development in the area is very limited below this depth.

6.1.2 Live coral cover estimates and their spatial variation

The influence of the spatial scale of IKONOS-based estimations of live coral cover is shown in Figure 6.6. The correlation with *in situ* estimates of live coral cover drops gradually as the spatial scale of IKONOS-based observations is increased beyond the 6 m radius. This is not surprising, since the spatial scale of the IKONOS-based observations becomes increasingly different from that of the *in situ* observations, which are based on substrate photos from the 5 metre-radius field site.

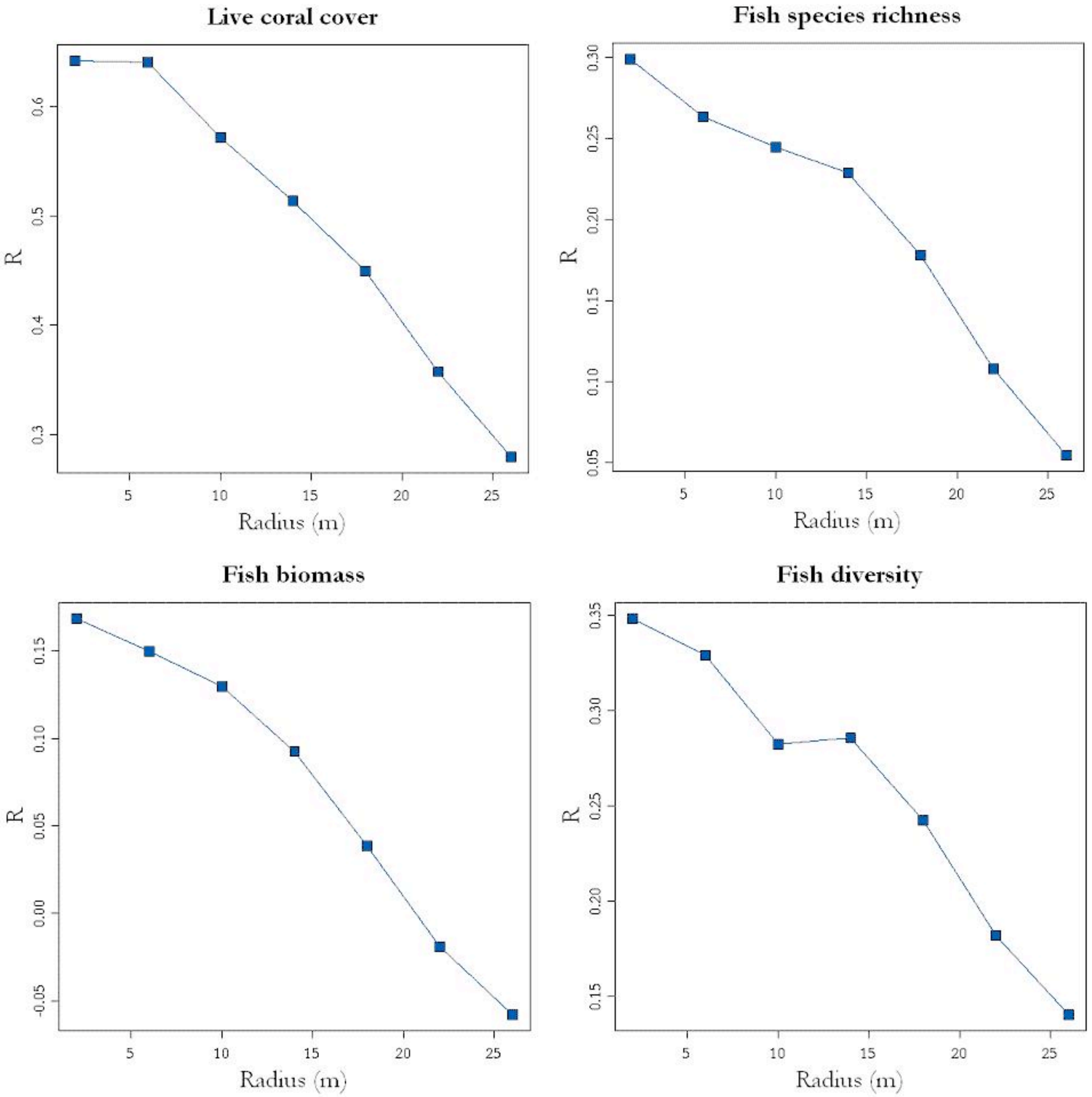

Figure 6.6: Correlations between IKONOS-based live coral cover estimates and *in situ* estimates of live coral cover, fish species richness, fish biomass and fish diversity, respectively. Spearman rank correlation coefficients shown on y-axis.

With only slight differences, similar trends are seen for correlations with the three fish variables. Compared to *in situ* live coral cover estimates, all correlations with fish variables are reduced when live coral cover is estimated from the IKONOS data (compare with Table 5.4). The trends illustrate that the reduced accuracy of IKONOS-based estimates affect the correlations with the fish variables, even when estimates are made at a similar spatial scale. More importantly, they illustrate that it is the coral cover in the immediate vicinity of the field site (e.g. within the 5 metre radius used in this study) that influences the fish community and that coral beyond this radius is of limited importance.

A likely explanation for this is that those fishes most dependent on the coral cover fall into two groups, the obligate corallivores that depend on the coral for food, and the herbivorous damselfishes that depend on the coral for shelter. Both these groups have limited mobility and are therefore little influenced by coral cover more than 5 m away from the centre of their territory.

6.2 Remote sensing of depth and rugosity

As outlined in chapter 4, remotely sensed estimates of depth form the foundation of rugosity estimates, and their precision is therefore paramount. *In situ* depth measurements were estimated from the IKONOS imagery with an RMSE of 1.07 m (R^2=0.73), which is in the range of previously reported results (Lyzenga et al. 2006; Muslim and Foody 2008; Purkis et al. 2008; Stumpf et al. 2003; Su et al. 2008). Several issues reduced the accuracy with which depth could be estimated in our study. The original IKONOS data contained a substantial amount of noise, and the study area contains a range of substrate types that differ strongly in spectral reflectance characteristics. This creates a difficult situation for remotely sensed depth estimation, only partly mitigated by existing methods (Su et al. 2008). In addition, the geolocation accuracy of the field sites (5-7 m) is a problem that may not have been completed resolved by the correction of GPS coordinates.

Remotely sensed estimates of rugosity differ between the *in situ* measures. At the 6 m radius, the depth range was estimated best (R^2=0.46), with both coarse rugosity (R^2=0.28) and fine rugosity (R^2=0.13) obtaining substantially lower coefficients of determination. Despite these relatively low values, all predictors performed substantially better than average-predictors when considering the RMSE values of predictions (1.08 vs. 1.59 for depth range, 1.02 vs. 1.20 for coarse rugosity, and 1.34 vs. 1.44 for fine rugosity).

6.2.1 Rugosity estimates and their spatial variation

The influence of the spatial scale of depth estimates was not investigated in a manner similar to that of live coral cover estimates, but rugosity estimates were derived from depth estimates at a range of scales. The remote estimations of rugosity and live coral cover differ in one important aspect. Estimates of live coral cover were formed by a simple averaging of the mapped live coral cover in the area in question, whereas the estimates of rugosity are calculated by a gradual coarsening of the pixel size. This coarsening will progressively reduce the importance of fine-scale depth variations (e.g. from large coral heads on the reef flat) while increasing the importance of coarse-scale variations (e.g. the transition from the shallow reef flat to the deep areas off the reef crest). In this way, a rugosity estimate is formed which is qualitatively different from a simple aggregation of *in situ* observations.

The influence of the spatial scale of these IKONOS-based estimates of rugosity is shown in Figure 6.7. Correlations show similar trends for most variables, with highest correlation coefficients at the smallest radius (6 m), decreasing sharply until the radius is around 100 m, then decreasing more slowly until the largest radius included in the study (300 m). However, there are a few notable exceptions to this rule. The correlation with coarse rugosity has a local maximum at the 112.5 m radius, after a local minimum at the 60 m radius. This pattern is unique to the combination of these two variables. The correlation with fish biomass is the only one to have its maximum at a large spatial scale (150 m radius). A map of rugosity at the 6 m and 150 m scales for Chumbe provides a visual clue to the difference in relative dominance by different reef features as the spatial scale changes (Figure 6.8). Smaller features on the reef flat west of the island are visible in the left image (6 m radius), along with artifacts from the island edge and the dense seagrass bed east of the island, where depth calculations suffered from large errors. These features are not discernable in the right image (150 m radius), where the reef edge is the all-dominating feature. Maximum correlations with the three fish variables are roughly similar, ranging from (Spearman) R=0.42 for fish diversity to (Spearman) R=0.51 for fish species richness.

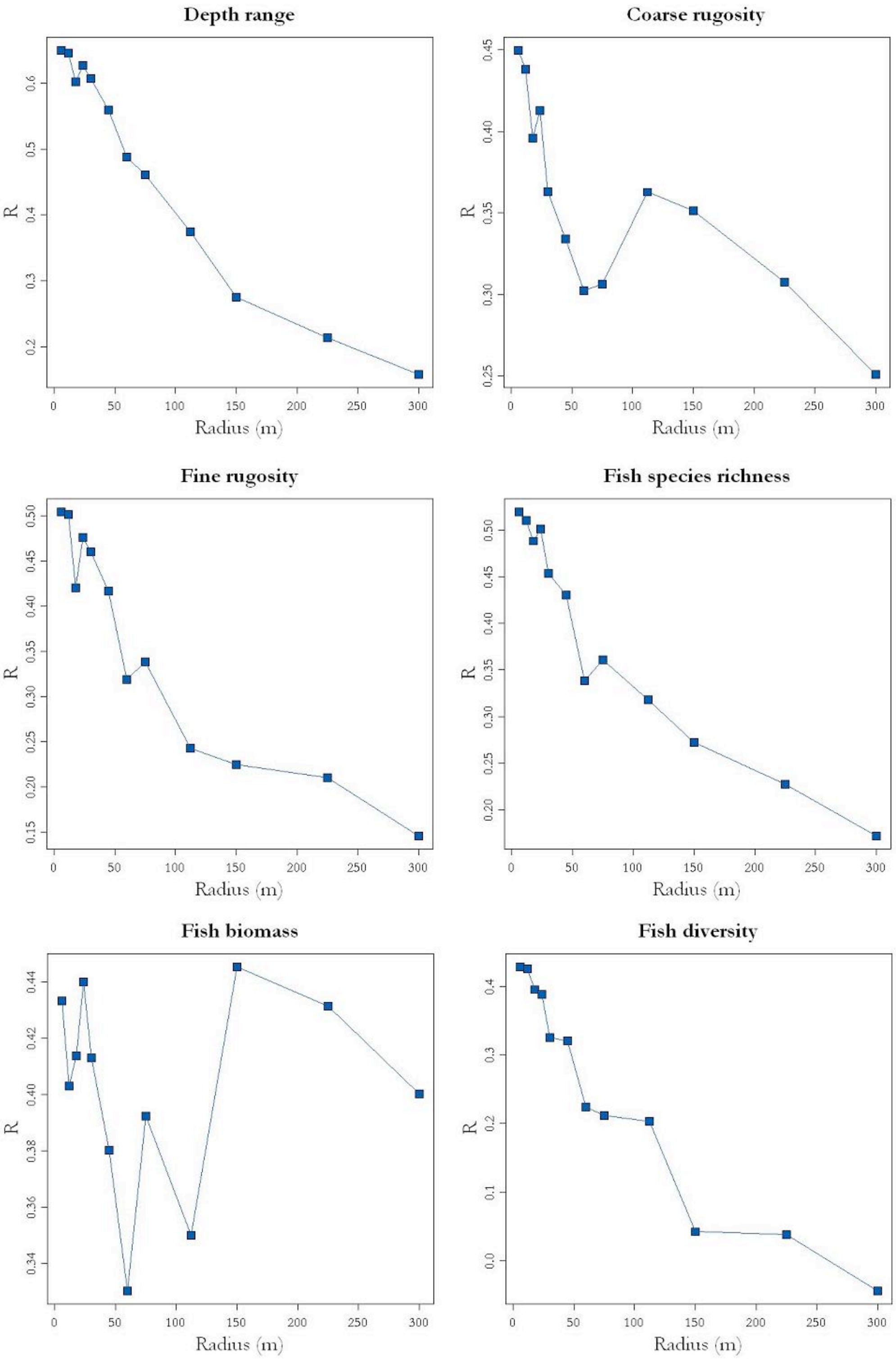

Figure 6.7: Correlations between IKONOS-based rugosity estimates and *in situ* estimates of depth range, coarse rugosity, fine rugosity, fish species richness, fish biomass and fish diversity, respectively. Spearman rank correlation coefficients shown on y-axis.

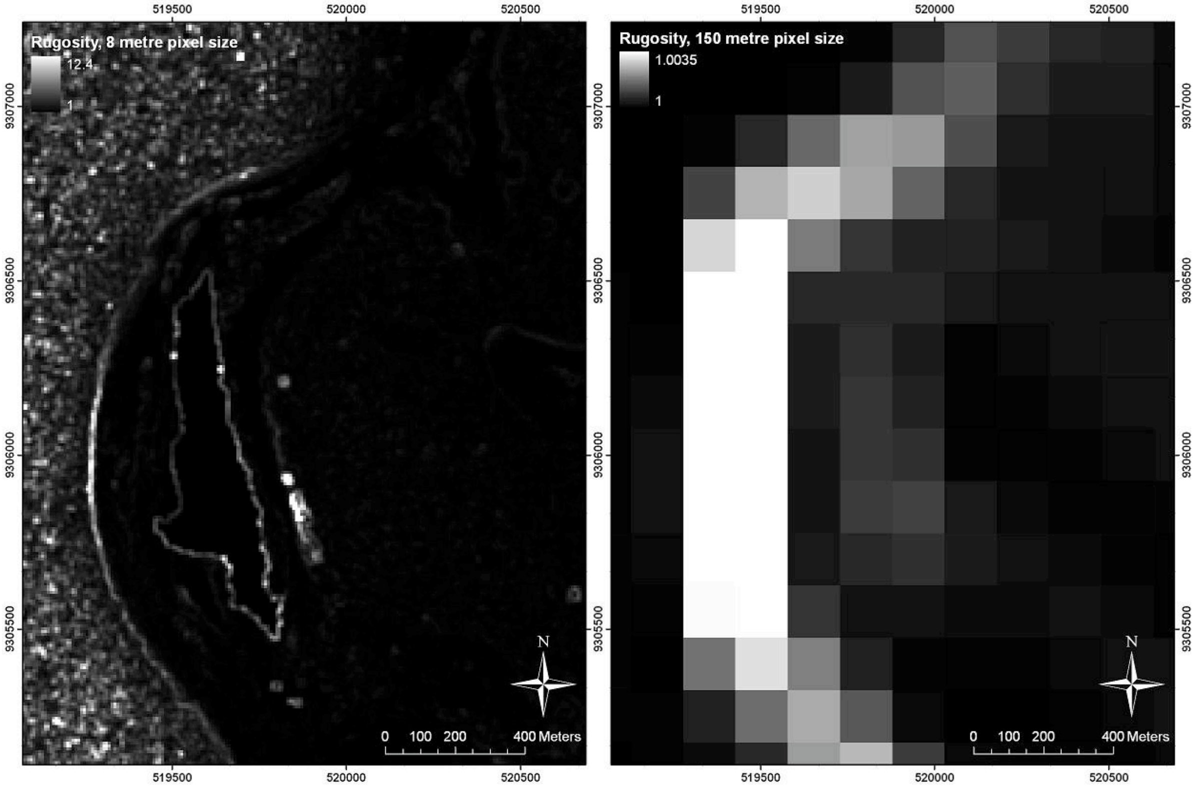

Figure 6.8: IKONOS-based rugosity estimates of areas around Chumbe Island, at pixel sizes of 8 and 150 m, respectively. The two maps cover the exact same area. Note the details on the reef flat west and south of the island in the left image (8 metre pixel size), and the complete dominance of the reef edge on the right image (150 metre pixel size).

6.3 Remote sensing of habitat variety

Remotely sensed estimation of the *in situ* substrate diversity variable was not considered feasible, and the results confirm this (see Figure 6.9). Regardless of the habitat variety measure used and the scale of calculation, the correlation with *in situ* substrate diversity is barely significant. This is not surprising, as the calculations of *in situ* and remotely sensed variety variables are based on different categories. The *in situ* estimates of substrate diversity are based on a large number of specific substrate types observed *in situ* (e.g. individual coral growth forms), whereas the remotely sensed estimates of habitat variety are based on broader habitat types, most of which contain a mix of substrates (e.g. 'sparse coral').

6.3.1 Habitat variety estimates and their spatial variation

The influence of spatial scale on correlations between habitat variety and four variables, *in situ* observations of substrate diversity and the three fish variables, is shown in Figure 6.9. Correlations are low for most variables, barely reaching statistical significance. The highest correlation is obtained with the fish diversity variable, which shows relatively high correlations with habitat diversity at the small spatial scales, the optimum being obtained using a 10 m radius. This correlation (Spearman R=0.27) is similar to that obtained by the *in situ* observations of substrate diversity (Spearman R=0.26, see Table 5.4). Given that the correlation with substrate diversity is near 0 at this scale (Figure 6.9), it is highly likely that the relationship between fish diversity and habitat diversity at the 10 m scale is based on an ecological relationship most appropriately observed at this spatial scale. The large difference in the spatial scale of measurement of *in situ* and remotely sensed habitat variables, and the fact that the substrate types that form the bases for the diversity calculations differ, suggest that these two variables represent somewhat unrelated quantifications of habitat diversity. Their very low correlation supports this view. The habitat variety measures should therefore not be seen as direct estimates of *in situ* substrate diversity, but rather as an extension of the diversity concept applied at large spatial scales.

6.3.1.1 Comparison of the variety measures

The two measures of habitat variety show some difference in their correlations with the three fish community variables, but neither generally outperforms the other. The habitat richness seems to perform better at large spatial scales whereas the habitat diversity measure performs better at small scales. This could suggest that the proportion of each substrate type (included in the diversity but not in the richness calculation) is important at small spatial scales (<100 m) but not at large ones (>100 m), though the reason for such a pattern is unclear. The habitat richness measure also produces more erratic correlations. This may be caused by a higher sensitivity to noise, which is reflected in the substrate classifications as individual incorrectly classified pixels. For these reasons, and due to the similar performance of the two measures, the habitat richness measure was excluded from further analysis and discussion.

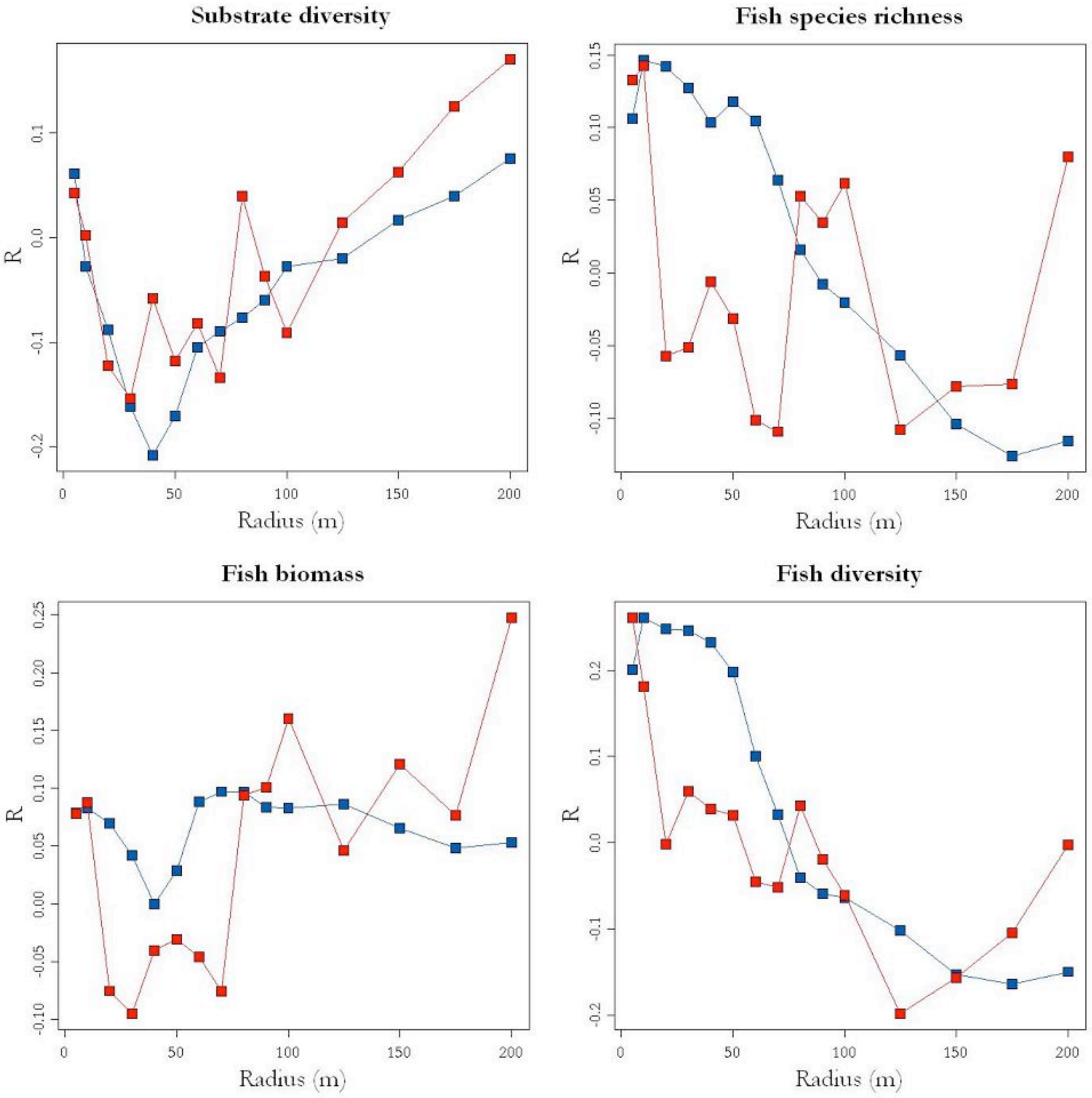

Figure 6.9: Correlations between IKONOS-based habitat variety estimates and *in situ* estimates of substrate diversity, fish species richness, fish biomass and fish diversity, respectively. The habitat richness measure is shown in red; the habitat diversity measure in blue. Spearman rank correlation coefficients shown on y-axis.

Due to a lack of functionality for diversity calculations in available software packages, it is currently easier to produce maps of habitat richness than for habitat diversity. Maps of habitat richness are shown in for Bawe in Figure 6.10 and for Chumbe in Figure 6.11. Both figures have been calculated using a 30 m radius. A summary of characteristic scales for each relationship is provided in Table 6.3.

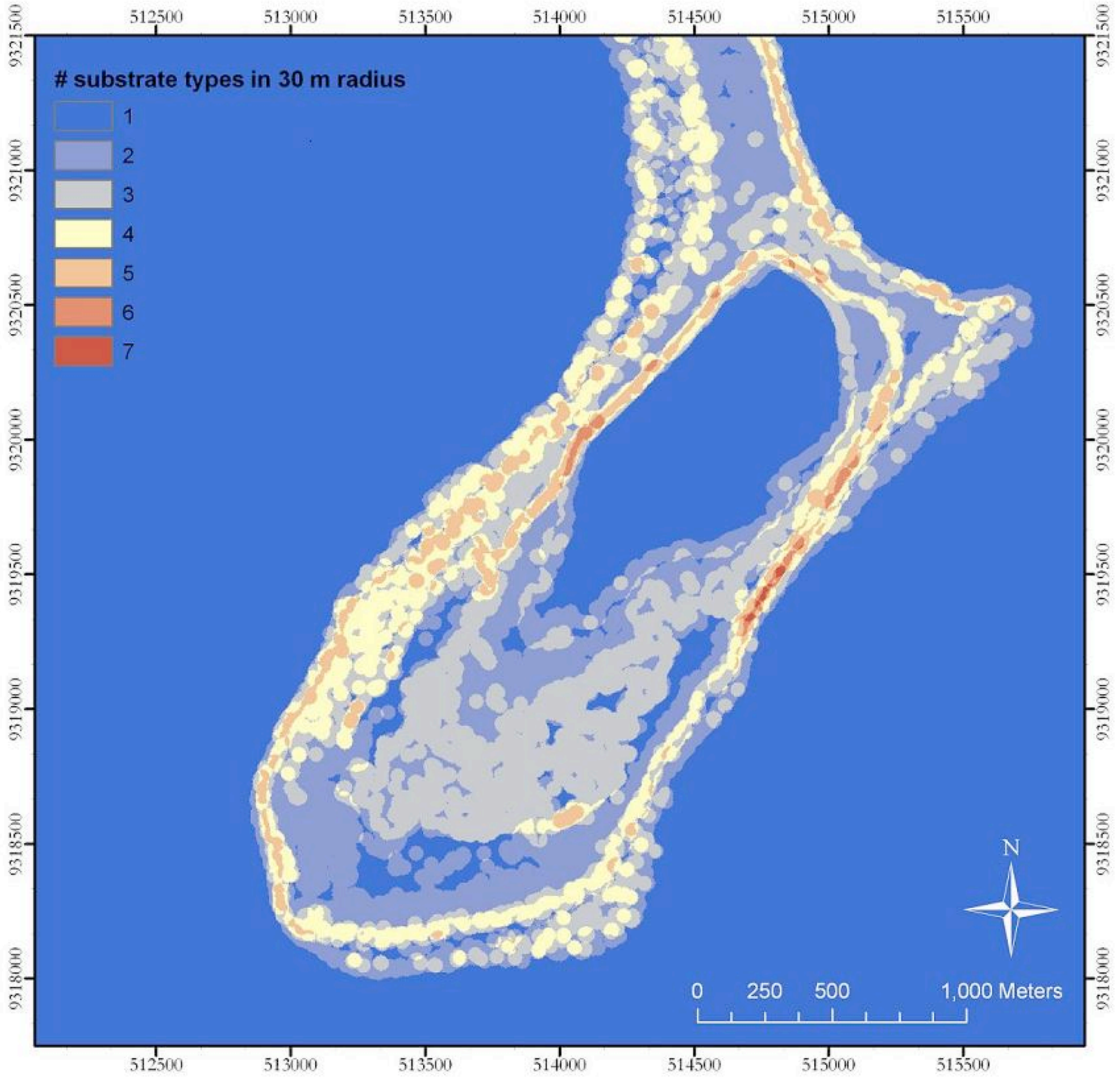

Figure 6.10: Map of habitat richness around Bawe Island. For each pixel, the map shows the number of different substrate types present within a 30 m radius.

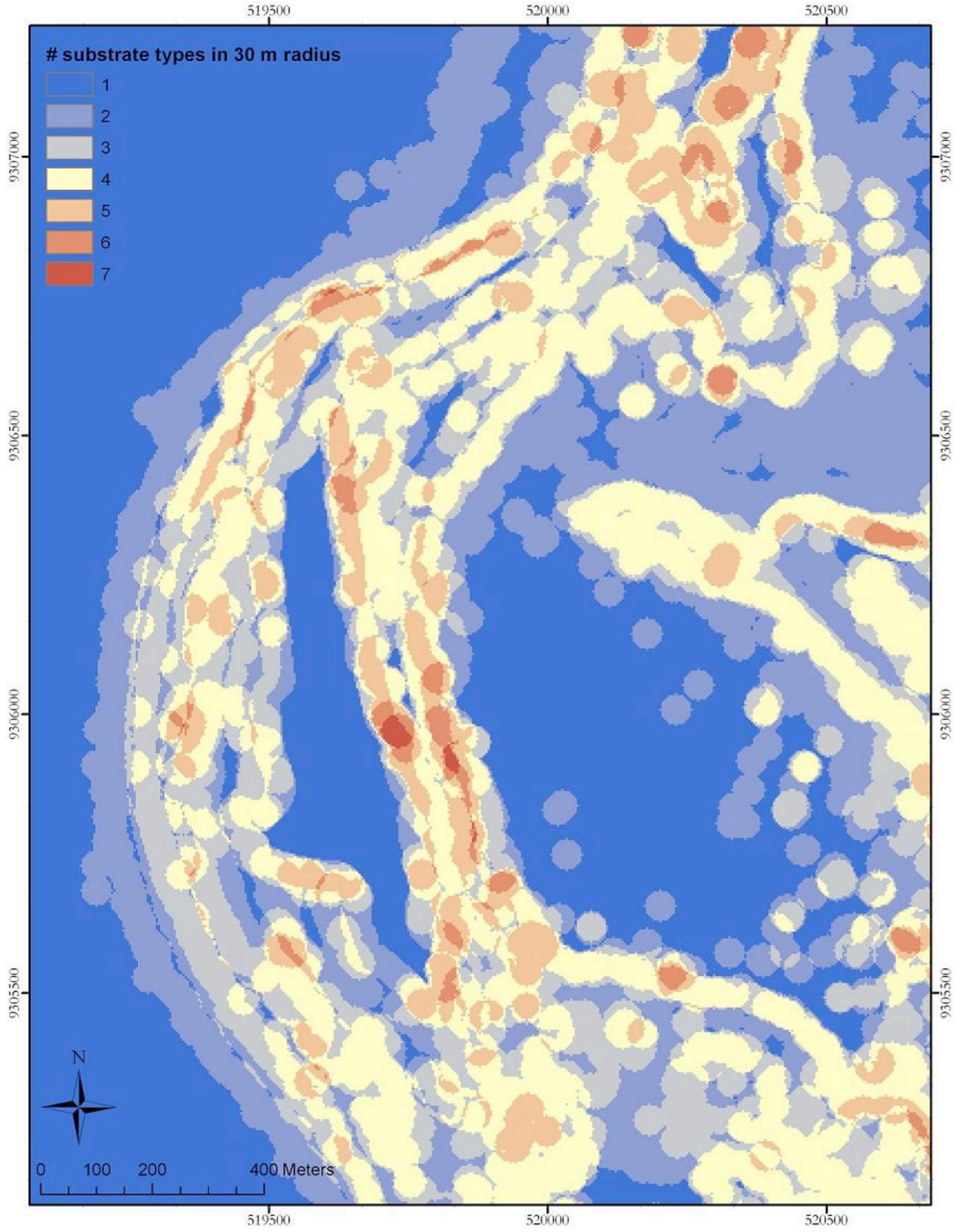

Figure 6.11: Map of habitat richness around Chumbe Island. For each pixel, the map shows the number of different substrate types present within a 30 m radius.

Table 6.3: Characteristic scales for fish-habitat relationships.

	Fish species richness	Fish biomass	Fish diversity
Live coral cover	2 m	2 m	2 m
Rugosity	6 m	150 m	6 m
Habitat diversity	10 m	60 m	10 m
Habitat richness	10 m	200 m	5 m

6.4 Discussion

6.4.1 Remote sensing of live coral cover

To my knowledge, this study is the first to estimate live coral cover directly from Lyzenga's (1978) depth-invariant index. This is surprising, given that a similar approach has proven effective for estimating seagrass biomass (Mumby et al. 1997a) and that coral's spectral signature differs substantially from the typical background signal (a sandy substrate), just as the spectral signature of seagrass does. We show that live coral cover can be estimated from the depth-invariant index, and that despite a relatively large RMSE the estimates are better than an average-predictor. In addition, we show that the resulting variables, as expected, are correlated with both live coral cover measured *in situ* and with the three response variables used here. The reason this result has been possible may be due to the very limited growth of algae in the study area. An area covered by algae is spectrally similar to coral, and when algae and coral grow in close proximity, the approach taken in this study is unlikely to work well. However, with the limited growth of algae, and a good spatial separation between coral and seagrass, the study area used here can be considered near optimum for IKONOS-based estimation of live coral cover. Hyperspectral sensors, however, have the ability to discriminate between the coral and algae using spectral unmixing (Goodman and Ustin 2007; Hedley et al. 2004), and are more likely to provide remotely sensed estimates of live coral cover in the future. However, the greatest potential for improvement mapping of live coral cover probably lies in combined hyperspectral/lidar systems, with their improved ability to separate the spectral influence of the water column from independent measurements of depth and water optical properties (Feygels et al. 2003; Tuell and Park 2004; Tuell et al. 2005), although practical applications of this technology have yet to emerge.

6.4.2 Remote sensing of depth and rugosity

The RMSE of depth estimates achieved (1.07 m) is comparable to that achieved by other studies deriving depth from IKONOS data in areas with spatial heterogeneity of substrate types and reflectance characteristics (Lyzenga et al. 2006; Muslim and Foody 2008; Stumpf et al. 2003; Su et al. 2008). Despite the theoretical potential to better separate the effects on reflectance of depth and substrate using hyperspectral imagery, improvements in the accuracy of depth estimates have been limited (Brando et al. 2009; Goodman and Ustin 2007; Klonowski et al. 2007; Lee et al. 2001; Lee et al. 1999). Lidar instruments, on the other hand, routinely produce spatially distributed depth estimates with RMSE values around 15 cm, and the development and application of this technology holds the greater potential for producing accurate high-density depth estimates over large areas. However, errors in depth estimates were shown by Su et al. (2008) to be spatially autocorrelated, which limits the impact of imprecise depth estimates on rugosity values. A comparison of prediction errors from models using rugosity values derived from IKONOS and lidar data, for the same set of response variables, would shed further light on the importance of precise depth predictions. In addition, lidar instruments are able to map water depth, and hence rugosity, as depths up to 70 m in clear water (Finkl et al. 2005), more than 3 times that of passive optical instruments. For a general application of the predictive models we worked with, that may prove to be of greater importance than the precision of depth estimates.

6.4.3 Remote sensing of habitat diversity

The habitat diversity variable used in our study is a measure of the alpha-diversity of habitats within the radius of observation, calculated on the basis of a habitat map with user-defined classes. In addition to investigating the influence of variations in radius (as in our study), there is thus room for experimentation with both thematic resolution of the map (number and kind of classes), as well as the quantification of diversity (e.g. the two measures used in our study). Although the influence of variations in thematic and spatial resolution has been investigated for habitat maps (Kendall and Miller 2008), the subsequent influence on predictability of fish community variables has not been investigated. In addition, quantification of habitat diversity can be expanded to take functional differences between habitat types into consideration, e.g. producing a higher diversity value in areas where both seagrass and coral is present than in areas where sparse coral and dense coral is present (Mumby 2001). In addition, a measure of beta-diversity (i.e. the change in habitat similarity along a gradient) may provide further information on the influence of habitat diversity in structuring fish communities (Harborne et al. 2006). The improvement in performance of predictive models that such habitat variables can bring has yet to be investigated.

6.4.4 Characteristic scale of the relationship between rugosity and fish species richness

Despite its utility for conservation planning, a limited number of studies have documented scale effects on fish-habitat relationships on coral reefs. The results presented here therefore add to a small but growing number of studies that together contribute to understanding the spatial scales of statistical relationships and ecological processes on coral reefs. A synthesis of research in this field is provided below. Due to the paucity of studies involving spatial scale variation and the variables live coral cover and fish biomass, the synthesis will focus on the remaining variables.

Purkis et al. (2008) investigated the influence of scale on relationships between IKONOS-based rugosity and fish community variables in Diego Garcia. Fish community variables included species richness as well as overall species abundance and a range of measures based on size, territoriality and diet. They found that the characteristic scale (kernel radius) for the rugosity-fish species richness relationship was 8 m. Characteristic scales for relationships with a number of other fish community variables in their study vary, but remain around the 8-20 m scale. These values correspond reasonably well with the results presented in chapters 6 and 7. Wedding et al. (2008), working with lidar-derived rugosity in Hawaii, reported slightly larger characteristic scales for rugosity-fish community relationships (37.5 m for both fish species richness and biomass, and no significant results for fish diversity), but very similar correlation coefficients at smaller scales do not allow confident inference of characteristic scales from their results. This is somewhat similar to our results, where correlation coefficients are similar for scales between 6 m and 24 m (see Figure 6.7). Also working with lidar-derived rugosity, Kuffner et al. (2007) found the highest correlations with fish species richness at a 2.5 m radius, though the coefficient of determination was very low. However, the environment in this study, a series of patch reefs in Florida, differ substantially from the continuous reefs studied by others, which is a likely cause of both the weak results and small characteristic scale. A study by Pittman et al. (2007) departs slightly from the others in using a regression tree approach as opposed to some form of linear regression, and by working at larger spatial scales. Using two measures of structural complexity (rugosity and bathymetric variance), the pruned regression tree used only two variables, rugosity at 42.5 m scale, and bathymetric variance at 22.5 m scale.

The results of these studies have been summarized in Table 6.4, but no specific scale can be derived as optimum from this set of studies. The difference in specific species surveyed in each study, and the range of environmental conditions such as reef type and depth, are likely explanations for this. Nevertheless, two observations can be made. Firstly, the smallest of the investigated scales only provide the best correlation in the present study, where differences between results up to the 24 m scale are negligible. In addition, two of the studies found the characteristic scale to be larger than 30 m. This could indicate that medium resolution data (15-20 m pixel size) may achieve similar results as those achieved in our study with IKONOS data. Secondly, the range of characteristic scales does not exceed 50 m for any study, despite 3 of the studies investigating variables at much larger scales. Although not pointing to a single characteristic scale where the relationship between rugosity and fish species richness is best observed, these observations suggest a range of likely characteristic scales for environments similar to those investigated in the reported studies.

Table 6.4: Characteristic scales for relationships between remotely sensed rugosity and fish species richness from recent studies. Note that for several studies, other scales than the one noted obtained very similar correlations between the two variables.

Study	Characteristic scale	Scale range investigated	Reef type	Depth range
Our study	6 m	6-300 m	Continuous reef	0-10 m
Kuffner et al. (2007)	2.5 m	1-5 m	Patch reefs	3.5-5.5 m
Pittman et al. (2007)	42.5 m	7.5-322.5 m	Mix	1-30 m
Purkis et al. (2008)	8 m	4-200 m	Continuous reef	4-6 m
Wedding et al. (2008)	37.5 m	6-37.5 m	Continuous reef	1-24 m

6.4.5 Characteristic scale of the relationship between habitat variety and fish species richness

Even fewer studies have investigated the influence of habitat variety on reef fish communities. Pittman et al. (2004) showed that several indices of landscape structure, particularly the abundance of seagrass and mangrove habitat, influence the juvenile fish and prawn community in mangroves and seagrasses in Deception Bay, Australia. The study also determined that using 300 m radii as the basis for calculations of habitat indices was significantly better than using 100 m radii, giving a first indication of characteristic scale. Purkis et al. (2008) also investigated the relationship between habitat evenness and fish community variables, and found the characteristic scale for the habitat evenness-fish species richness relationship to be 40 m, although the statistical significance of the relationship was not investigated. Another study (Pittman et al. 2007) investigated but found no significant correlations between habitat richness and the fish community. Results from the available studies are summarized in Table 6.5. Although a characteristic scale is presented for the present study, it should be kept in mind that following multiple-testing correction, the correlation coefficients found in this study are not statistically significant.

Table 6.5: Characteristic scales for relationships between remotely sensed habitat diversity and fish species richness from recent studies. Note: Habitat variety has been quantified using different approaches in all four studies. N.S. = no significant relationships found.

Study	Characteristic scale	Scale range investigated	Reef type	Depth range
Our study	10 m (N.S.)	5-200 m	Continuous reef	0-10 m
Pittman et al. (2004)	300 m	100 m, 300 m	Tidal flat	3.5-5.5 m
Pittman et al. (2007)	N.S.	31.8 m	Mix	1-30 m
Purkis et al. (2008)	40 m	4-200 m	Continuous reef	4-6 m

As for the rugosity variable, these studies do not converge on a common characteristic scale, but they do suggest that habitat variety is best quantified at larger spatial scales than rugosity. It is also noteworthy that the one other study that failed to find significant relationships (Pittman et al. 2007) was conducted at a large number of sites covering a mix of reef types, and it is possible that a single measure was unable to adequately describe habitat diversity for this range of environments. More studies, covering a range of habitat variety measures, reef environments, and spatial scales are necessary to enable conclusions to be drawn in this area.

6.4.6 The targeted landscape approach

A number of other studies have specifically targeted fish-habitat relationships that are based on known ecological links, such as the relationship between fish species richness/abundance on a reef and the presence of nearby spawning/nursery grounds (Nagelkerken et al. 2000). However, most have not been spatially explicit enough to infer the characteristic scales of those relationships. Dorenbosch et al. (2004) showed an influence from nearby bays with seagrass and mangrove habitats on the abundance of a number of reef fish species in Curaçao, but focused on fish species with a known seagrass dependence and did not analyse sites based on exact distances to the bays. Similarly, Mumby et al. (2004a) similarly showed influence on reef fish community structure from nearby mangroves, but contrasted reefs that had nearby mangroves with reefs separated from mangroves by more than 15 km. Though such studies are valuable, they do not allow inference of the spatial scales (distances) at seagrass beds and mangroves can function as nursery habitats for reef fish. One exception is the study by Grober-Dunsmore et al. (2007), who specifically investigated the spatial scale of seagrass-reef fish relationships, and found an influence on reef fish species richness from the amount of seagrass cover in radii as great as 1 km, though the characteristic scale was 250 m (range: 100 m-1 km). The characteristic scale may in this case reflect the distance young adults migrating from seagrass beds to reefs typically are able to cover. Such studies are needed for more substrate types (e.g. mangroves), from more regions of the world, and for functional groups and individual fish species.

6.5 Summary

Sub-question B was "*How accurately can habitat variables be estimated remotely, and at what spatial scales are these variables most predictive for the fish community?*". The results presented in this chapter answer the first part of that question for the variables live coral cover, depth, structural complexity ('depth range'), and substrate diversity. RMSE values for IKONOS-based estimates of these variables are roughly 20 percentage points for live coral cover, 1 m for depth, and 1 m for depth range, all with Spearman R values above 0.6. Although further improvement of these values is possible with improved data sources, our results indicate that IKONOS data do provide a means to create fairly accurate and spatially explicit estimates (maps) of these three variables. This is not the case for the *in situ* measure of substrate diversity, with which the IKONOS-based measures of habitat richness and diversity are not significantly correlated. Nevertheless, the remotely sensed measures of habitat variety show significant correlations with the three fish community variables.

The second part of the question was answered by comparing correlations between the fish community variables and the remotely sensed habitat variables at a range of spatial scales. For both fish species richness and diversity, the fine spatial scales (radii < 10 m) produced the highest correlations for all habitat variables, indicating that it is the immediate environment that influences these two aspects of the fish community. For fish biomass, however, calculations at coarse spatial scales produced the highest correlations for rugosity (radius = 150 m) and habitat diversity (radius = 60 m, not statistically significant), although it was still the finest spatial scale that produced the highest correlation with live coral cover (radius = 2 m). The biomass variable is highly influenced by large individuals and roaming schools of medium-sized fish, both of which are most commonly found near the reef edge. The coarse scale rugosity calculations most likely achieve improved correlations with the biomass variable by eliminating high frequency noise as well as the fine scale rugosity on the reef flat, and focusing on the coarse scale rugosity caused by the change in depth from the reef flat to the deep areas outside the reef.

CHAPTER 7: REMOTE SENSING OF REEF FISHES

In this chapter we bring together the predictive model types compared in chapter 5 and the remotely sensed estimates of habitat variables from chapter 6. Only results obtained with cross-validation resampling are presented in this chapter. The performance of all six model types is compared using four different sets of habitat variables as input data: 1) the *in situ* data already analyzed in chapter 5, 2) the full range of habitat variables derived from IKONOS data, 3) a subset of the habitat variables, limited to the spatial scales obtainable from Landsat TM imagery, called the simulated Landsat data, and 4) all habitat variables, both *in situ* and remotely sensed.

We then focus on the question of which remotely sensed habitat variables are important for predictions of the fish community. The volume of the variable importance results makes reporting all of them unfeasible, and thus only selected results are reported here, though summary plots of all results are available upon request. First, the Bagging model is used to illustrate those variables important for models based on the IKONOS data set, because it is the median-performing tree-based models and provides near-optimum prediction accuracy. For spatially distributed predictive models of the fish community, the identified variables will be important to derive as accurately as possible from IKONOS (or other) satellite imagery. Secondly, results from all model types predicting species richness using the "All" data set are used to further illustrate the influence of model type on variable importance. The "All" data set is used to also derive those remotely sensed habitat variables that not only estimate important *in situ* variables, but provide complementary information to the "*in situ*" data set.

Together, the comparison of prediction accuracies and habitat variable importance provides insight into the ability of remote sensing to produce spatially distributed predictions of fish community variables, and its ability to derive habitat variables of importance to the fish community, both those that operate inside and outside the range of spatial scales accessible with conventional fieldwork. Together, these investigations provide an answer to sub-question C. Finally, as an example of what the approach can produce, and map of species richness around Chumbe Island, predicted with the Bagging model and the IKONOS data set, is presented.

7.1 Comparison of predictive model accuracy by data set

The distributions of prediction accuracies for all models types, data set and response variables are shown in Figure 7.1; median values are also provided in Table 7.1. The prediction accuracies of models developed from *in situ* data have already been discussed in chapter 5, and they form a benchmark against which to compare the accuracies produced by models based on remote sensing data.

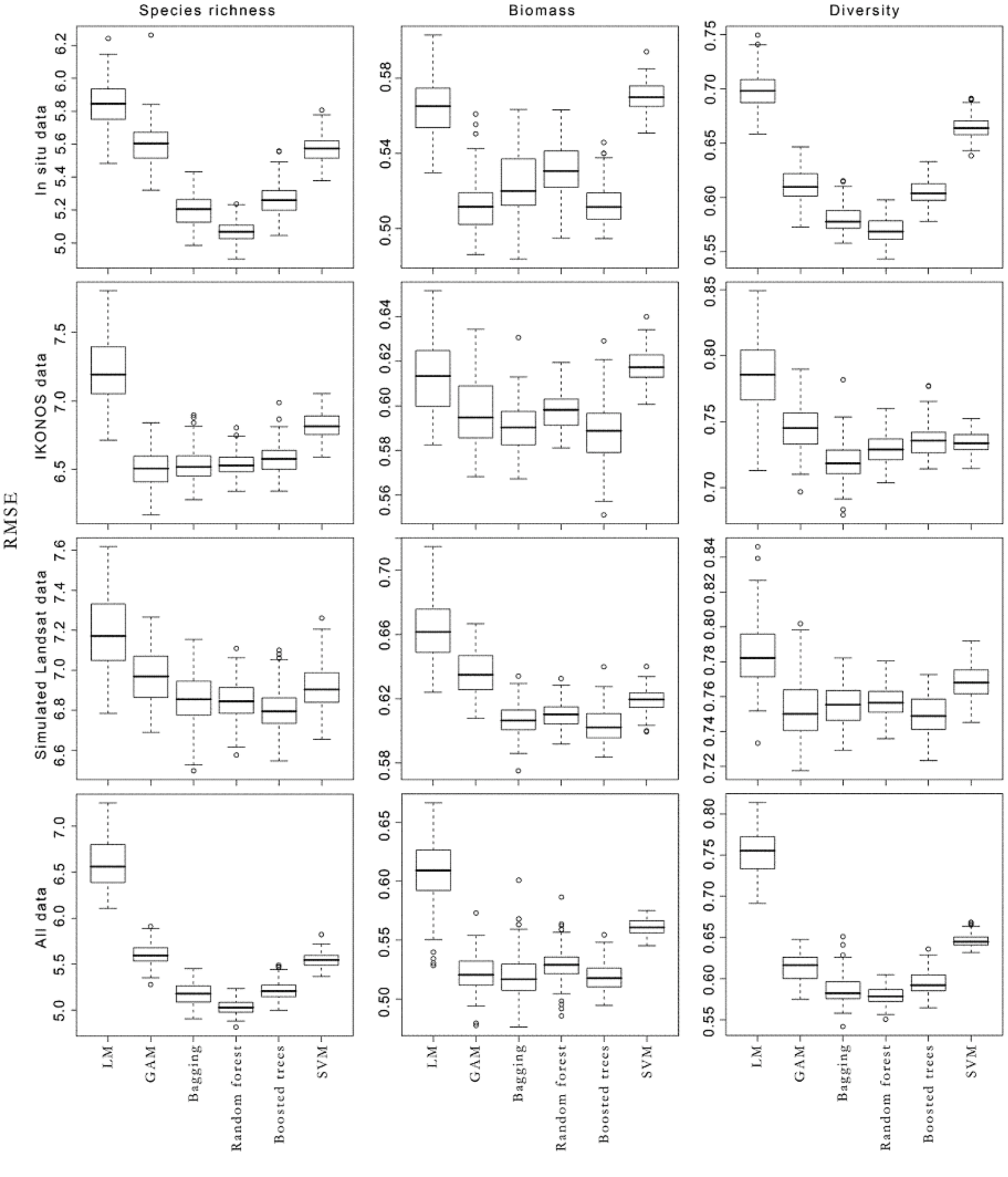

Figure 7.1: Prediction accuracies for all models types, input data, and response variable. Note the variation in y-axis ranges between plots.

The other benchmark against which to compare these models is the prediction accuracy of an "average-predictor", shown in Table 5.3, only predictive models with significantly lower RMSE values than the "average-predictor" can be worth developing.

The pattern of relative performance of model types based on *in situ* data is generally also found when the models are based on the other 3 data sets. As such, the LM consistently performs worse than any other model type, only occasionally beating the SVM, which itself is also generally outperformed by the GAM and the tree-based models. The tree-based models generally outperform the GAM, with a few exceptions including the following combinations: IKONOS - species richness, *in situ* – biomass, and simulated Landsat – diversity. The Random forest consistently performs worse than Bagging and Boosted trees when predicting biomass, while no other consistent trend is seen among the tree-based models.

Table 7.1: Median RMSE values for predictive models by input data, model type, and response variable.

Species Richness						
	LM	GAM	Bagging	Random forest	Boosted trees	SVM
in situ	5.84	5.60	5.20	5.07	5.26	5.57
IKONOS	7.22	6.51	6.53	6.53	6.58	6.82
Landsat	7.19	6.97	6.85	6.84	6.80	6.91
All	6.61	5.61	5.17	5.03	5.21	5.55
Biomass						
	LM	GAM	Bagging	Random forest	Boosted trees	SVM
in situ	0.564	0.513	0.523	0.531	0.513	0.570
IKONOS	0.613	0.598	0.591	0.598	0.588	0.618
Landsat	0.663	0.636	0.607	0.610	0.604	0.619
All	0.607	0.522	0.519	0.530	0.519	0.561
Diversity						
	LM	GAM	Bagging	Random forest	Boosted trees	SVM
in situ	0.699	0.611	0.580	0.570	0.605	0.665
IKONOS	0.745	0.745	0.719	0.729	0.735	0.735
Landsat	0.753	0.753	0.755	0.757	0.750	0.768
All	0.614	0.614	0.587	0.579	0.595	0.646

The model types differ in their ability to utilize the remotely sensed data in addition to the *in situ* data. One extreme is shown by the GAM, which produces less accurate predictions of all three response variables when using the “All” as opposed to the “*in situ*” data set. This may be due to the necessary exclusion of variables when working with the “All” data set, a problem that could be mitigated by iteratively searching for the variable combination that would produce the optimum result. The same pattern is seen for the LM models of species richness and biomass, and the (Bagging – diversity), (Random forest – diversity) and (Boosted trees – biomass) combinations. The other extreme is shown by the SVM, which produces the most accurate predictions for all response variables when using the “All” data set.

7.1.1 Models predicting species richness

Predictions from the Bagging model have been used to compare the results obtained with the 4 different data sets. The Bagging model has been chosen because it is the median-performing tree-based model for most predictions. All differences described in this section are statistically significant ($p<0.01$). IKONOS-based Bagging models of fish species richness produce significantly higher prediction errors than do the models based on *in situ* data, IKONOS-based models (RMSE=6.53, R^2=0.30) providing only 47% of the reduction in RMSE over the average-predictor (RMSE=7.72) seen in models based on *in situ* data (RMSE=5.20) (Figure 7.2). As expected, the models based on simulated Landsat data produce the highest errors. The improvement in predictive performance from adding remotely sensed data to the *in situ* data (the “All” data set) is very modest, reducing the RMSE from 5.20 to 5.17.

7.1.2 Models predicting biomass

A similar trend is seen for predictive models of biomass, where the IKONOS-based models (RMSE=0.591, R^2=0.25) provide only 35% of the improvement over the average-predictor (RMSE=0.628) compared to models based on *in situ* data (RMSE=0.523). Again, models based on simulated Landsat data produce the least accurate predictions, and the improvement in accuracy gained from adding remotely sensed data to the *in situ* data is modest, reducing RMSE to 0.519 (Figure 7.3).

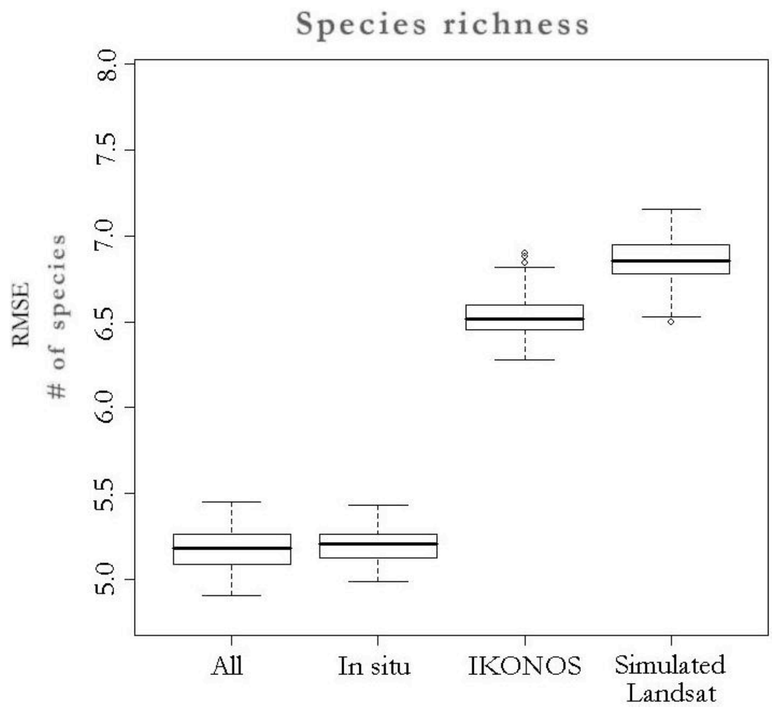

Figure 7.2: Comparison of species richness prediction accuracy from Bagging models using the four different data sets.

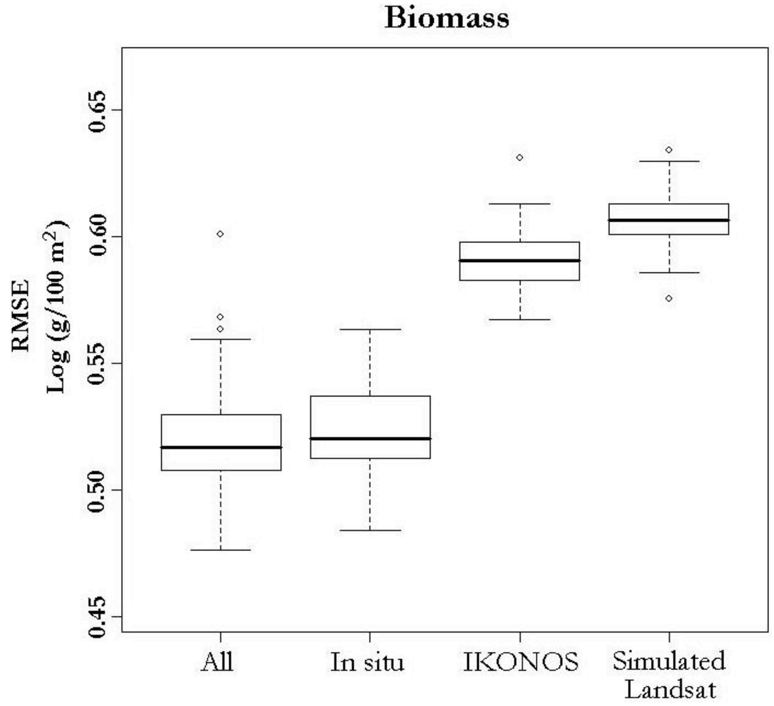

Figure 7.3: Comparison of biomass prediction accuracy from Bagging models using the four different data sets.

7.1.3 Models predicting diversity

The pattern for predictions of diversity is also similar to that of species richness. IKONOS-based models (RMSE=0.719, R^2=0.23) provide 34% of the accuracy improvement over the average-predictor (RMSE=0.792) as compared to the *in situ*-based models (RMSE=0.580), and the models based on simulated Landsat perform slightly worse than those based on IKONOS data. However, the addition of remotely sensed data to the *in situ* data actually reduces accuracy for predictions of diversity, increasing RMSE to 0.587 (Figure 7.4), probably due to overfitting with the large number of variables in the "All" data set.

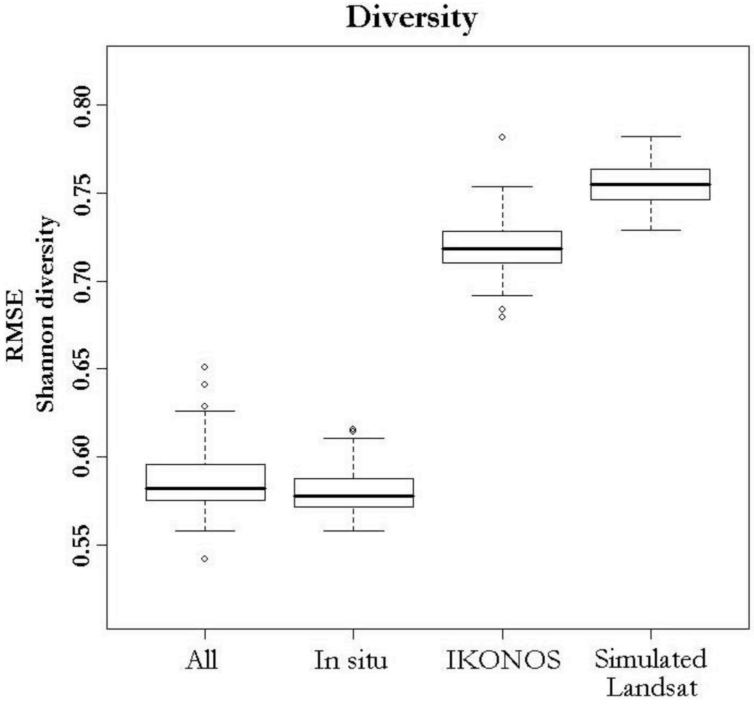

Figure 7.4: Comparison of diversity prediction accuracy from Bagging models using the four different data sets.

Together, these results illustrate not only the importance of spatial resolution, but also underline that an improved remote sensing-based estimation of the habitat variables observed *in situ* is likely to improve predictions. Such improved estimation is likely to come from a combination of improved spatial and spectral resolution of future sensors, most promisingly the improved derivation of depth and substrate composition possible with hyperspectral data.

7.2 Variable importance results

The results presented in this section will be limited to two foci. The first part will focus on the variables important for models based on the IKONOS data set, using the Bagging model. The second part will focus on those remotely sensed variables important in models using the “All” dataset, which will give an indication of which variables provide complementary information to that which can be obtained *in situ*.

7.2.1 Important variables from the IKONOS data set

As seen with the *in situ* data set, the important variables for predicting species richness and diversity are similar, and the importance is concentrated on fewer variables than is the case for predicting biomass. The dominant predictor for both species richness and diversity is rugosity at the finest scale possible (r=6 m), which is also the IKONOS-based variable most highly correlated with the *in situ* ‘depth range’ (Spearman R=0.65). Other variables provide minor contributions, including rugosity at the second-finest scale (r=12 m), depth, and substrate class, as well as habitat diversity at relatively fine scales (r=10 m and r=20 m) and, to a very small degree, live coral cover (r=2 m) (Figure 7.5). For biomass predictions, the fine scale rugosity (r=6m) is also the most important variable, but is followed closely by rugosity at two coarse scales (r=225m and r=300m), as well as fine scale habitat diversity (r=10m) and depth. Substrate class, along with a few other variables (rugosity (r=45m), habitat diversity (r=20m)), provide minor contributions.

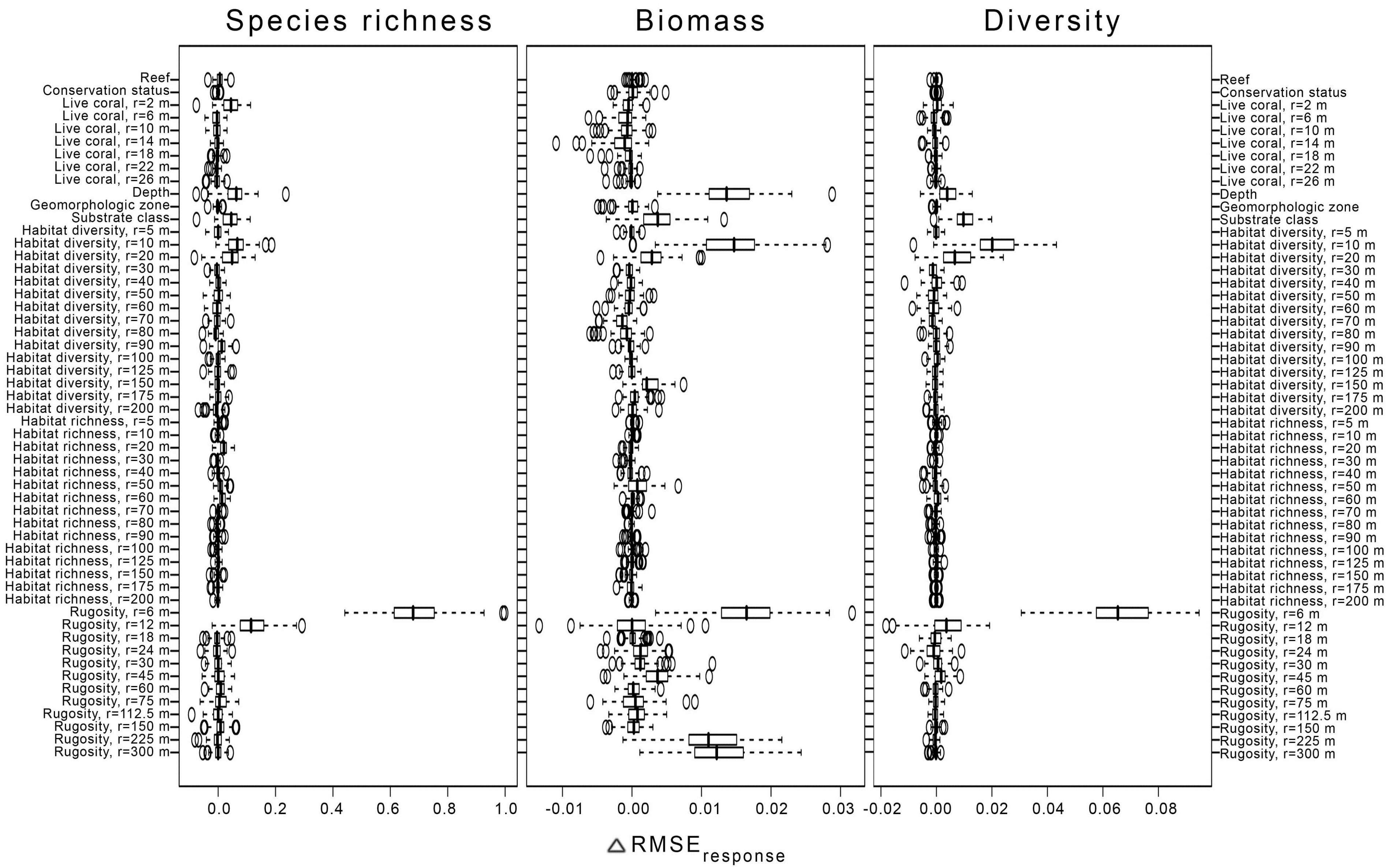

Figure 7.5: Variable importance from Bagging models based on the IKONOS data set.

7.2.2 Remotely sensed variables with complementary information from the "All" data set

Variable importance for all models predicting species richness with the "All" data set is shown in Figures 34-36. As presented earlier for models based on *in situ* data (Figure 5.5), variables identified as important in the LM differs substantially from those identified in other models. As such, the two most important variables in the LM model are identified as remotely sensed 'habitat diversity' (r=60 and 70 m). Remotely sensed live coral cover (r=14 and 18 m) have a small, though highly variable, importance, along with several variables measured *in situ* ('depth', 'depth range', 'substrate diversity', live coral cover', and '# of growth forms'). Important variables in the GAM model more closely resemble those from the models based on *in situ* data, with 'depth range' being the dominant variable and minor importance seen for 'depth', 'coarse rugosity', 'substrate diversity' and 'live coral'. Of the remotely sensed variables, only the 'substrate class' has minor importance (Figure 7.6). However, interpretation of the results from the GAM model is complicated by the computational limitation of 17 habitat variables in this model. Models not included in the model have zero importance, as seen in Figure 7.6. Results from the Random forest model (Figure 7.7) resemble those of the GAM, with only rugosity (r=6 m), of the remotely sensed variables, having minor importance.

In the Bagging model, no remotely sensed variables have importance significantly different from zero. Similar results are also seen in the Boosted trees model Figure 7.8, where the remotely sensed variables have very little importance. Variable importance in the SVM is different, though, with importance spread over a larger number of variables, and both rugosity (r=45 m) and remotely sensed depth have some importance.

Across all model types except LM, remotely sensed variables thus have little importance, i.e. they provide little complementary information to that already contained in the *in situ* variables. The importance of 'habitat diversity' (r=60 and 70 m) in the LM model is probably an artifact caused by the sensitivity of this model to variables with outliers, which the 'habitat diversity' variable has at the 60 and 70 m radii. The best performing models (the tree-based models in general, Random forest in particular), show that of the remotely sensed variables only rugosity (at varying radii depending on model) has some importance.

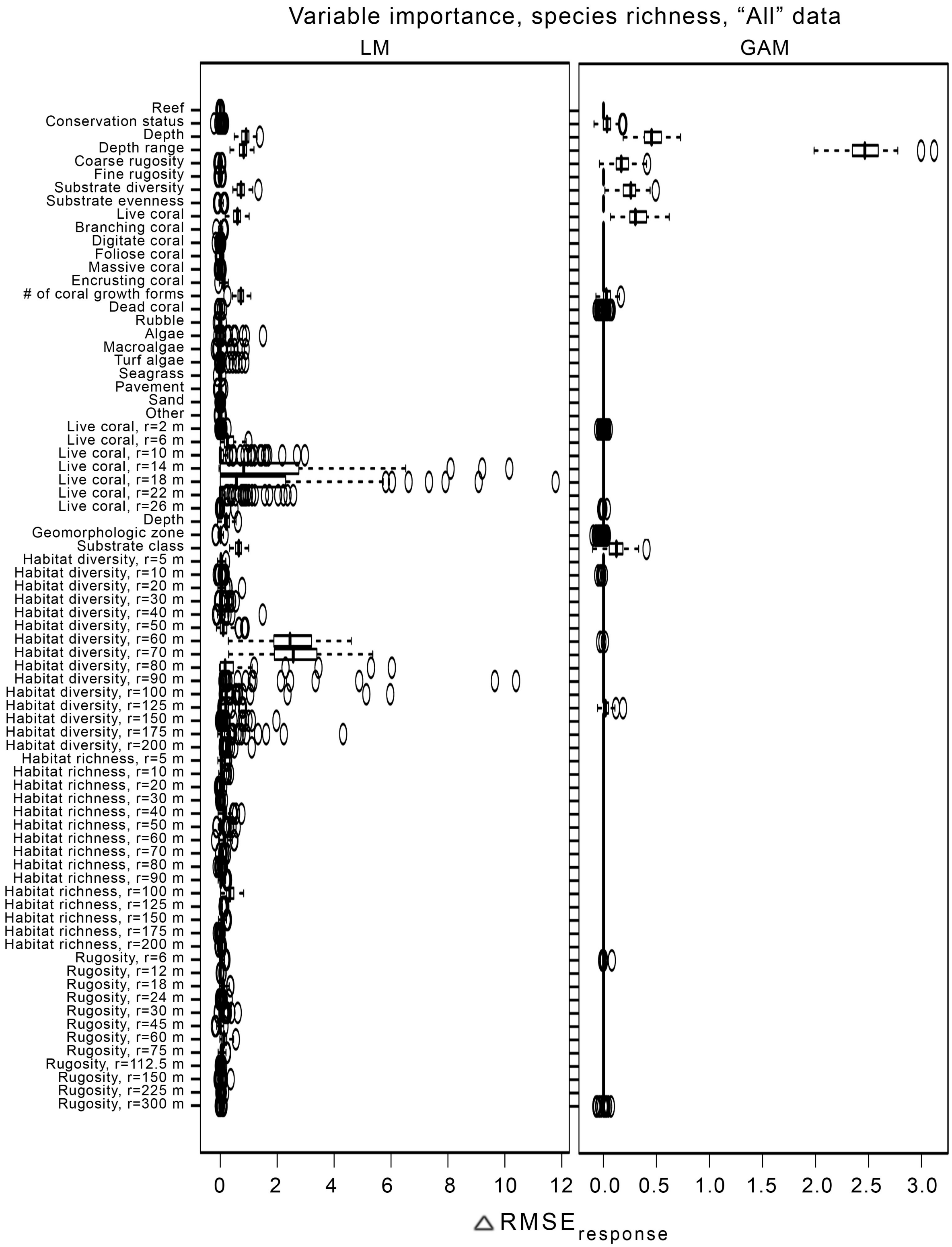

Figure 7.6: Variable importance for species richness prediction using the LM and GAM models, based on "All" data. The many variables with zero influence in the GAM models are caused by the maximum of 17 variables that could be included in this model.

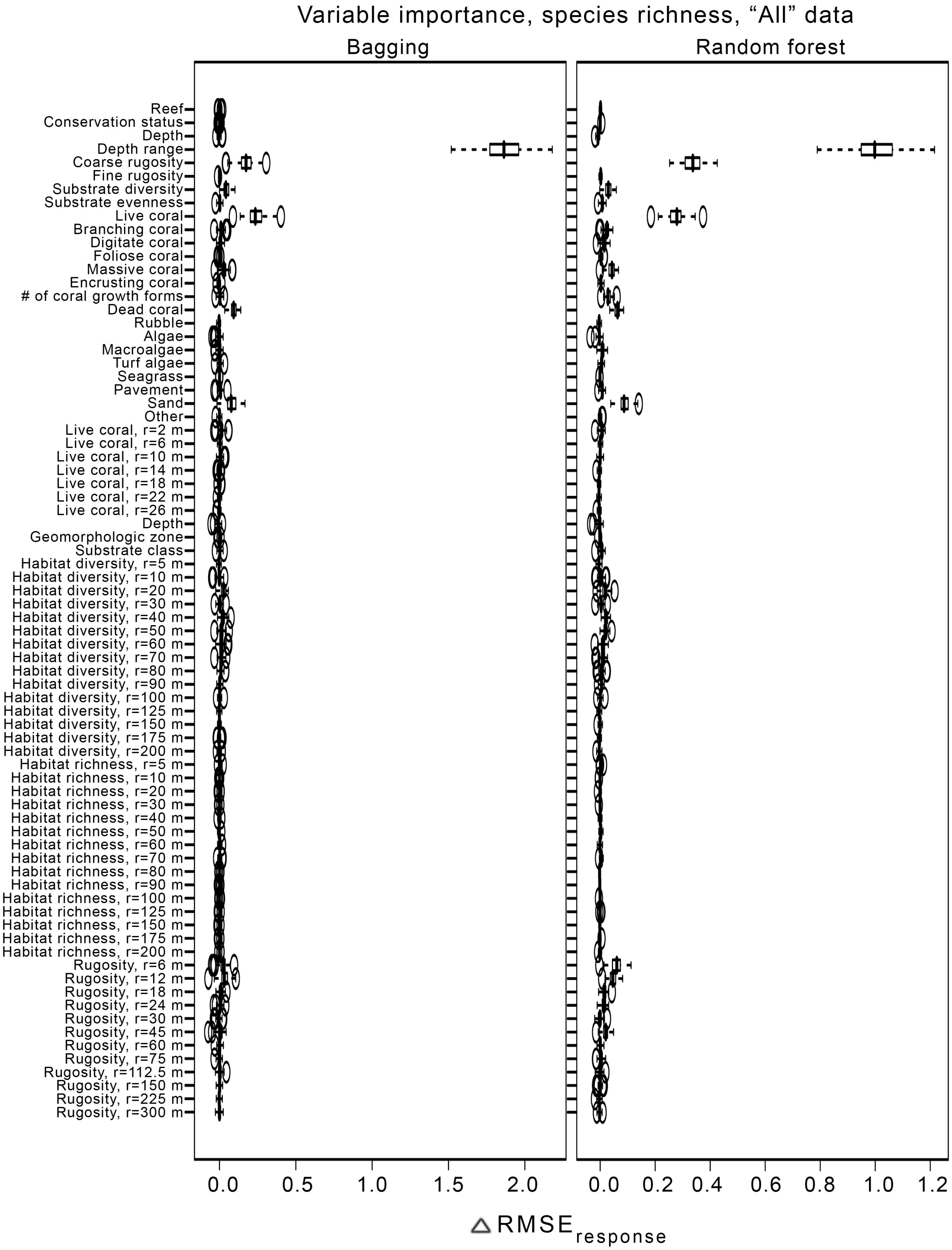

Figure 7.7: Variable importance for species richness prediction using the Bagging and Random forest models, based on "All" data.

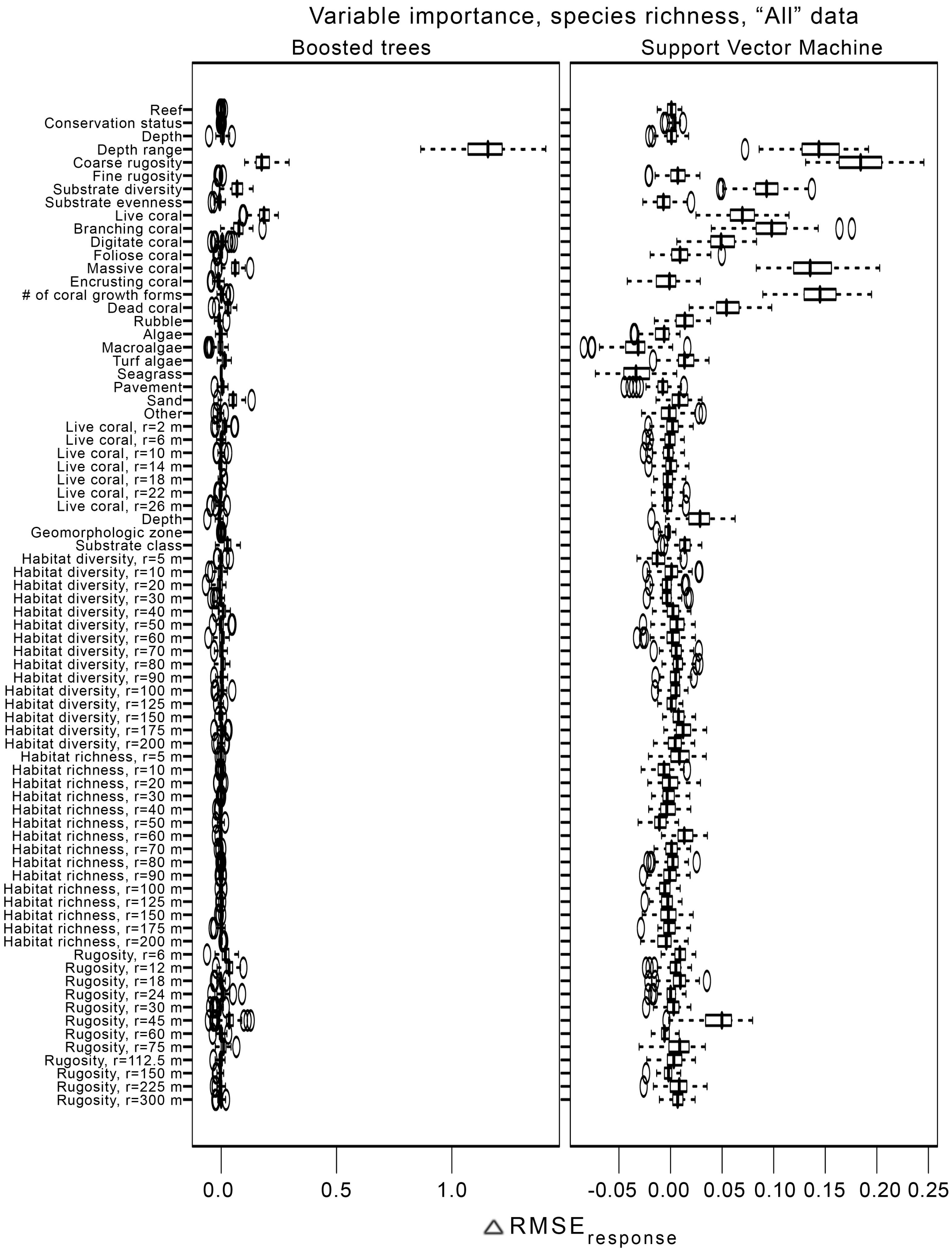

Figure 7.8: Variable importance for species richness prediction using the Boosted trees and SVM models, based on "All" data.

7.3 Map of predicted species richness, Chumbe Island

Using the Bagging model and the IKONOS data set, Figure 7.9 shows spatially distributed predictions of fish species richness in the reef areas around Chumbe Island. Based on knowledge of the area, the Bagging model seems to produce very reasonable predictions for the area not covered by the field data. The map shows many fish species near the edge of the reef west of the island (green area), and fewer species in the deeper waters off the reef (pale blue area). Sandy areas stretching north-east from the island's north tip has few species (blue areas), except at the seagrass-covered edges of the sand bar (green and orange areas).

However, the predictions in the lagoon east of Chumbe (large grey area) are probably too high. This area has a flat and sandy bottom, and a very sparse fish fauna. Some noise is also seen in the south-west corner of the image, where high predictions result from erroneously high rugosity estimates.

7.4 Discussion

The largest contribution of remote sensing to predictive modelling of fish community variables is undoubtedly the spatial coverage of remotely sensed data, which allows predictive models to become spatially distributed and explicit. However, our results show that this spatial coverage comes at a price – increased prediction errors. Prediction errors increase because, compared to the habitat variables derived from *in situ* data, habitat variables derived from the remotely sensed information are relatively poorer estimates of the aspects of the habitat that influence the fish community. However, the spatial coverage of remotely sensed data also allows users to derive information about the habitat at spatial scales not measurable with traditional field-based methods. The net effect on prediction error therefore depends on how closely remotely sensed habitat variables can quantify the aspects of habitat that influence the fish community (chapter 6), as well as the ability of predictive models to utilize the additional information remotely sensed data provide at coarse spatial scales.

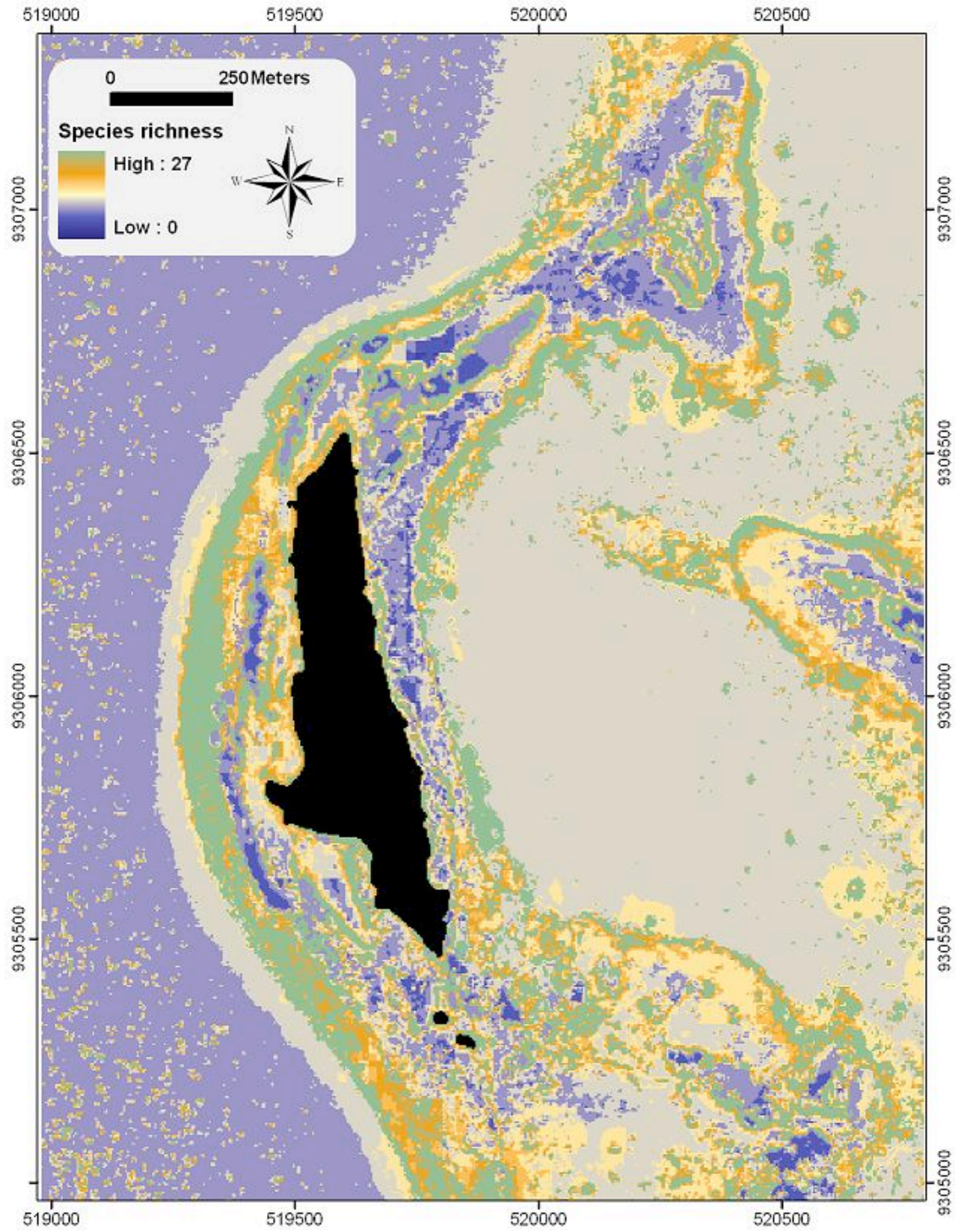

Figure 7.9: Map showing predicted fish species richness around Chumbe Island. Calculations done with Bagging model and IKONOS data set.

7.4.1 Predictive performance of IKONOS-based models

For all three response variables, the predictive models using IKONOS data performed significantly worse than the ones using *in situ* data. Using the "average-predictor" as a benchmark, they reduced prediction errors by only 47%, 35%, and 34% when compared to the models using *in situ* data, for species richness, biomass, and diversity, respectively. A comparison with results from other studies is shown in Table 7.2, but the comparison is not straight-forward because all other studies in the table report coefficients of determination calculated on all or part of the data set used to train the predictive model, and use a variety of models. As such, studies that fit models closely to the training data will report higher correlations than studies that use more parsimonious model types. For the sake of comparison, R^2 values from this study, calculated on the training set, have therefore been provided in Table 7.2 in parentheses, and the model type employed by each study indicated. The difference between results achieved with different model types is also illustrated in Table 7.3, where the coefficients of determination of values of this study (based on IKONOS data) are reported as calculated on both training and test data sets. Further complicating a comparison, the studies compared in Table 7.2 use a variety of remotely sensed data, including bathymetric lidar and a coastal relief model derived from a combination of data sources for depth estimation.

Table 7.2: Comparison of predictive performance achieved by recent studies predicting fish community variables from remotely sensed data. Coefficients of determination for this study are provided along with values (in parentheses) obtained from testing accuracy on the training set. * Wedding et al. (2008) report their results with Spearman rank correlations (R) – these values have been squared here for a rough comparison with the coefficients of determination from other studies.

Study	Response variable	Coefficient of determination	Model type	Data type
Our study	Spp. richness	0.30 (0.74)	Bagging	IKONOS
Pittman et al. (2009)	Spp. richness	0.64	Boosted trees	Lidar
Pittman et al. (2007)	Spp. richness	0.48-0.56	Regression Tree	TIN
Wedding et al. (2008)*	Spp. richness	0.44	LM	Lidar
Our study	Biomass	0.25 (0.64)	Bagging	IKONOS
Pittman et al. (2009)	Biomass	0.46	Boosted trees	Lidar
Wedding et al. (2008)*	Biomass	0.42	LM	Lidar
Our study	Diversity	0.23 (0.69)	Bagging	IKONOS
Wedding et al. (2008)*	Diversity	0.17	LM	Lidar

Table 7.3: Coefficients of determination for IKONOS-based models of all three fish community variables. Calculations based on test data outside parentheses, calculations based on training data in parentheses.

	Species richness	Biomass	Diversity
LM	**0.23 (0.54)**	**0.24 (0.50)**	**0.16 (0.46)**
GAM	**0.32 (0.50)**	**0.26 (0.47)**	**0.20 (0.43)**
Bagging	**0.30 (0.74)**	**0.25 (0.64)**	**0.23 (0.69)**
Random forest	**0.30 (0.93)**	**0.23 (0.91)**	**0.21 (0.93)**
Boosted trees	**0.29 (0.60)**	**0.26 (0.66)**	**0.19 (0.50)**
SVM	**0.24 (0.50)**	**0.18 (0.45)**	**0.20 (0.49)**

From this variety of data sources, model types and results reporting, the one discernable pattern is that higher coefficients of determination are obtained for species richness than for the other response variables. Reasons for this pattern are speculative, but possibly related to a lower variability in field observations of species richness, which is less sensitive to the passing of schools of fish during data collection. More comparable studies are needed to better establish the predictions errors and coefficients of determination that can be expected from spatial predictive mapping of fish communities, but the studies reviewed in Table 7.2, representing reef environments in the Indian Ocean, the Pacific, and the Caribbean, suggest that prediction is indeed possible, across regions. Furthermore, the comparison suggests that, although coefficients of determination are small (0.23 – 0.30) when tested on test data as in this study, IKONOS data can provide predictions that are comparable to the more expensive lidar data. This is encouraging for practical application of spatial predictive models.

7.4.2 Variable importance in IKONOS-based models

7.4.2.1 Rugosity

Regardless of response variable, the most important habitat variable in the Bagging model was rugosity at the finest scale obtainable with IKONOS imagery (r=6 m). This supports the dominant influence of structural complexity in shaping the fish communities, which is also shown by the high importance of the 'depth range' variable for all models based on *in situ* data. Although the 'substrate class' variable also has substantial importance for biomass predictions (along with the 'depth' variable itself), this indicates that a precise derivation of depth, used for rugosity calculations, is crucial for IKONOS-based predictions of any of the response variables. For example, the noise seen in the south-west corner of Figure 7.9 is directly caused by image noise that resulted in unreasonably high rugosity estimates for some pixels.

7.4.2.2 Live coral cover

The importance of live coral cover was low for predictions of species richness, and near zero or even slightly negative for predictions of biomass and diversity (although live coral cover at some radius was the most important variable in both LM and GAM models of diversity). Given its importance in models based on *in situ* data, where it was the second-most important variable for all response variables, the reduced importance when estimated from the IKONOS data must be attributed to the lower precision with which it can be estimated remotely as compared to the other (competing) variables. Despite the correlations obtained between the depth-invariant index and live coral cover in this study, predictions performed only marginally better than an average-predictor, and correlations between the response variables and the remotely sensed live coral cover variable were substantially lower than those obtained with live coral cover measured *in situ* (Table 5.4).

7.4.2.3 Habitat diversity

Using IKONOS data and the Bagging model, habitat diversity showed very modest importance for species richness, but substantial importance for both biomass and diversity, where habitat diversity at r=10 m was the second most important variable. This variation in the importance of habitat diversity for the different response variables mirrors the variation in correlations between habitat diversity and each response variable individually (see Figure 6.9). It is interesting to note that although these individual correlations are lower for habitat diversity (Figure 6.9) than they are for live coral cover (Figure 6.6), the importance of the habitat diversity variable in the Bagging models is higher than that of live coral cover. This is most likely due to the two variables' different collinearity with rugosity. The highest values of live coral cover on the two reefs is found near the reef edges, where rugosity is also high. The collinearity of rugosity (r=6 m) and live coral cover (r=2 m) (Spearman R=0.27) therefore reduces the importance of the latter in a multi-variable model. Habitat diversity (r=10 m), on the other hand, is higher away from the reef edge (see Figure 6.10 and Figure 6.11), though these display habitat richness the spatial pattern of habitat diversity can be assumed to be similar). This variable therefore suffers less from colinearity with rugosity (r=6 m) (Spearman R=-0.06), and provides more complementary.

7.4.3 Complementarity of remotely sensed information

The low importance of all remotely sensed variables for models based on the "All" data set (except for the LM model) is an indication that all three response variables are influenced mainly by their immediate environment and that larger-scale habitat effects, at least those investigated in this study, have limited influence. The few studies available so far have not converged on a characteristic scale at which the response variable used in this study respond to their habitat (see Table 6.4 and Table 6.5), but the differences observed in correlations between predictor and response variable with changing scale (see Figure 6.6, Figure 6.7 and Figure 6.9) support the conclusion that the most immediate environment influences the fish community most strongly, which reduces the role of remote sensing to that of providing spatial coverage of predictions, not new insight into the effect of habitat on structuring fish communities. However, specific influences of habitat on the fish community, over large spatial scales, have been documented in studies designed to investigate the importance of seagrasses and mangroves, (see section 6.4.6). In order to improve predictive models such influences should be specifically tested for, and incorporated in predictive models as variables, along other habitat variables observed at an increasing range of spatial scales.

7.5 Summary

Sub-question C was "*How does remote sensing compare to traditional fieldwork for mapping a coral reef fish community?*". Part of the answer to this question follows directly from the fact that the coverage of field data is limited to the sites at which the fieldwork has been carried out, whereas remote sensing provides a spatially continuous coverage of data over large areas. Because of the influence the habitat has on the fish community, and because of the heterogeneous spatial patterns in which coral reef habitats exist, the use of interpolation approaches to create spatially distributed predictions of fish community variables from field data is unlikely to be successful. A short answer to sub-question C therefore is that traditional fieldwork cannot produce data from which maps of the fish community can be made, whereas remote sensing can.

The more relevant question, therefore, is how the performance of remote sensing's spatially explicit predictive models of the fish community compare to those that can be produced from *in situ* data. The results shown in this chapter clearly indicate that the inaccuracy introduced when estimating habitat variables with IKONOS data has led to reduced predictive performance of models based on IKONOS data, compared to those based on *in situ* data. For all predictive models based on IKONOS data, rugosity at the finest spatial scale (6 m radius) was identified as the most important predictor. But this variable only approximates the 'depth range' variable available to *in situ*-based models, which was identified as the most important in all tree-based models based on the "All" data set. This approximation leads to higher RMSE values in the models based on IKONOS data. As a quantitative measure of relative model performance, we have used the reduction in RMSE value that a model achieves when compared to an average-predictor. Using that measure, models based on IKONOS data achieve only 47% of the performance that a model based on *in situ* data does. The same values for fish biomass and diversity are 35% and 34%, respectively. Models based on simulated Landsat data perform even worse. Despite these low numbers, models based on IKONOS data can still produce credible maps of the spatial distribution of fish community variables, as seen in the map of fish species richness around Chumbe Island (Figure 7.9). Results from other studies suggest that this is also the case for reef environments in other parts of the world, and when using airborne lidar data instead of IKONOS images. The last question, whether or not such maps are sufficiently accurate and precise to be useful in a management context, is dealt with in the following, and last, chapter.

CHAPTER 8: CONCLUSION

In this thesis, we have presented a series of steps for the development of spatially explicit predictive models that use information about the coral reef habitat to predict three variables describing the fish community: species richness, biomass, and diversity. We have tested the predictive performance of six model types that include both traditional parametric and new non-parametric machine-learning approaches, and we have used a permutation-based approach to quantify the contribution of each variable to the model predictions, thereby both identifying the habitat variables that are most important for predicting the fish community variables, and illustrating the profound influence of model structure on variable importance. We have investigated the importance of the spatial scale of the habitat data used for model development, and ultimately produced a map predicting fish species richness in the nearshore environment around Chumbe Island, Zanzibar. The thesis thus forms a comprehensive answer to the research question posed in chapter 1:

How can remote sensing be used to map coral reef fish communities?

The research conducted to answer this question, described in the preceding chapters, has led to several results that contribute to the body of knowledge in the fields of geography, remote sensing in particular, and coral reef studies. The most important contributions to these areas are summarised below:

- The complex ecological relationships between fish communities and their habitat require equally complex approaches for their modeling. Assumptions of linearity and additive effects do not hold true, and the use of simplistic model types such as multiple linear regression lead to unnecessarily poor predictions of fish community variables. Of the model types tested in our study, tree-based ensemble techniques generally outperform others, and their adoption in the ecological modeling community is therefore likely to improve predictive models of both coral reef fish communities and other dependent variables with complex relationships to their predictors.

- The importance of a habitat variable in a multi-variable predictive model is dependent on a number of factors. These include the frequency distribution and quantification (binary, discrete, or continuous) of the variable, collinearity with other variables available for model development, and model type. Interpretation of variable importance is therefore far from straight-forward, and uncritical interpretation of model outputs can lead to unqualified and misleading conclusions. This is particularly important because the derivation of important variables from multi-variable predictive models is used by the conservation community to target "important" features of the coral reef habitat for protection. Another application of variable importance measures is to "correct for habitat influences" in order to identify the importance of other factors such as management regimes (e.g. protection). Such corrections are likely to lead to misleading conclusions unless the numerous factors that influence variable importance are properly accounted for.

- Keeping these limitations in mind, structural complexity stands out as an aspect of habitat that has a large influence on the fish community. For all fish community variables and all model types other than the LM, 'depth range' was the most important variable in models based on in situ data, and similarly rugosity, calculated at the smallest spatial scale, was the most important variable in models based on IKONOS data. This not only confirms numerous field studies that show similar results, but also show that IKONOS data are able, albeit imperfectly, to estimate the structural complexity that influences the fish community.

- Conservation status, whether a site is located inside or outside an MPA, has negligible importance for fish species richness and diversity, whereas its importance for fish biomass is significant. This conclusion is surprising, and contradicted both by the personal experience of the author and by species lists compiled for the two reefs studied. Its validity may be limited to the measure of alpha-diversity (and richness) and the small field sites used in this study, and merits further attention.

- The fish community is best predicted (and hence considered mainly influenced) by habitat at the local scale, although fish biomass is also influenced by proximity to the reef edge which is best quantified at larger spatial scales.

- It is possible to derive estimates of water depth, structural complexity and live coral cover from IKONOS data. Although these estimates do not correlate as well with the fish community variables as their *in situ* counterparts do, they all contribute to IKONOS-based predictions of at least one of these variables.

- In addition, measures of habitat diversity can also be derived from substrate classifications based on IKONOS data. Although not all of these measures are significantly correlated with the fish community variables themselves, they nevertheless contribute significantly to IKONOS-based predictions of all the fish community variables.

- Spatially explicit predictive models of fish species richness, biomass, and diversity, can be produced from IKONOS data. Although their predictive performance is limited when compared to models using *in situ* data, the resulting maps produce reasonable predictions, and are likely to be useful for management.

These results are encouraging, not only because they individually contribute to the body of knowledge in geography, remote sensing and coral reef studies, but because together they illustrate how maps of the fish community can be produced with IKONOS data (and how accurate such maps can be). They can thus be expected to accelerate the adoption of remote sensing data in spatial ecology, including spatially explicit predictive modeling. However, the research presented in this thesis also points to areas that merit further attention. The importance of structural complexity suggests that the use of sensors that enable precise mapping of depth at sufficiently fine spatial scales (e.g. airborne lidar) can improve predictive performance, but a direct comparison between prediction errors from models based on IKONOS and lidar data has yet to be made. Similarly, the importance of scale for remotely sensed habitat variables suggests that improved spatial resolution (e.g. from Geoeye-1 launched in 2008 or Worldview-2 to be launched October 6, 2009, both with <2 m spatial resolution) may improve predictive performance. In addition, the development of habitat diversity measures that better quantify diversity as it is relevant for structuring the fish community are likely to lead to improvements. Better incorporation of the landscape ecology approach, identifying areas that provide habitat for juveniles, feeding grounds, spawning aggregation sites etc., also has the potential to improve spatial predictive models.

A different approach, predicting presence/absence of individual species, also holds great potential because of its immediate applicability to conservation of endangered/keystone/icon species. A species-based approach is likely to lead to greater ecological insight, as individual species respond to different aspects of the habitat in different ways, and at different spatial scales. Important habitat variables, and the characteristic scales of response to them, are therefore more easily linked to their causal mechanisms (such as daily migration distance, critical habitat for a particular life stage or activity) when studied at the species level. The maximum distance and the minimum spatial extent at which seagrasses and mangroves influences reef fish also needs to be more clearly identified, and differentiated by fish species. The species-based approach to spatial predictive modeling, mainly employed as species distribution modeling, is already relatively mature in terrestrial environments, often employed to create scenarios of future change in the distribution of a specific species with projected climate change. This approach has yet to be applied to coral reef environments, and field studies of fish-habitat relationships have mostly focused on aggregate measures such as those used in our study, or on functional groups based on diet. However, the importance of individual species for ecosystem function (e.g. parrotfishes as dominant grazers on Caribbean reefs and triggerfishes as the last predators of sea urchins in East Africa) is becoming increasingly clear, and predictions of the change in their distributions, with climate change or with other human impacts, will be important for conservation management. For example, the poleward movement of coral species distributions has already been documented on the Great Barrier Reef, and similar changes are likely to be happening, unnoticed and undocumented, on most reefs around the world. An improved understanding of how such changes in habitat, caused by direct or indirect human impacts, influences the distributions of species will be an important contribution of spatial predictive modeling, along with any practical application of the maps that it can produce.

Ultimately it remains to be seen whether the predictive performance of models based on IKONOS data is sufficient for the approach to be widely adopted in coral reef management efforts. Given the numerous potential avenues of improvement discussed above, it is encouraging that the approach has already been adopted by the Wildlife Conservation Society for their ecosystem-based coral reef conservation project in Fiji. Part of this project involves the further development of an existing network of MPAs, a task for which the spatial distribution of fish community variables will provide important input. The application of the research approach in Fiji will also provide an evaluation of its performance in another environment, an environment with more than 500 species of reef fish spread over almost 1000 km^2, and with a complex pattern of fishing effort and a third form of conservation status – seasonal fishing closures.

In conclusion, there is potential for improvements in both the sources of remote sensing data, and the derivation of habitat variables from such data, to reduce errors in spatially explicit predictive models, but even at the current level of predictive performance IKONOS-based maps of fish community variables are sufficiently accurate to be useful for coral reef management efforts.

BIBLIOGRAPHY

Adjeroud, M., Andréfouët, S., Payri, C., & Orempuller, J. (2000). Physical factors of differentiation in macrobenthic communities between atoll lagoons in the Central Tuamotu Archipelago (French Polynesia). *Marine Ecology-Progress Series, 196*, 25-38

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

Andréfouët, S., Claereboudt, M., Matsakis, P., Pages, J., & Dufour, P. (2001). Typology of atoll rims in Tuamotu Archipelago (French Polynesia) at landscape scale using SPOT HRV images. *International Journal of Remote Sensing, 22*, 987-1004

Andréfouët, S., & Guzman, H.M. (2005). Coral reef distribution, status and geomorphology-biodiversity relationship in Kuna Yala (San Blas) archipelago, Caribbean Panama. *Coral Reefs, 24*, 31-42

Andréfouët, S., Kramer, P., Torres-Pulliza, D., Joyce, K.E., Hochberg, E.J., Garza-Perez, R., Mumby, P.J., Riegl, B., Yamano, H., White, W.H., Zubia, M., Brock, J.C., Phinn, S.R., Naseer, A., Hatcher, B.G., & Muller-Karger, F.E. (2003). Multi-site evaluation of IKONOS data for classification of tropical coral reef environments. *Remote Sensing of Environment, 88*, 128-143

Beger, M., Jones, G.P., & Munday, P.L. (2003). Conservation of coral reef biodiversity: a comparison of reserve selection procedures for corals and fishes. *Biological Conservation, 111*, 53-62

Bellwood, D.R., & Hughes, T.P. (2001). Regional-scale assembly rules and biodiversity of coral reefs. *Science, 292*, 1532-1534

Bellwood, D.R., Wainwright, P.C. (2002). The History and Biogeography of Fishes on Coral Reefs. In P.F. Sale (Ed.), *Coral Reef Fishes: Dynamics and Diversity in a Complex Ecosystem* (p. 549). San Diego: Academic Press

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate - a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 57*, 289-300

Bohnsack, J.A., & Bannerot, S.P. (1986). A stationary visual census technique for quantitatively assessing community structure of coral reef fishes. NOAA, Technical Report NMFS 41, pp. 17

Brainerd, T. (1994). Socioeconomic research on fisheries and aquaculture in Africa. In A. Charles, T. Brainerd, A. Bermudez, H. Montalvo & R. Pomeroy (Eds.), *Fisheries Socioeconomics in the Developing World: Regional Assessments and an Annotated Bibliography* (pp. 12-37). Ottawa: IDRC

Brando, V.E., Anstee, J.M., Wettle, M., Dekker, A.G., Phinn, S.R., & Roelfsema, C. (2009). A physics based retrieval and quality assessment of bathymetry from suboptimal hyperspectral data. *Remote Sensing of Environment, 113*, 755-770

Breiman, L. (1996). Bagging predictors. *Machine Learning, 24*, 123-140

Breiman, L. (2001a). Random forests. *Machine Learning, 45*, 5-32

Breiman, L. (2001b). Statistical Modeling: The Two Cultures. *Statistical Science, 16*, 199-215

Breiman, L., Friedman, J.H., Olshen, R.A., & Stone, C.G. (1984). *Classification and Regression Trees*. Belmont, USA: Wadsworth International Group,

Brenning, A. (2009). Benchmarking classifiers to optimally integrate analysis and multispectral remote sensing in automatic rock glacier detection. *Remote Sensing of Environment, in press.*

Briggs, J.C. (1999). Coincident biogeographic patterns: Indo-West Pacific Ocean. *Evolution, 53*, 326-335

Brock, J.C., Wright, C.W., Clayton, T.D., & Nayegandhi, A. (2004). LIDAR optical rugosity of coral reefs in Biscayne National Park, Florida. *Coral Reefs, 23*, 48-59

Bruno, J.F., & Selig, E.R. (2007). Regional Decline of Coral Cover in the Indo-Pacific: Timing, Extent, and Subregional Comparisons. *PLoS ONE, 2*, e711

Buja, K. (2008). Diversity Calculator, ArcMap 9.2 extension. http://arcscripts.esri.com/details.asp?dbid=15258

Capolsini, P., Andréfouët, S., Rion, C., & Payri, C. (2003). A comparison of Landsat ETM+, SPOT HRV, Ikonos, ASTER, and airborne MASTER data for coral reef habitat mapping in south pacific islands. *Canadian Journal of Remote Sensing, 29*, 187-200

Carr, M.H., & Reed, D.C. (1993). Conceptual issues relevant to marine harvest refuges - examples from temperate reef fishes. *Canadian Journal of Fisheries and Aquatic Sciences, 50*, 2019-2028

Cesar, H., Burke, L., & Pet-Soede, L. (2003). The economics of worldwide coral reef degradation. Cesar Environmental Economics Consulting, pp. 23

Chang, C.C., & Lin, C.J. (2009). LIBSVM - A library for support vector machines, R package library.

Chapman, M.R., & Kramer, D.L. (1999). Gradients in coral reef fish density and size across the Barbados Marine Reserve boundary: effects of reserve protection and habitat characteristics. *Marine Ecology-Progress Series, 181*, 81-96

Congalton, R. (1991). A review of assessing the accuracy of classifications of remotely sensed data. *Remote Sensing of Environment, 37*, 35-46

Connell, J.H. (1978). Diversity in tropical rain forests and coral reefs - high diversity of trees and corals is maintained only in a non-equilibrium state. *Science, 199*, 1302-1310

Connell, J.H., Hughes, T.P., & Wallace, C.C. (1997). A 30-year study of coral abundance, recruitment, and disturbance at several scales in space and time. *Ecological Monographs, 67*, 461-488

De'ath, G. (2007). Boosted trees for ecological modeling and prediction. *Ecology, 81*, 3178-3192

De'ath, G., & Fabricius, K. (2000). Classification and regression trees: A powerful yet simple technique for ecological data analysis. *Ecology, 81*, 3178-3192

Dickman, M. (1968). Some indices of diversity. *Ecology, 49*, 1191-&

Dorenbosch, M., van Riel, M.C., Nagelkerken, I., & van der Velde, G. (2004). The relationship of reef fish densities to the proximity of mangrove and seagrass nurseries. *Estuarine Coastal and Shelf Science, 60*, 37-48

Dormann, C.F., McPherson, J.M., Araujo, M.B., Bivand, R., Bolliger, J., Carl, G., Davies, R.G., Hirzel, A., Jetz, W., Kissling, W.D., Kuhn, I., Ohlemuller, R., Peres-Neto, P.R., Reineking, B., Schroder, B., Schurr, F.M., & Wilson, R. (2007). Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. *Ecography, 30*, 609-628

EcoAfrica (2005a). Mnemba Island and Chwaka Bay Conservation Areas: A preliminary Situational Assessment. Department of Fisheries, Zanzibar, pp. 58

EcoAfrica (2005b). Rapid Assessment of the Menai Bay Conservation Area (MBCA). pp. 107

Efron, B., & Gong, G. (1983). A leisurely look at the bootstrap, the jackknife, and cross-validation. *The American Statistician, 37*, 36-48

Efron, B., & Tibshirani, R. (1993). *An introduction to the bootstrap*. New York: Chapman & Hall, pp. 225

Elith, J., Graham, C.H., Anderson, R.P., Dudik, M., Ferrier, S., Guisan, A., Hijmans, R.J., Huettmann, F., Leathwick, J.R., Lehmann, A., Li, J., Lohmann, L.G., Loiselle, B.A., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., Overton, J.M., Peterson, A.T., Phillips, S.J., Richardson, K., Scachetti-Pereira, R., Schapire, R.E., Soberon, J., Williams, S., Wisz, M.S., & Zimmermann, N.E. (2006). Novel methods improve prediction of species' distributions from occurrence data. *Ecography, 29*, 129-151

Elith, J., Leathwick, J.R., & Hastie, T. (2008). A working guide to boosted regression trees. *Journal of Animal Ecology, 77*, 802-813

Feygels, V.I., Kopilevich, Y.I., Surkov, A., Yungel, J.K., & Behrenfeld, M.J. (2003). Airborne lidar system with variable field-of-view receiver for water optical properties measurement. *SPIE, 5155*, 12-21

Finkl, C.W., Benedet, L., & Andrews, J.L. (2005). Interpretation of seabed geomorphology based on spatial analysis of high-density airborne laser bathymetry. *Journal of Coastal Research, 21*, 501-514

Friedlander, A.M., & Parrish, J.D. (1998). Habitat characteristics affecting fish assemblages on a Hawaiian coral reef. *Journal of Experimental Marine Biology and Ecology, 224*, 1-30

Friedman, J.H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics, 29*, 1189-1232

Froese, R., & Pauly, D. (2008). Fishbase. *2008.* www.fishbase.org

Garpe, K.C., & Ohman, M.C. (2003). Coral and fish distribution patterns in Mafia Island Marine Park, Tanzania: fish-habitat interactions. *Hydrobiologia, 498*, 191-211

Garpe, K.C., Yahya, S.A.S., Lindahl, U., & Ohman, M.C. (2006). Long-term effects of the 1998 coral bleaching event on reef fish assemblages. *Marine Ecology-Progress Series, 315*, 237-247

Gladstone, W. (2002). The potential value of indicator groups in the selection of marine reserves. *Biological Conservation, 104*, 211-220

Goodman, J.A., & Ustin, S.L. (2007). Classification of benthic composition in a coral reef environment using spectral unmixing. *Journal of Applied Remote Sensing, 1*

Goreau, T.J., & Hayes, R.L. (1994). Coral Bleaching and Ocean Hot-Spots. *Ambio, 23*, 176-180

Graham, M.H. (2003). Confronting multicollinearity in ecological multiple regression. *Ecology, 84*, 2809-2815

Graham, N.A.J., Wilson, S.K., Jennings, S., Polunin, N.V.C., Bijoux, J.P., & Robinson, J. (2006a). Dynamic fragility of oceanic coral reef ecosystems. *Proceedings of the National Academy of Sciences of the United States of America, 103*, 8425-8429

Graham, N.A.J., Wilson, S.K., Jennings, S., Polunin, N.V.C., Bijoux, J.P., & Robinson, J. (2006b). Dynamic fragility of oceanic coral reef ecosystems. *Proceedings of the National Academy of Sciences, USA, 103*, 8425-8429

Green, E.P., Mumby, P.J., Edwards, A.J., & Clark, C.D. (2000). *Remote Sensing Handbook for Tropical Coastal Management*. Paris: UNESCO, pp. 316

Grober-Dunsmore, R., Frazer, T.K., Beets, J.P., Lindberg, W.J., Zwick, P., & Funicelli, N.A. (2008). Influence of landscape structure, on reef fish assemblages. *Landscape Ecology, 23*, 37-53

Grober-Dunsmore, R., Frazer, T.K., Lindberg, W.J., & Beets, J.P. (2007). Reef fish and habitat relationships in a Caribbean seascape: the importance of reef context. *Coral Reefs, 26*, 201-216

Grootenhuis, F., & Lopez, J. (2003). Household economy analysis for Zanzibar. Report to the Revolutionary Government of Zanzibar and Save the Children Tanzania. Save the Children Tanzania,

Guisan, A., Zimmermann, N.E., Elith, J., Graham, C.H., Phillips, S., & Peterson, A.T. (2007). What matters for predicting the occurrences of trees: Techniques, data, or species' characteristics? *Ecological Monographs, 77*, 615-630

Halpern, B.S. (2003). The impact of marine reserves: do reserves work and does reserve size matter? *Ecological Applications, 13*, S117-S137

Harborne, A.R., Mumby, P.J., Zychaluk, K., Hedley, J.D., & Blackwell, P.G. (2006). Modeling the beta diversity of coral reefs. *Ecology, 87*, 2871-2881

Hastie, T. (2008). gam: Generalized Additive Models., R package.

Hastie, T., Tibshirani, R., & Friedman, J.H. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* New York: Springer-Verlag, pp. 763

Hedley, J.D., Harborne, A.R., & Mumby, P.J. (2005). Simple and robust removal of sun glint for mapping shallow-water benthos. *International Journal of Remote Sensing, 26*, 2107-2112

Hedley, J.D., & Mumby, P.J. (2003). A remote sensing method for resolving depth and subpixel composition of aquatic benthos. *Limnology and Oceanography, 48*, 480-488

Hedley, J.D., Mumby, P.J., Joyce, K.E., & Phinn, S.R. (2004). Spectral unmixing of coral reef benthos under ideal conditions. *Coral Reefs, 23*, 60-73

Hochberg, E.J., Andréfouët, S., & Tyler, M.R. (2003a). Sea surface correction of high spatial resolution Ikonos images to improve bottom mapping in near-shore environments. *Ieee Transactions on Geoscience and Remote Sensing, 41*, 1724-1729

Hochberg, E.J., & Atkinson, M.J. (2003). Capabilities of remote sensors to classify coral, algae, and sand as pure and mixed spectra. *Remote Sensing of Environment, 85*, 174-189

Hochberg, E.J., Atkinson, M.J., & Andréfouët, S. (2003b). Spectral reflectance of coral reef bottom-types worldwide and implications for coral reef remote sensing. *Remote Sensing of Environment, 85*, 159-173

Hoegh-Guldberg, O. (1999). Climate change, coral bleaching and the future of the world's coral reefs. *Marine and Freshwater Research, 50*, 839-866

Holden, H., & LeDrew, E. (1999). Hyperspectral identification of coral reef features. *International Journal of Remote Sensing, 20*, 2545-2563

Holland, J.D., Bert, D.G., & Fahrig, L. (2004). Determining the spatial scale of species' response to habitat. *Bioscience, 54*, 227-233

Horrill, J.C., Kamukuru, A., Mgaya, Y.D., & Risk, M. (2000). Northern Tanzania, Zanzibar and Pemba. In T.R. McClanahan, C.R.C. Sheppard & D.O. Obura (Eds.), *Coral Reefs of the Indian Ocean* (pp. 167-198). Oxford: Oxford University Press

Hubbard, D.K. (1997). Reefs as Dynamic Systems. In C. Birkeland (Ed.), *Life and Death of Coral Reefs* (pp. 43-67). New York: Chapman and Hall

Hughes, T.P. (1994). Catastrophes, Phase-Shifts, and Large-Scale Degradation of a Caribbean Coral-Reef. *Science, 265*, 1547-1551

Hughes, T.P., Baird, A.H., Bellwood, D.R., Card, M., Connolly, S.R., Folke, C., Grosberg, R., Hoegh-Guldberg, O., Jackson, J.B.C., Kleypas, J., Lough, J.M., Marshall, P., Nystrom, M., Palumbi, S.R., Pandolfi, J.M., Rosen, B., & Roughgarden, J. (2003). Climate change, human impacts, and the resilience of coral reefs. *Science, 301*, 929-933

Huston, M.A. (1994). *Biological diversity: The coexistence of species on changing landscapes*. Cambridge: Cambridge University Press, pp. 681

Isoun, E., Fletcher, C., Frazer, N., & Gradie, J. (2003). Multi-spectral mapping of reef bathymetry and coral cover; Kailua Bay, Hawaii. *Coral Reefs, 22*, 68-82

ITT Visual Information Solutions (2007). ENVI 4.4.

Jackson, J.B.C. (2008). Ecological extinction and evolution in the brave new ocean. *Proceedings of the National Academy of Sciences of the United States of America, 105*, 11458-11465

Jackson, J.B.C., Kirby, M.X., Berger, W.H., Bjorndal, K.A., Botsford, L.W., Bourque, B.J., Bradbury, R.H., Cooke, R., Erlandson, J., Estes, J.A., Hughes, T.P., Kidwell, S., Lange, C.B., Lenihan, H.S., Pandolfi, J.M., Peterson, C.H., Steneck, R.S., Tegner, M.J., & Warner, R.R. (2001). Historical overfishing and the recent collapse of coastal ecosystems. *Science, 293*, 629-638

Jenness, J. (2002). Surface Areas and Ratios from Elevation Grid (surfgrids.avx), Extension for ArcView 3.x. http://jennessent.com/arcview/surface_areas.htm

Jiddawi, N.S., & Khatib (2007).

Jiddawi, N.S., & Ohman, M.C. (2002). Marine fisheries in Tanzania. *Ambio, 31*, 518-527

Jones, G.P., McCormick, M.I., Srinivasan, M., & Eagle, J.V. (2004a). Coral decline threatens fish biodiversity in marine reserves. *Proceedings of the National Academy of Sciences of the United States of America, 101*, 8251-8253

Jones, G.P., McCormick, M.I., Srinivasan, M., & Eagle, J.V. (2004b). Coral decline threatens fish biodiversity in marine reserves. *Proceedings of the National Academy of Sciences, USA, 101*, 8251-8253

Jones, G.P., & Syms, C. (1998). Disturbance, habitat structure and the ecology of fishes on coral reefs. *Australian Journal of Ecology, 23*, 287-297

Joyce, K. (2004a). A method for mapping live coral cover using remote sensing. *School of Geography, Doctor of Philosophy*, 137

Joyce, K. (2004b). A method for mapping live coral cover using remote sensing. *School of Geography, Planning and Architecture, Doctor of Philosophy*, 137

Joyce, K.E., & Phinn, S.R. (2002). Bi-directional reflectance of corals. *International Journal of Remote Sensing, 23*, 389-394

Joyce, K.E., Phinn, S.R., Scarth, P.F., & Roelfsema, C.M. (2003). A method for determining live coral cover using remote sensing. *International symposium for remote sensing of the environment*

Kendall, M.S., Buja, K.R., Christensen, J.D., Kruer, C.R., & Monaco, M.E. (2004). The seascape approach to coral ecosystem mapping: An integral component of understanding the habitat utilization patterns of reef fish. *Bulletin of Marine Science, 75*, 225-237

Kendall, M.S., & Miller, T. (2008). The influence of thematic and spatial resolution on maps of a coral reef ecosystem. *Marine Geodesy, 31*, 75-102

Kleypas, J.A., Buddemeier, R.W., Archer, D., Gattuso, J.P., Langdon, C., & Opdyke, B.N. (1999). Geochemical consequences of increased atmospheric carbon dioxide on coral reefs. *Science, 284*, 118-120

Klonowski, W.M., Fearns, P., & Lynch, M.J. (2007). Retrieving key benthic cover types and bathymetry from hyperspectral imagery. *Journal of Applied Remote Sensing, 1*

Knowlton, N. (2001). The future of coral reefs. *Proceedings of the National Academy of Sciences of the United States of America, 98*, 5419-5425

Knudby, A., & LeDrew, E. (2007). Measuring Structural Complexity on Coral Reefs. *AAUS Annual Symposium*

Knudby, A., LeDrew, E., & Newman, C. (2007). Progress in the use of remote sensing for coral reef biodiversity studies. *Progress in Physical Geography, 31*, 421-434

Knudby, A., Newman, C., & LeDrew, E. (2008). Remote sensing for studies of the spatial distribution of coral reef fishes. *International Coral Reef Symposium*

Kohler, K.E., & Gill, S.M. (2006). Coral Point Count with Excel extensions (CPCe): A Visual Basic program for the determination of coral and substrate coverage using random point count methodology. *Computers & Geosciences, 32*, 1259-1269

Kohn, A.J., & Leviten, P.J. (1976). Effect of Habitat Complexity on Population-Density and Species Richness in Tropical Intertidal Predatory Gastropod Assemblages. *Oecologia, 25*, 199-210

Kuffner, I.B., Brock, J.C., Grober-Dunsmore, R., Bonito, V.E., Hickey, T.D., & Wright, C.W. (2007). Relationships between reef fish communities and remotely sensed rugosity measurements in Biscayne National Park, Florida, USA. *Environmental Biology of Fishes, 78*, 71-82

Kutser, T., Dekker, A.G., & Skirving, W. (2003). Modeling spectral discrimination of Great Barrier Reef benthic communities by remote sensing instruments. *Limnology and Oceanography, 48*, 497-510

Lanshammar, F. (2004). Efficiency of Measures Introduced at Chumbe Island Coral Park (CHICOP), with Regard to Fish Communities, Zanzibar, Tanzania. *Department of Zoo Ecology, M.Sc.*, 27

Lapointe, B.E. (1997). Nutrient thresholds for bottom-up control of macroalgal blooms on coral reefs in Jamaica and southeast Florida. *Limnology and Oceanography, 42*, 1119-1131

Leathwick, J.R., Elith, J., Francis, M.P., Hastie, T., & Taylor, P. (2006). Variation in demersal fish species richness in the oceans surrounding New Zealand: an analysis using boosted regression trees. *Marine Ecology Progress Series, 321*, 267-281

Lee, Z., Carder, K.L., Chen, R.F., & Peacock, T.G. (2001). Properties of the water column and bottom derived from Airborne Visible Infrared Imaging Spectrometer (AVIRIS) data. *Journal of Geophysical Research-Oceans, 106*, 11639-11651

Lee, Z.P., Carder, K.L., Mobley, C.D., Steward, R.G., & Patch, J.S. (1999). Hyperspectral remote sensing for shallow waters: 2. Deriving bottom depths and water properties by optimization. *Applied Optics, 38*, 3831-3843

Liaw, A., & Wiener, M. (2002). randomForest: Classification and Regression with Random Forest, R package.

Lieske, E., & Myers, R. (2001). *Corel Reef Fishes*. Princeton, USA: Princeton University Press, pp. 420

Lillesand, T.M., & Kiefer, R.W. (1994). *Remote Sensing and Image Interpretation*. New York, USA: John Wiley & Sons Inc.,

Lindahl, U., Ohman, M.C., & Schelten, C.K. (2001). The 1997/1998 mass mortality of corals: Effects on fish communities on a Tanzanian coral reef. *Marine Pollution Bulletin, 42*, 127-131

Luckhurst, B.E., & Luckhurst, K. (1978). Analysis of influence of substrate variables on coral-reef fish communities. *Marine Biology, 49*, 317-323

Ludwig, J.A., & Reynolds, J.F. (1988). *Statistical Ecology*. New York: John Wiley and Sons, pp. 337

Lyzenga, D.R. (1978). Passive Remote-Sensing Techniques for Mapping Water Depth and Bottom Features. *Applied Optics, 17*, 379-383

Lyzenga, D.R. (1981). Remote sensing of bottom reflectance and water attenuation parameters in shallow water using aircraft and Landsat data. *International Journal of Remote Sensing, 2*, 71-82

Lyzenga, D.R., Malinas, N.R., & Tanis, F.J. (2006). Multispectral bathymetry using a simple physically based algorithm. *Ieee Transactions on Geoscience and Remote Sensing, 44*, 2251-2259

McClanahan, T.R. (1988). Coexistence in a Sea-Urchin Guild and Its Implications to Coral-Reef Diversity and Degradation. *Oecologia, 77*, 210-218

McCormick, M.I. (1994). Comparison of field methods for measuring surface-topography and their associations with a tropical reef fish assemblage. *Marine Ecology Progress Series, 112*, 87-96

Minghelli-Roman, A., Chisholm, J.R.M., Marchioretti, M., & Jaubert, J.M. (2002). Discrimination of coral reflectance spectra in the Red Sea. *Coral Reefs, 21*, 307-314

Mobley, C.D. (1994). *Light and Water: Radiative Transfer in Natural Waters*. San Diego: Academic Press, Moguerza, J.M., & Muñoz, A. (2006). Support vector machines with applications. *Statistical Science, 21*, 322-336

Moisen, G.G., Freeman, E.A., Blackard, J.A., Frescino, T.S., Zimmermann, N.E., & Edwards, T.C. (2006). Predicting tree species presence and basal area in Utah: A comparison of stochastic gradient boosting, generalized additive models, and tree-based methods. *Ecological Modelling, 199*, 176-187

Moisen, G.G., & Frescino, T.S. (2002). Comparing five modelling techniques for predicting forest characteristics. *Ecological Modelling, 157*, 209-225

Mumby, P., Green, E., Edwards, A., & Clark, C. (1997a). Measurement of seagrass standing crop using satellite and digital airborne remote sensing. *Marine Ecology Progress Series, 159*, 51-60

Mumby, P.J. (2001). Beta and habitat diversity in marine systems: a new approach to measurement, scaling and interpretation. *Oecologia, 128*, 274-280

Mumby, P.J., Broad, K., Brumbaugh, D.R., Dahlgren, C.P., Harborne, A.R., Hastings, A., Holmes, K.E., Kappel, C.V., Micheli, F., & Sanchirico, J.N. (2008). Coral reef habitats as surrogates of species, ecological functions, and ecosystem services. *Conservation Biology, 22*, 941-951

Mumby, P.J., Clark, C.D., Green, E.P., & Edwards, A.J. (1998). Benefits of water column correction and contextual editing for mapping coral reefs. *International Journal of Remote Sensing, 19*, 203-210

Mumby, P.J., & Edwards, A.J. (2002). Mapping marine environments with IKONOS imagery: enhanced spatial resolution can deliver greater thematic accuracy. *Remote Sensing of Environment, 82*, 248-257

Mumby, P.J., Edwards, A.J., Arias-Gonzalez, J.E., Lindeman, K.C., Blackwell, P.G., Gall, A., Gorczynska, M.I., Harborne, A.R., Pescod, C.L., Renken, H., Wabnitz, C.C.C., & Llewellyn, G. (2004a). Mangroves enhance the biomass of coral reef fish communities in the Caribbean. *Nature, 427*, 533-536

Mumby, P.J., Green, E.P., Edwards, A.J., & Clark, C.D. (1997b). Coral reef habitat-mapping: how much detail can remote sensing provide? *Marine Biology, 130*, 193-202

Mumby, P.J., Hedley, J.D., Chisholm, J.R.M., Clark, C.D., Ripley, H., & Jaubert, J. (2004b). The cover of living and dead corals from airborne remote sensing. *Coral Reefs, 23*, 171-183

Mumby, P.J., Skirving, W., Strong, A.E., Hardy, J.T., LeDrew, E.F., Hochberg, E.J., Stumpf, R.P., & David, L.T. (2004c). Remote sensing of coral reefs and their physical environment. *Marine Pollution Bulletin, 48*, 219-228

Murray, K., & Conner, M.M. (2009). Methods to quantify variable importance: implications for the analysis of noisy ecological data. *Ecology, 90*, 348-355

Muslim, A.M., & Foody, G.M. (2008). DEM and bathymetry estimation for mapping a tide-coordinated shoreline from fine spatial resolution satellite sensor imagery. *International Journal of Remote Sensing, 29*, 4515-4536

Muthiga, N., Riedmiller, S., Carter, E., van der Elst, R., Mann-Lang, J., Horrill, C., & McClanahan, T.R. (2000). Management Status and Case Studies. In T.R. McClanahan, C.R.C. Sheppard & D.O. Obura (Eds.), *Coral Reefs of the Indian Ocean* (pp. 473-505). Oxford: Oxford University Press

Nagelkerken, I., Dorenbosch, M., Verberk, W., de la Moriniere, E.C., & van der Velde, G. (2000). Importance of shallow-water biotopes of a Caribbean bay for juvenile coral reef fishes: patterns in biotope association, community structure and spatial distribution. *Marine Ecology Progress Series, 202*, 175-192

Newman, C., Knudby, A., & LeDrew, E. (2007). Assessing the effect of management zonation on live coral cover using multi-date IKONOS satellite imagery. *Journal of Applied Remote Sensing, 1*, 26 December 2007

Newman, C., & LeDrew, E. (2008). A socio-remote sensing approach to coral reef management. *Coastal Management, submitted*

Ngoile, M.A.K. (1990). Ecological baseline surveys of coral reefs and intertidal zones around Mnemba Island and Zanzibar Town. The Commission for Lands and Environment, Zanzibar,

Ngusaru, A. (2002). Geological history. In M.D. Richmond (Ed.), *A Field Guide to the Seashores of Eastern Africa and the Western Indian Ocean Islands* (p. 461). Milano: SIDA/SAREC - UDSM

Ormond, R.F.G., Roberts, J.M., & Jan, R.Q. (1996). Behavioural differences in microhabitat use by damselfishes (Pomacentridae): Implications for reef fish biodiversity. *Journal of Experimental Marine Biology and Ecology, 202*, 85-95

Paulay, G. (1997). Diversity and Distribution of Reef Organisms. In C. Birkeland (Ed.), *Life and Death of Coral Reefs* (pp. 298-353). New York: Chapman & Hall

PCI Geomatics (2003). Geomatica Focus 9.1.0.

Persson, M., & Tryman, K. (2003). Coral reef status comparisons between Chumbe, Bawe and Changuu islands off Zanzibar, Tanzania. Swedish University of Agricultural Sciences, ISSN 1402-3237, pp. 18

Peters, A., & Hothorn, T. (2007). ipred: Improved predictors, R package.

Philpot, W.D. (1989). Bathymetric Mapping with Passive Multispectral Imagery. *Applied Optics, 28*, 1569-1578

Phinn, S., Roelfsema, C., Dekker, A., Brando, V., & Anstee, J. (2008). Mapping seagrass species, cover and biomass in shallow waters: An assessment of satellite multi-spectral and airborne hyper-spectral imaging systems in Moreton Bay (Australia). *Remote Sensing of Environment, 112*, 3413-3425

Pittman, S.J., Christensen, J.D., Caldow, C., Menza, C., & Monaco, M.E. (2007). Predictive mapping of fish species richness across shallow-water seascapes in the Caribbean. *Ecological Modelling, 204*, 9-21

Pittman, S.J., Costa, M.B., & Battista, T.A. (2009). Using lidar bathymetry and boosted regression trees to predict the diversity and abundance of fish and corals. *Journal of Coastal Research, in press*

Pittman, S.J., McAlpine, C.A., & Pittman, K.M. (2004). Linking fish and prawns to their environment: a hierarchical landscape approach. *Marine Ecology-Progress Series, 283*, 233-254

Prasad, A.M., Iverson, L.R., & Liaw, A. (2006). Newer classification and regression tree techniques: Bagging and random forests for ecological prediction. *Ecosystems, 9*, 181-199

Purkis, S.J., Graham, N.A.J., & Riegl, B.M. (2008). Predictability of reef fish diversity and abundance using remote sensing data in Diego Garcia (Chagos Archipelago). *Coral Reefs, 27*, 167-178

Purkis, S.J., Myint, S.W., & Riegl, B.M. (2006). Enhanced detection of the coral Acropora cervicornis from satellite imagery using a textural operator. *Remote Sensing of Environment, 101*, 82-94

R Core Development Team (2008). R: a language and environment for statistical computing. http://www.R-project.org/

Reaka-Kudla, M.L. (1997). The Global Biodiversity of Coral Reefs: A Comparison with Rain Forests. In M.L.W.D.E.W.E.O. Reaka-Kudla (Ed.), *Biodiversity II* (pp. 83-107). Washington DC: Joseph Henry Press

Ridgeway, G. (2007). gbm: Generalized Boosted Regression Models, R package.

Risk, M.J. (1972). Fish diversity on a coral reef in the Virgin Islands. *Atoll Research Bulletin, 153*, 1-6

Roberts, C.M., McClean, C.J., Veron, J.E.N., Hawkins, J.P., Allen, G.R., McAllister, D.E., Mittermeier, C.G., Schueler, F.W., Spalding, M., Wells, F., Vynne, C., & Werner, T.B. (2002). Marine biodiversity hotspots and conservation priorities for tropical reefs. *Science, 295*, 1280-1284

Sale, P.F. (2002). The science we need to develop for more effective management. In P.F. Sale (Ed.), *Coral reef fishes: dynamics and diversity in a complex ecosystem* (pp. 564-596). London: Academic Press

Sale, P.F., & Dybdahl, R. (1975). Determinants of Community Structure for Coral-Reef Fishes in an Experimental Habitat. *Ecology, 56*, 1343-1355

Salm, R.V., & Clark, J.R. (2000). *Marine and Coastal Protected Areas*. Cambridge: IUCN, pp. 370

Shannon, C.E. (1948). A mathematical theory of communication. *AT&T Technical Journal, 27*, 379-423

Simes, R.J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika, 73*, 4

Smith, V.E., Rogers, R.H., & Reed, L.E. (1975). Automated mapping and inventory of Great Barrier Reef zonation with Landsat*, 1*, 775-780

Smith, V.E., Rogers, R.H., Reed, L.E. (1975). Automated mapping and inventory of Great Barrier Reef zonation with Landsat. *IEEE Ocean Conference, 1*, 775-780

Spaceimaging (2001). IKONOS Relative Spectral Response and Radiometric Cal Coefficients. Spaceimaging, pp.?

Stevens, T., & Connolly, R.M. (2004). Testing the utility of abiotic surrogates for marine habitat mapping at scales relevant to management. *Biological Conservation, 119*, 351-362

Strobl, C., Boulesteix, A.L., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics, 8*

Stumpf, R.P., Holderied, K., & Sinclair, M. (2003). Determination of water depth with high-resolution satellite imagery over variable bottom types. *Limnology and Oceanography, 48*, 547-556

Su, H.B., Liu, H.X., & Heyman, W.D. (2008). Automated Derivation of Bathymetric Information from Multi-Spectral Satellite Imagery Using a Non-Linear Inversion Model, 281-298

Suzuki, H., Matsakis, P., Andréfouët, S., & Desachy, J. (2001). Satellite image classification using expert structural knowledge: a method based on fuzzy partition computation and simulated annealing*, (CD-ROM)*

Syms, C., & Jones, G.P. (2000). Disturbance, habitat structure, and the dynamics of a coral-reef fish community. *Ecology, 81*, 2714-2729

Taylor, A.E. (2009). Statistical enhancement of support vector machines. *Department of Statistics, Doctor of Philosophy*, 156

Tuell, G., & Park, J.Y. (2004). Use of SHOALS bottom reflectance images to constrain the inversion of a hyperspectral radiative transfer model*, 5412*, 185-193

Tuell, G., Ramnath, V., Park, J.Y., Feygels, V., Aitken, J., & Kopilevich, Y. (2005). Fusion of SHOALS Bathymetric Lidar and Passive Spectral Data for Shallow Water Rapid Environmental Assessment*, 2*, 1046-1051

Turner, J., & Klaus, R. (2005). Coral reefs of the Mascarenes, Western Indian Ocean. *Philosophical Transactions of the Royal Society of London, Series A, 363*, 229-250

Tyler, E.H.M. (2005). The effect of fully and partially protected marine reserves on coral reef fish populations on Zanzibar, Tanzania. *Linacre College, PhD*, 215

Veron, J.E.N. (1995). *Corals in Space and Time: The Biogeography and Evolution of the Scleractinia*. Ithica: Cornell University Press,

Veron, J.E.N. (2000). *Corals of the World*. Townsville: AIMS,

Veron, J.E.N. (2008). Mass extinctions and ocean acidification: biological constraints on geological dilemmas. *Coral Reefs, 27*, 459-472

Ward, T.J., Vanderklift, M.A., Nicholls, A.O., & Kenchington, R.A. (1999). Selecting marine reserves using habitats and species assemblages as surrogates for biological diversity. *Ecological Applications, 9*, 691-698

Wedding, L., Friedlander, A.M., McGranaghan, M., Yost, R., & Monaco, M. (2008). Using bathymetric LIDAR to define nearshore benthic habitat complexity: Implications for management of reef fish assemblages in Hawaii. *Remote Sensing of Environment, 112*, 4159-4165

Wilhm, J.L. (1968). Use of biomass units in Shannon's formula. *Ecology, 49*, 153-&

Wilkinson, C. (2004). *Status of Coral Reefs of the World*. Townsville: Australian Institute of Marine Science, pp. 316

Wilson, E.O. (1997). Introduction. In M.L. Reaka-Kudla, Wilson, Don E., Wilson, Edward O. (Ed.), *Biodiversity II* (p. 551). Washington DC: Joseph Henry Press

Wilson, S.K., Graham, N.A.J., & Polunin, N.V.C. (2007). Appraisal of visual assessments of habitat complexity and benthic composition on coral reefs. *Marine Biology, 151*, 1069-1076

Wood, E.M. (1983). *Corals of the world*. Neptune City: T.F.H. Publications, pp. 256

Wood, R. (1998). The ecological evolution of reefs. *Annual Review of Ecology and Systematics, 29*, 179-206

Yonge, C.M. (1940). *The biology of reef building corals*. London: British Museum,